



Laboratorio Nacional de Políticas Públicas



CENTRO DE INVESTIGACIÓN
Y DOCENCIA ECONÓMICAS A.C.

ESTADÍSTICA

LABORATORIO I

MAESTRÍA EN PERIODISMO DE POLÍTICAS PÚBLICAS.

JORGE JUVENAL CAMPOS FERREIRA.

Presentación

Jorge Juvenal Campos Ferreira

M.C. En Economía por el Colegio de México.

Asistente de Investigación en el LNPP desde 2018.

Twitter: @JuvenalCamposF

Github: JuveCampos 

¿Por qué R?



¿Por qué R?



1. **¡Es gratis!, *open source*, y disponible para los principales sistemas operativos.**
2. **R tiene una comunidad muy grande de usuarios y programadores, generando material de ayuda y de referencia.**
3. **Existe una gran cantidad de paquetes y librerías.**
4. **Tiene herramientas de visualización y generación de reportes muy buenas.**

Fuente: Wickham, 2019. Advanced R. Introducción.

¿Por qué R?



5. Contamos con el IDE RStudio.

6. R se utiliza en desarrollos de Inteligencia Artificial y Machine Learning.

7. Entre otras ventajas (Vectorización, productos de RStudio, Metaprogramación, conexión con C, C++, etc).

Fuente: Wickham, 2019. Advanced R. Introducción.

Algunos inconvenientes de R.



- 1. La gente que usa R no prioriza la eficiencia o elegancia en su código. Se enfoca más en el resultado que en el proceso.**
- 2. No es un lenguaje de programación particularmente rápido** (aunque a nivel básico esto no representará un problema).
- 3. R utiliza una considerable cantidad de memoria de la computadora (RAM).**

Fuente: Wickham, 2019. Advanced R. Introducción.



Alternativas a R

Python. Lenguaje de programación muy utilizado en ciencia de datos, pero también en tareas de programación pura y dura, especialmente para páginas web. Al igual que R, es gratuito, libre y de código abierto.



Stata. Paquete de software estadístico muy utilizado en economía, sociología, ciencias políticas y algunas ramas de la medicina, y sirve para hacer estadística y econometría. No es libre, hay que pagar licencia, y no tiene una base de usuarios tan activa.







Excel. El programa de hojas de cálculo por excelencia. Nos permite limpiar y analizar bases de datos, así como hacer visualizaciones de datos (igual que R), pero no es libre, ni gratuito, ni permite la replicabilidad ni la automatización de procesos.



Otros más: Julia, SPSS, eViews, SAS, etc.

Alternativas a R

	Libre y Gratuito	Procesos replicables	Hojas de Cálculo	Usuarios	Librerías
 R	Si	Si	No	Muy usado en entornos académicos y de ciencias sociales	Muchas, comunidad muy activa
 Python	Si	Si	No	Muy usado en Modelos de Inteligencia Artificial y por programadores	Muchas, muchos programadores.
 Stata	No	Si	Algunas funciones	Usado mayoritariamente sólo para econometría y para ciencias sociales	Algunas, las elaboran principalmente los empleados de la empresa.
 Excel	No	No	Si	Ampliamente usado por todo mundo	Algunos plugins para hacer cosas avanzadas.

