ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Introducción a la Ciencia de Datos con R

Miguel Jorquera

Educación Profesional
Escuela de Ingeniería

RESUMEN

INTERACCIÓN CON APIs

¿Que es una API?

- API: **A**pplication **P**rogramming **I**nterface
- Son instrucciones programadas sobre cómo interactuar con una pieza de software, por ejemplo:
  - Un package e R/Python u otro lenguaje,
  - Otra API (pública por ejemplo)
  - Una base de datos
  - Un sistema operativo

INTRO: WEB SCRAPING CON R

EDUCACIÓN PROFESIONAL

Conceptos básicos

Para esta sección utilizaremos el siguiente recurso online

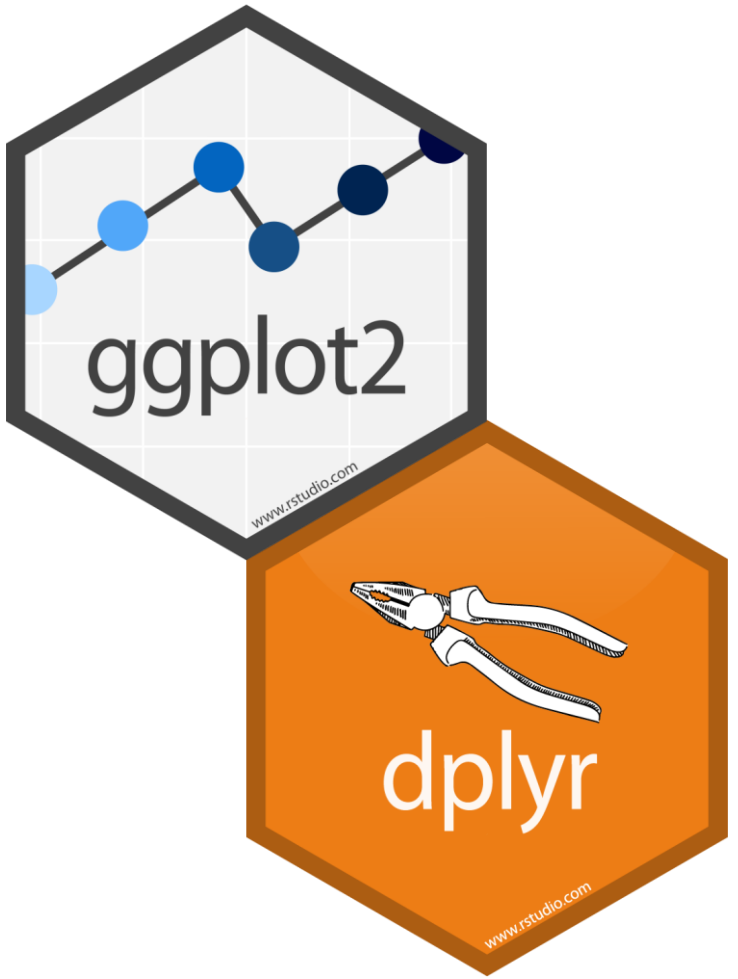
- https://cfss.uchicago.edu/notes/web-scraping/

TEMAS PARA HOY

RESUMEN

Manipulación de tablas

- Análisis exploratorio de datos
  - `dplyr`:
    - Manipulación de tablas
  - `ggplot2`:
    - Gramática de gráficos