Fuente: Yo

Alternativas a RStudio

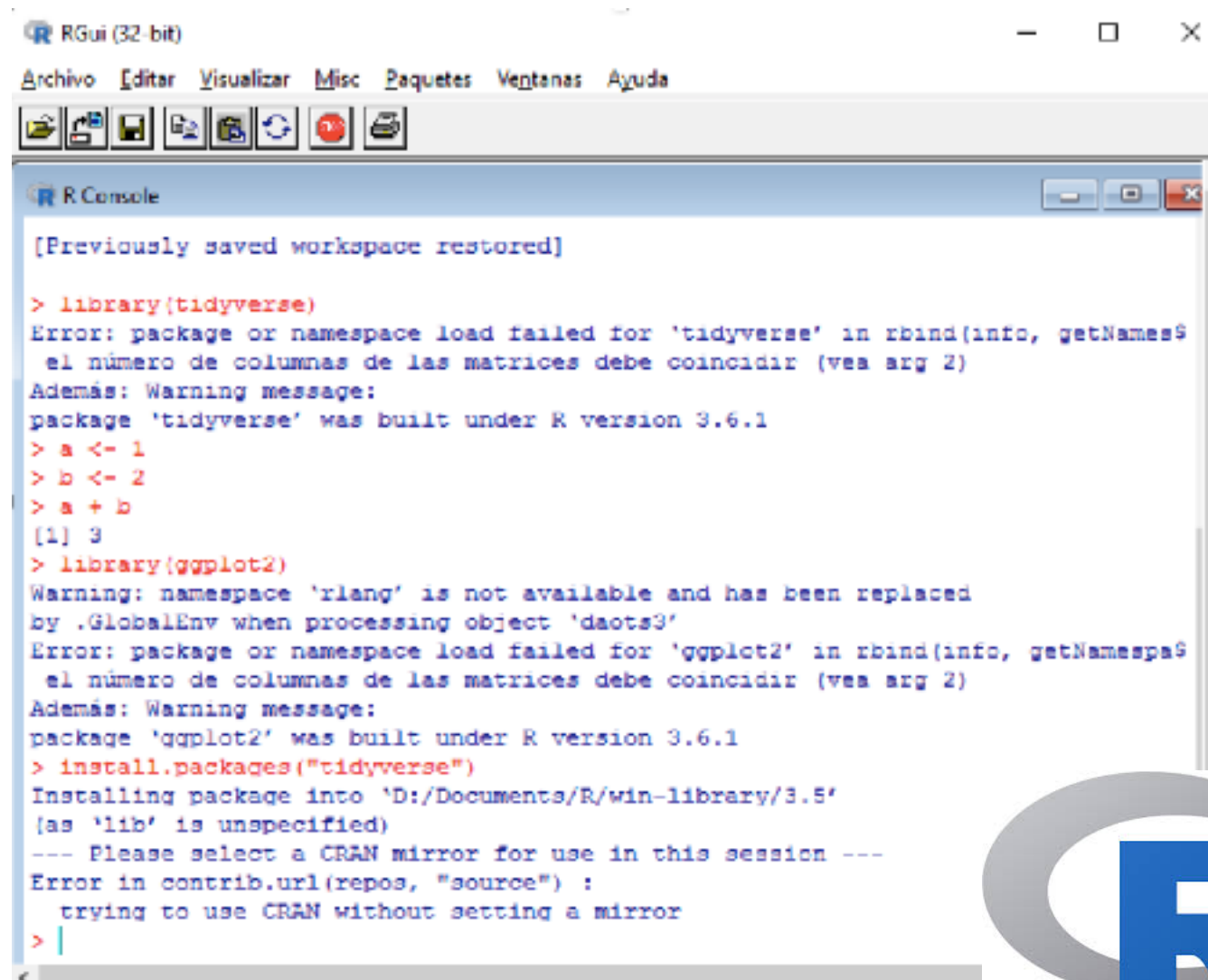


Alternativas a RStudio



(Recordemos que R != RStudio)