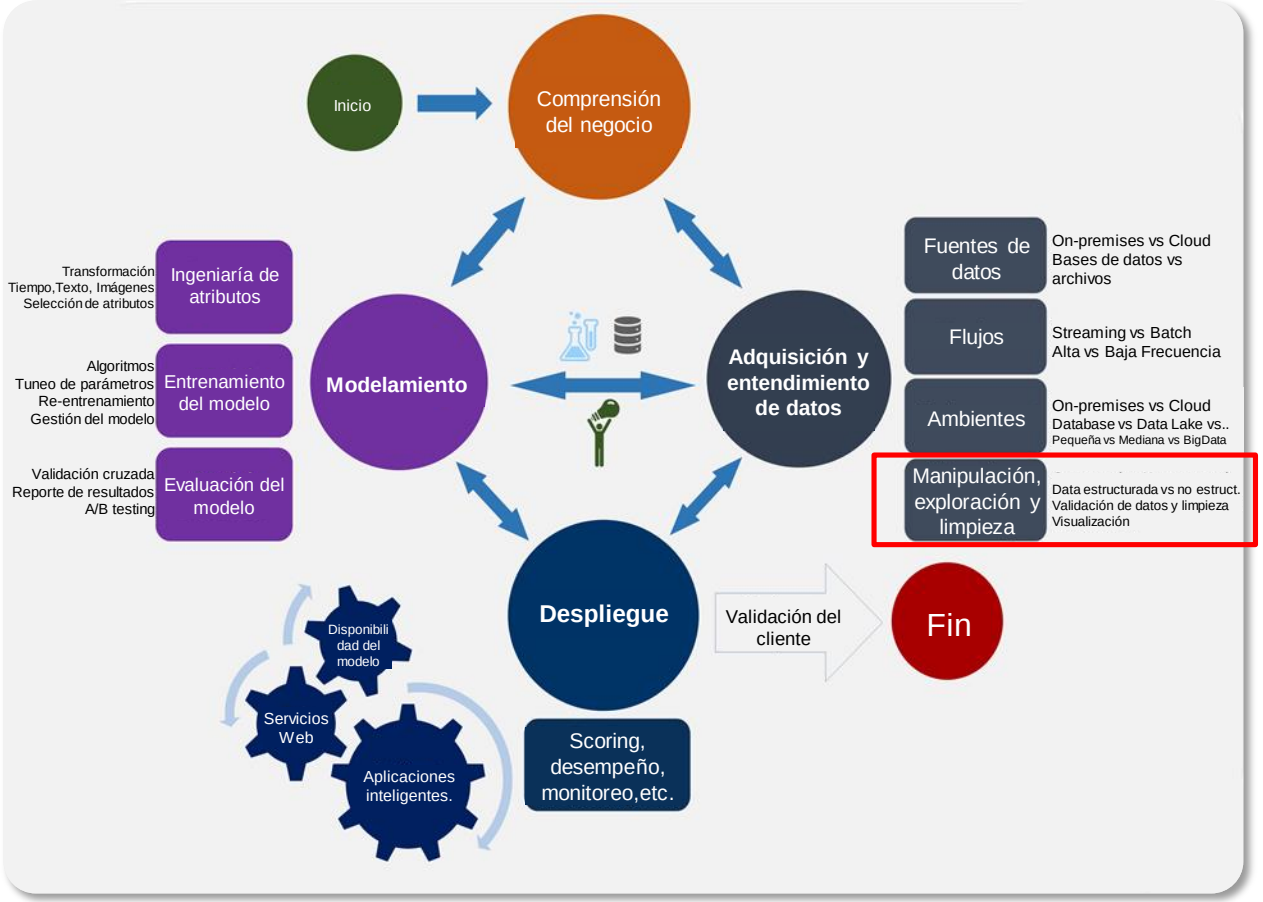
CONTEXTO

EDUCACIÓN PROFESIONAL

TDSP

La metodología TDSP (Team Data Science Process) propuesta por Microsoft es una metodología ágil, iterativa y eficiente, que promueve la colaboración entre los distintos miembros del equipo de desarrollo así como la interacción permanente con el cliente.