Consola de R. Es la consola default que instalamos al instalar R (no RStudio). No proporciona las facilidades que proporciona RStudio (editor de texto, visualizador web, etc).



```
RGui (32 bit)
Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda

R Console

[Previously saved workspace restored]

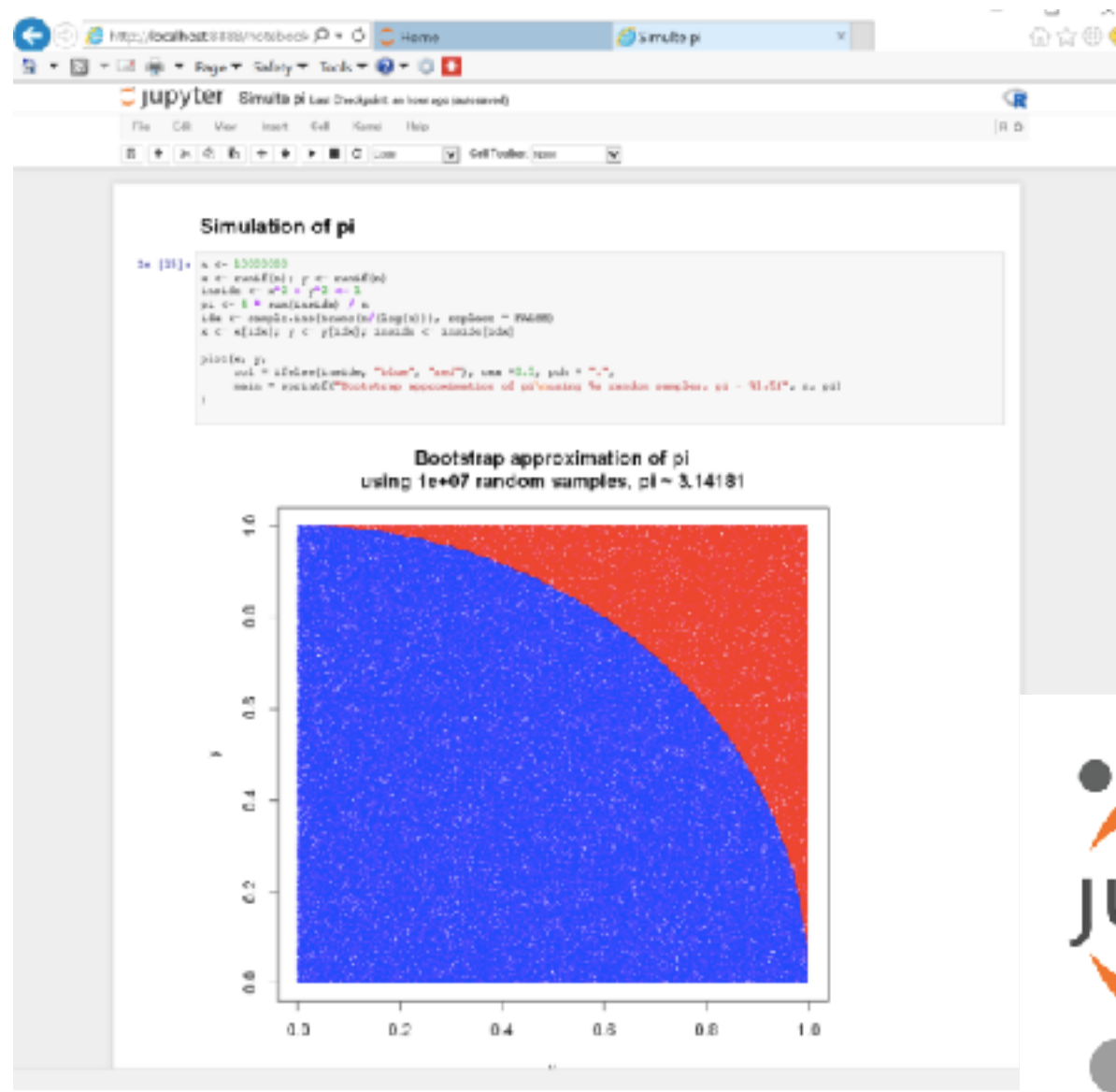
> library(tidyverse)
Error: package or namespace load failed for 'tidyverse' in rbind(info, getNames9
  el número de columnas de las matrices debe coincidir (vea arg 2)
Además: Warning message:
package 'tidyverse' was built under R version 3.6.1
> a <- 1
> b <- 2
> a + b
[1] 3
> library(ggplot2)
Warning: namespace 'rlang' is not available and has been replaced
by .GlobalEnv when processing object 'daots3'
Error: package or namespace load failed for 'ggplot2' in rbind(info, getNamespa9
  el número de columnas de las matrices debe coincidir (vea arg 2)
Además: Warning message:
package 'ggplot2' was built under R version 3.6.1
> install.packages("tidyverse")
Installing package into 'D:/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
Error in contrib.url(repos, "source") :
  trying to use CRAN without setting a mirror
> |
```



Alternativas a RStudio



Jupyter Notebook. Aplicación web *Open Source* que permite crear y compartir documentos que contienen código, ecuaciones. Visualizaciones y texto narrativo. Tiene varias de las funciones que tiene RStudio, pero la instalación es mucho más complicada, al igual que el arranque.