Flujo de trabajo en Data Science

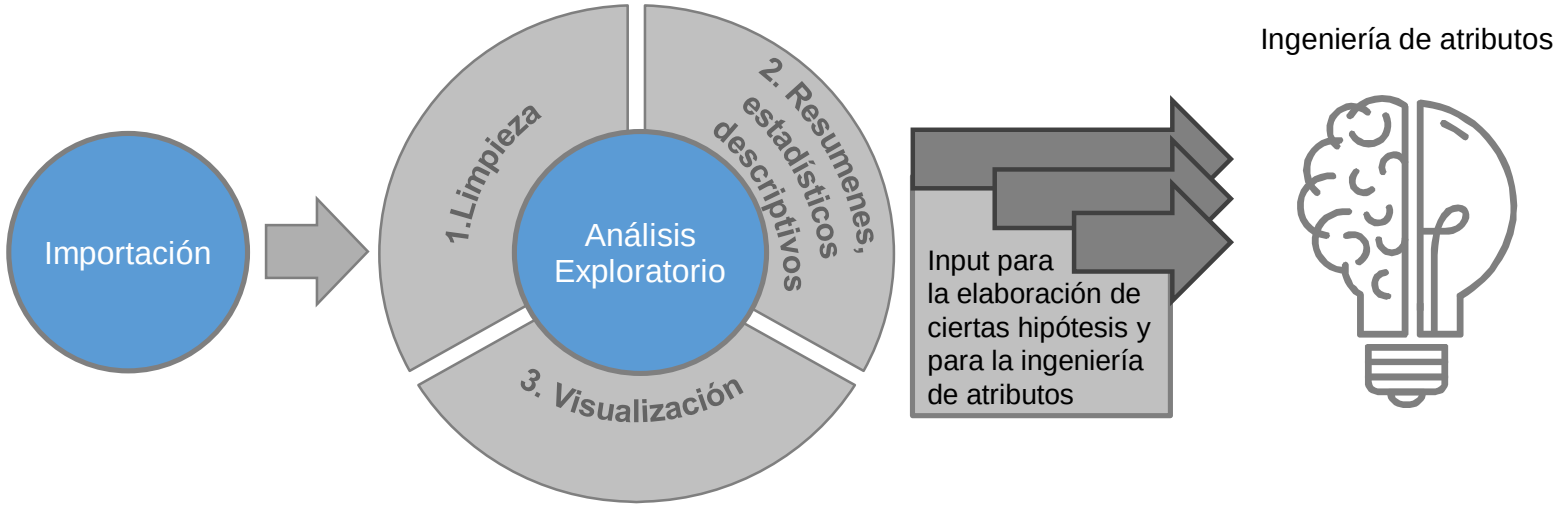


ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

CONTEXTO

Manipulación, Exploración y Limpieza

- Esta fase la denominaremos como fase exploratoria. En ella se llevarán a cabo los siguientes procesos
  - Se valida la consistencia de los datos proporcionados
  - Se describen los datos importados a nivel estadístico y visual.
  - Se plantean hipotesis sobre las relaciones entre las variables existentes.
  - Se da paso a la ingeniería de atributos



LIMPIEZA YMANIPULACIÓN

Tareas usuales

- Al importar datos a un nuevo ambiente (R en nuestro caso), es de utilidad chequear los siguientes aspectos
  - Nombres de las columnas en formato estándar.
  - Validar la consistencia de los tipos de dato de cada campo.
  - Verificar total de registros importados.
  - Analizar e imputar datos faltantes