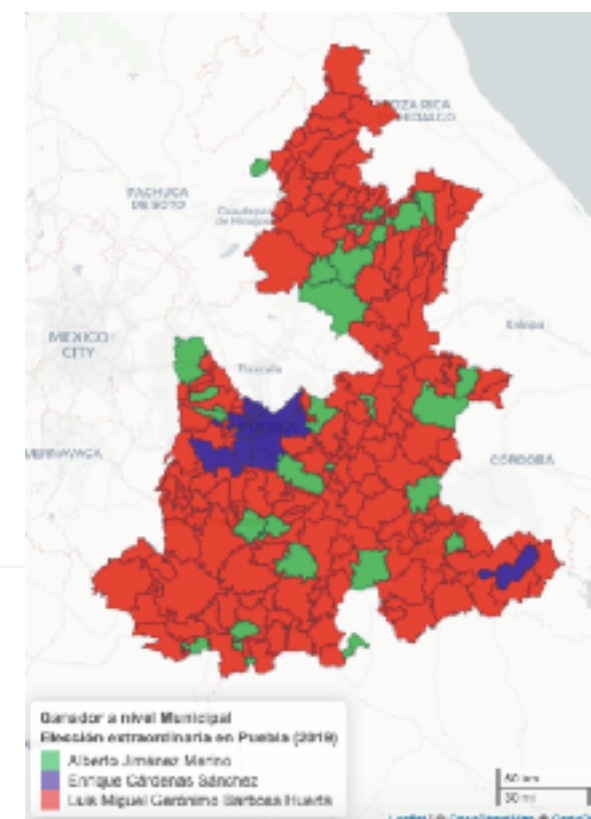
¿Qué se puede hacer con R?

Esto lo vamos a ver en la clase 😊

1. Manejo y manipulación de datos en R.

`library(tidyverse)`

```
72 datos <- prep %>%
73   select(ECS, AJM, LMGBH, TOTAL_VOTOS, LISTA_NOMINAL, MUNICIPIO, DISTRITO) %>%
74   filter(!is.na(MUNICIPIO)) %>%
75   group_by(MUNICIPIO) %>%
76   summarise(ECS = sum(ECS, na.rm = TRUE),
77             AJM = sum(AJM, na.rm = TRUE),
78             LMGBH = sum(LMGBH, na.rm = TRUE),
79             Total_Votos = sum(TOTAL_VOTOS, na.rm = TRUE),
80             ListaNominal = sum(LISTA_NOMINAL, na.rm = TRUE)
81   )
```



¿Qué se puede hacer con R?

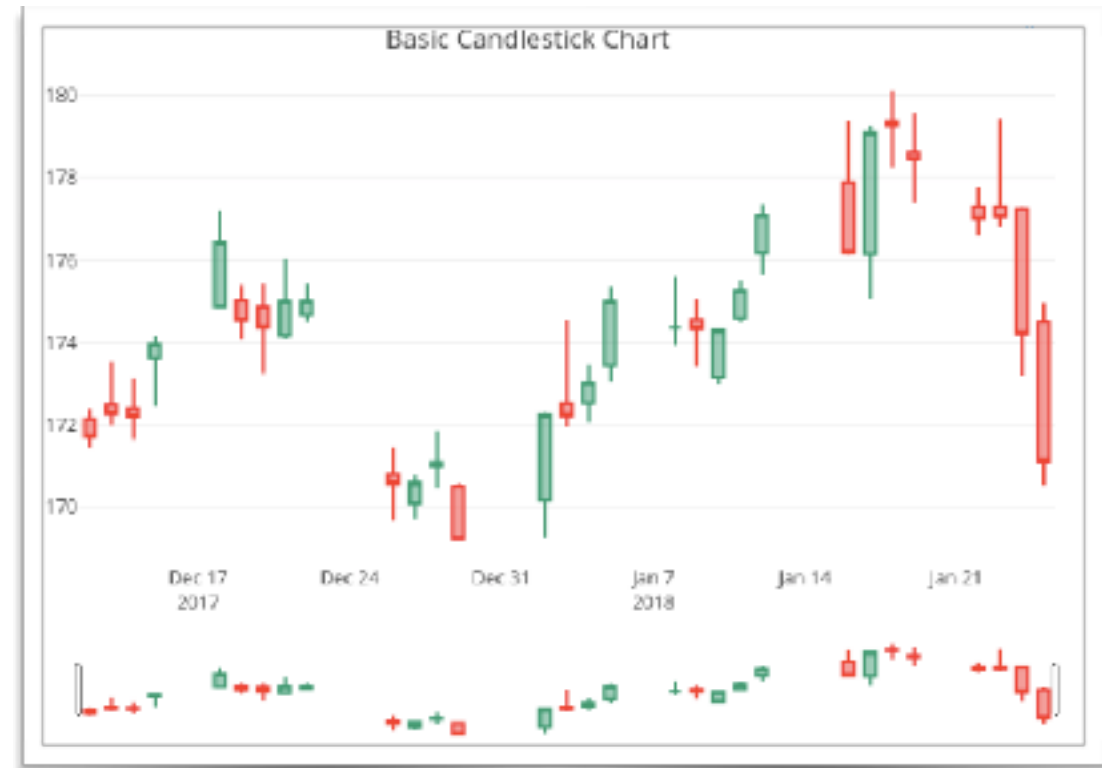
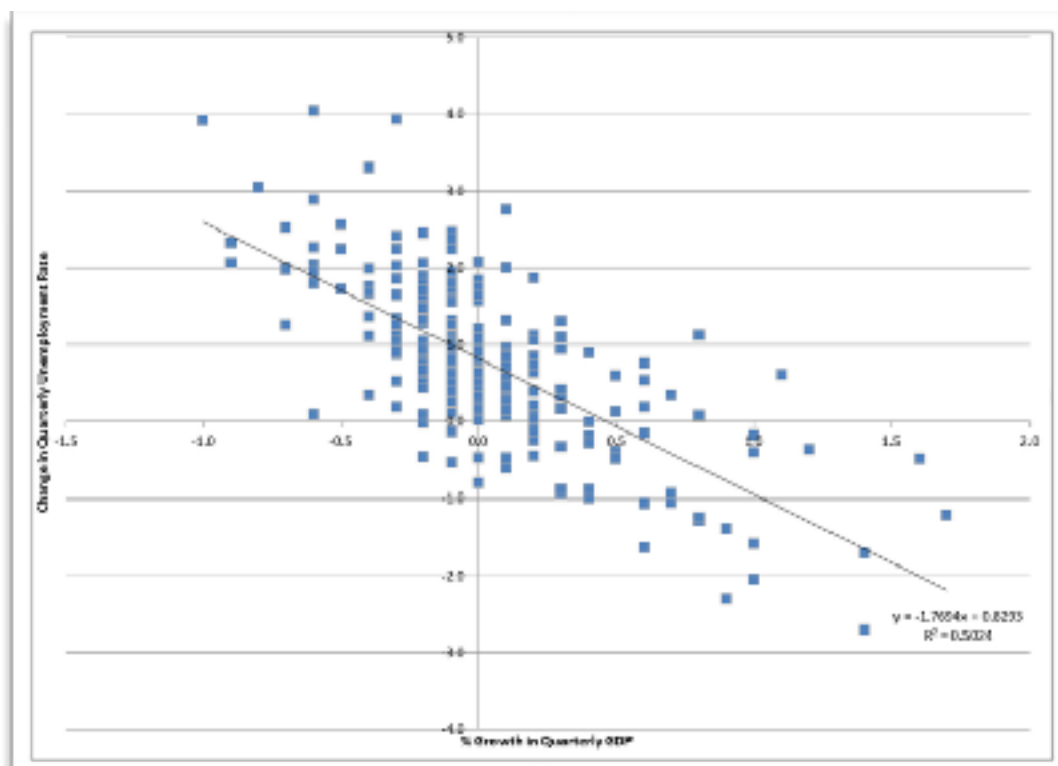