EXPLORACIÓN Y LIMPIEZA

Herramientas disponibles

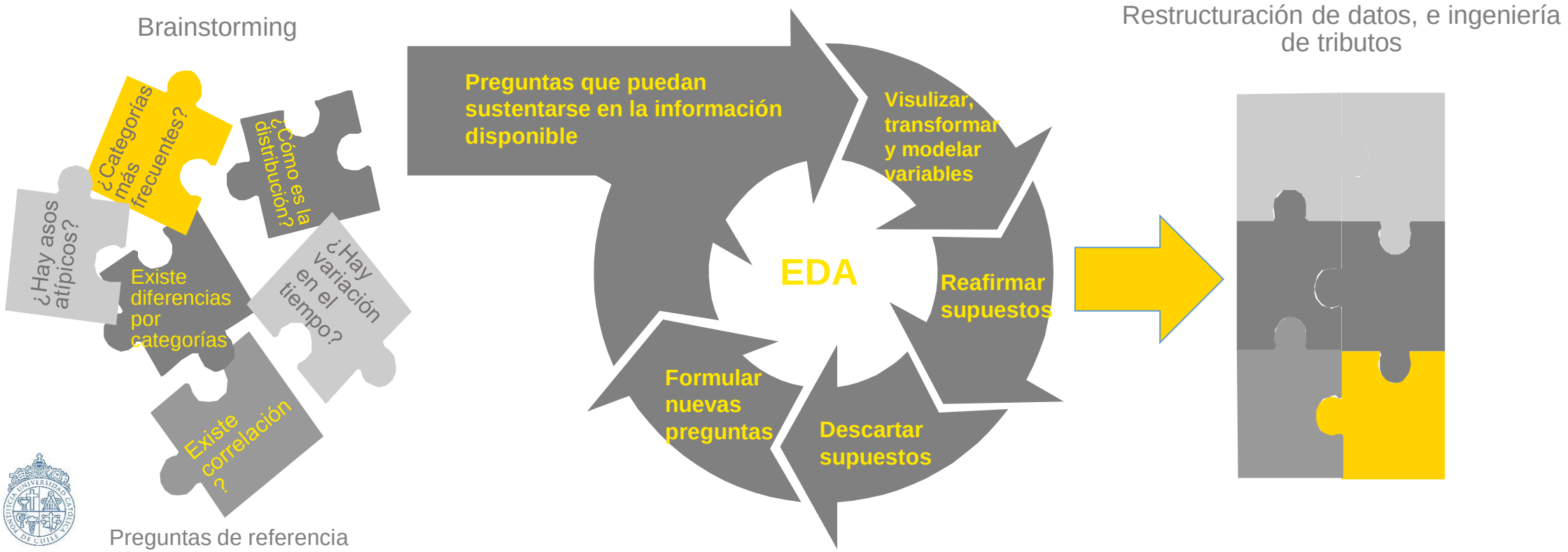
- Si bien hay variedad de herramientas para llevar acabo la fase exploratoria, nosotros nos centraremos en la utilización de dos packages principalmente
  - `dplyr` para consultas
    - Generación de información agregada.
    - Tablas de frecuencia.
    - Facilita el cálculo de estadísticos descriptivos en general
  - `ggplot2` para visualización
    - Gráficos univariados y bivariados.
    - Visualización de variables continuas y categóricas
    - Gráficos de dispersión
    - Gráficos temporales
    - Visualización de distribuciones.

EXPLORATORY DATA ANALYSIS (EDA)

EDUCACIÓN PROFESIONAL

Proceso Iterativo

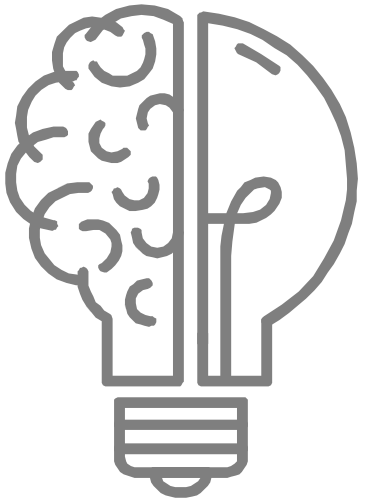
- La fase de análisis exploratorio, es un proceso iterativo que se caracteriza por la formulación de preguntas de interés que permitan guiar el análisis cuyas respuestas tenga sustento en la información disponible



EXPLORATORY DATA ANALYSIS (EDA)

Características

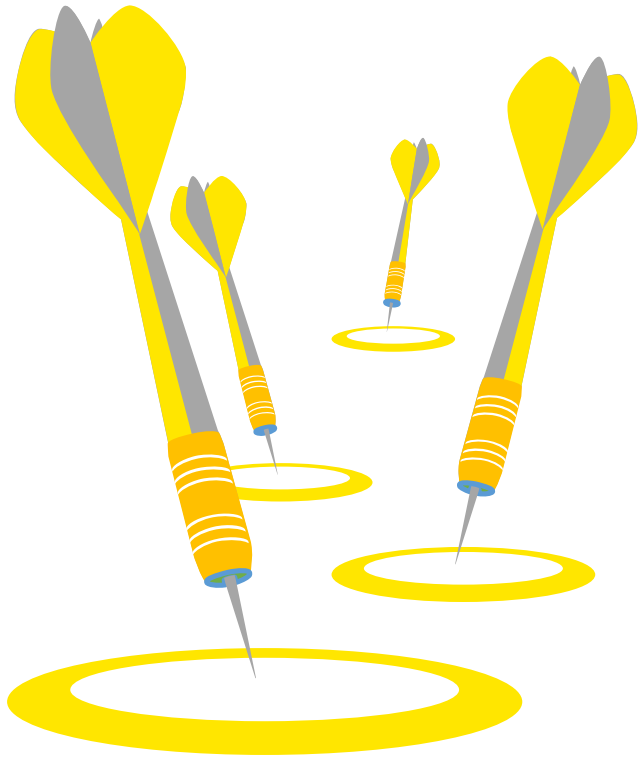
- EDA no es un proceso formal con reglas estrictas. Es un **proceso creativo**!
  - Hay libertad de explorar toda idea inicial. Algunas llegará a punto muerto, otras serán dignas de ser comunicadas.
  - La limpieza es parte de la exploración. Podemos **visualizar**, **transformar** e incluso **modelar** en esta fase!.