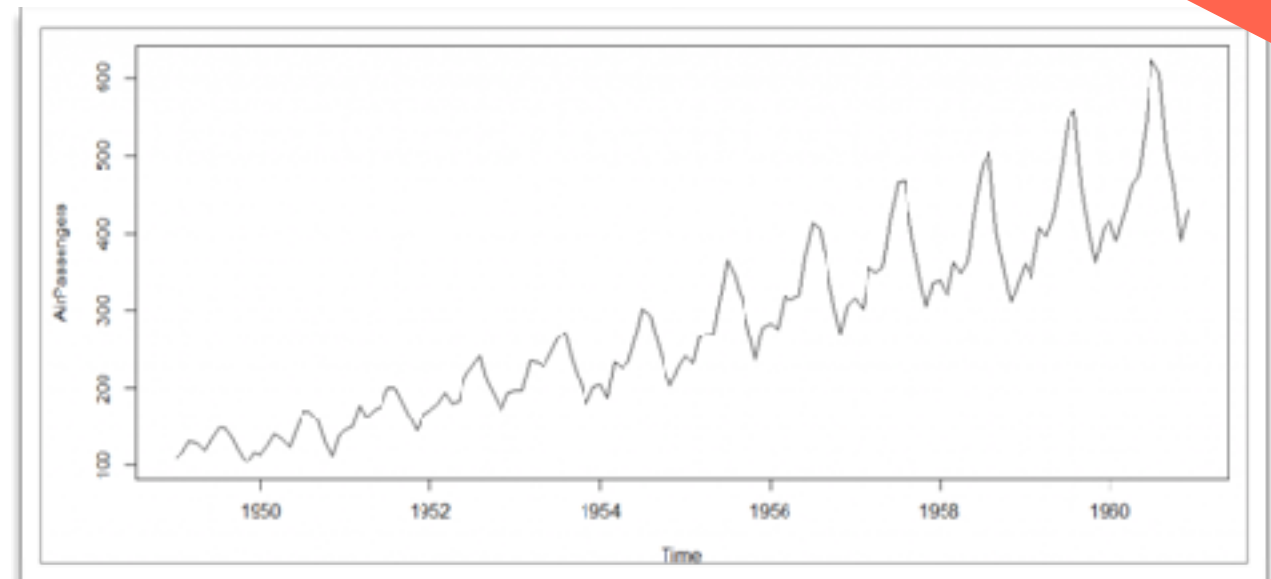


Esto no lo vamos a ver en la clase 😅

2. Análisis estadístico y econometría.

`library(base)`

`library(MASS)`



¿Qué se puede hacer con R?

Esto no lo vamos a ver en la clase 😊

3. Machine Learning y Deep Learning.

`library(e1071)`

`library(tensorflow)`

`library(caret)`

`library(rpart)`

etc.

Classification



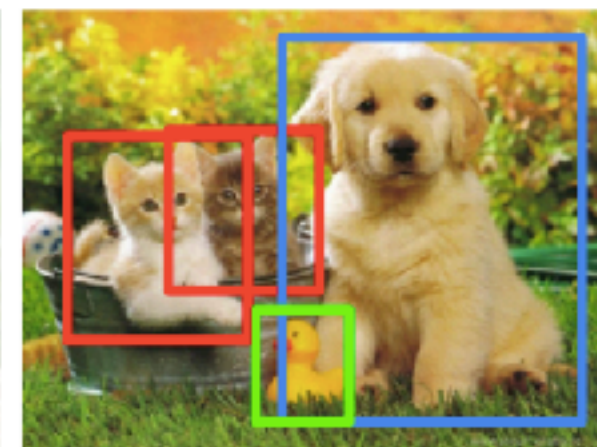
CAT

**Classification
+ Localization**



CAT

Object Detection



CAT, DOG, DUCK

**Instance
Segmentation**



CAT, DOG, D

Single object

Multiple objects

Fuente: <https://www.mocomakers.com/cats-versus-dogs/>

Esto lo vamos a ver en la clase 😊

library(tm)

library(stringr)



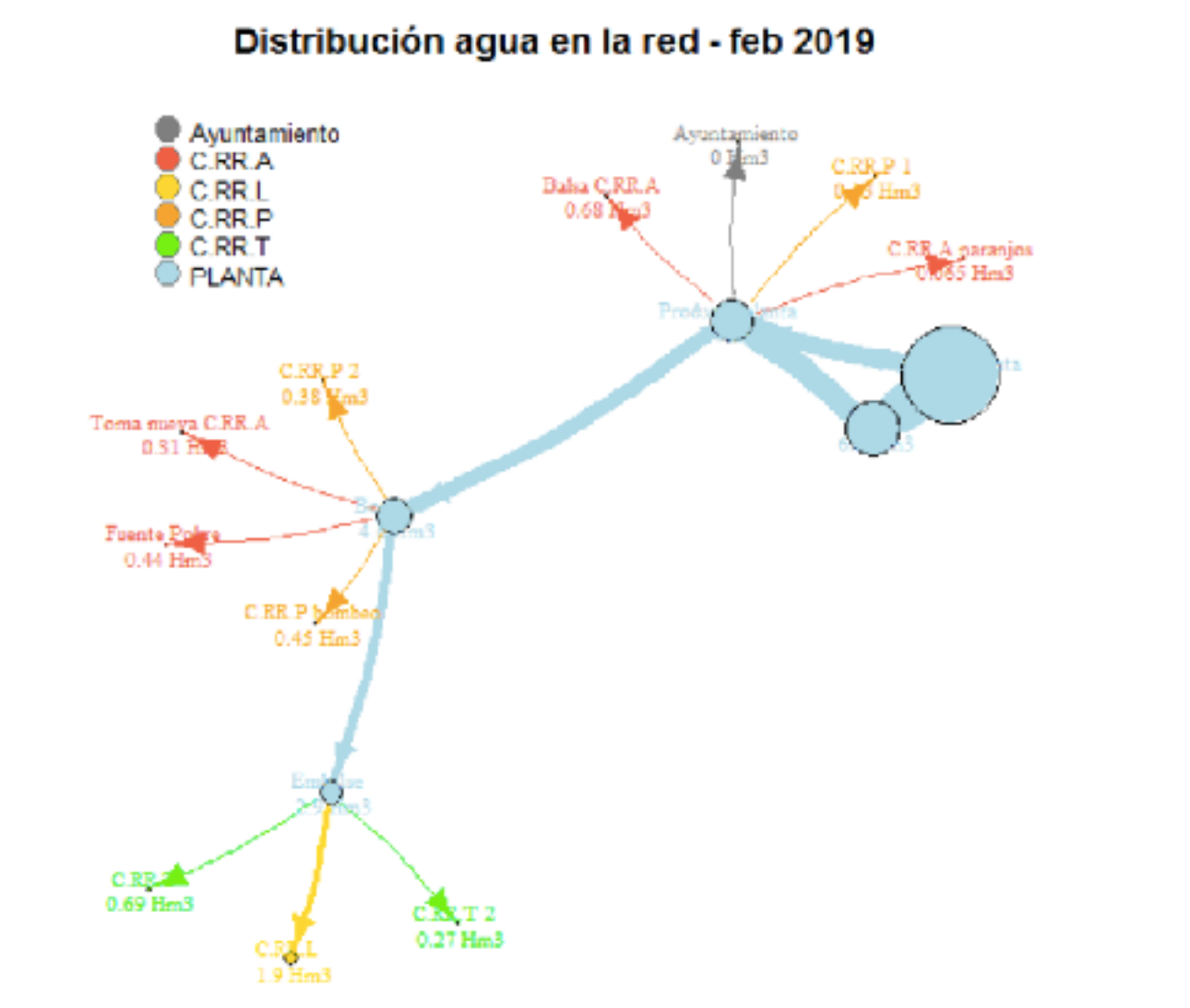
Nube de palabras.