EXPLORATORY DATA ANALYSIS (EDA)

EDUCACIÓN PROFESIONAL

Cómo guiar el análisis



- ¿Cómo generar buenas preguntas?
  - **Generando muchas preguntas!**
- ¿Cómo comenzar? Dos focos iniciales:
  - **¿Qué tipos de variaciones presentan mis variables?**
  - **¿Qué tipo de co-variación existe entre mis variables?**

EXPLORATORY DATA ANALYSIS (EDA)

EDUCACIÓN PROFESIONAL

Cómo guiar el análisis

- Para medir variaciones usualmente es de utilidad:
  - Histogramas y gráficos de barra.
  - Tablas de frecuencia
  - Polígonos de frecuencia
  - Boxplots y gráficos de violín

  → Se caracterizan los valores típicos.
  → Se observan posibles casos atípicos y anomalías.

- Para el caso de la co-variación
  - Medidas de correlación
  - Gráficos de dispersión
  - Tablas de contingencia

  → Búsqueda de patrones
    - Tendencias
    - Clusters

    → ¿Coincidencia?
    → ¿Es posible describir el posible patrón?
    → ¿Qué tan "fuerte" es la relación de dependencia dada por el patrón?
    → ¿Otras variables podrían afectar al patrón observado?
    → ¿Cambia la relación si se observan subgrupos idividuales?

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

INTRODUCCIÓN A GGPLOT

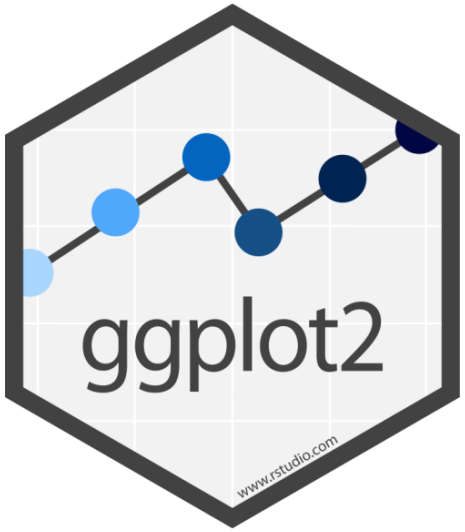
GRAMÁTICA DE GRÁFICOS

ggplot2

Ggplot2 es un package basado en la gramática de gráficos, creada por Wilkinson (2015). Básicamente la gramática de gráficos nos dice que un gráfico estadístico corresponde a un mapeo entre la data hacia atributos estéticos (como el color, forma o tamaño) de objetos geométricos (como puntos, líneas o barras).

Referencias útiles

- https://ggplot2.tidyverse.org/reference/index.html
- https://r4ds.had.co.nz/data-visualisation.html
- https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0



https://ggplot2.tidyverse.org/

Otras referecias

highcharter

- Official package website: http://jkunst.com/highcharter
- Replicating Highcharts Demos: https://cran.rstudio.com/web/packages/highcharter/vignettes/replicating-highcharts-demos.html
- CRAN site: https://cran.r-project.org/web/packages/highcharter/.
- Shiny demo code: https://github.com/jbkunst/shiny-apps/tree/master/highcharter.
- Referencia oficial (No R): http://highcharts.com

leaflet

- https://rstudio.github.io/leaflet/
- https://www.rdocumentation.org/packages/leaflet/versions/2.0.3
- https://github.com/rstudio/leaflet
- Referencia oficial (No R): https://leafletjs.com/reference-1.6.0.html

Shiny

- https://shiny.rstudio.com/
- https://www.shinyapps.io/

GRAMÁTICA DE GRÁFICOS

Vamos!