Plan Nacional de Desarrollo. 2019.

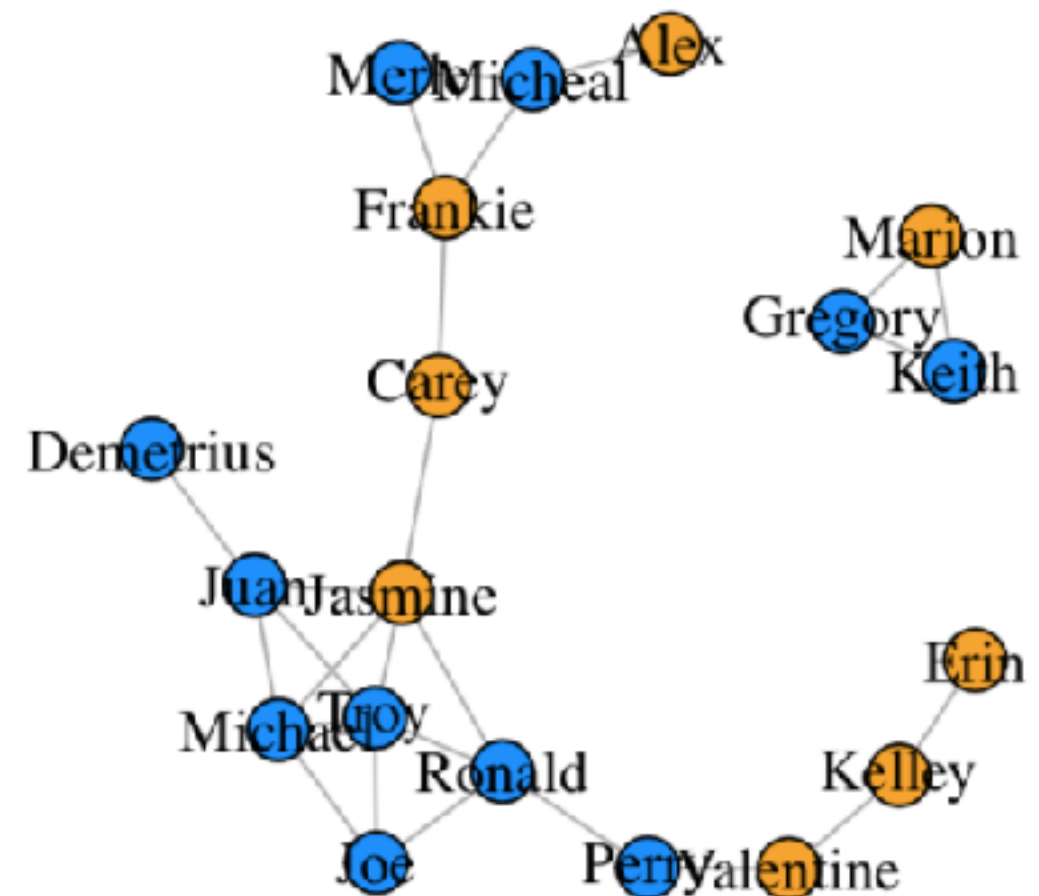
¿Qué se puede hacer con R?

5. Análisis de redes.

`library(igraph)`



Red de distribución de Agua



Red social de amigos en una prepa

¿Qué se puede hacer con R?

6. Visualización de datos.

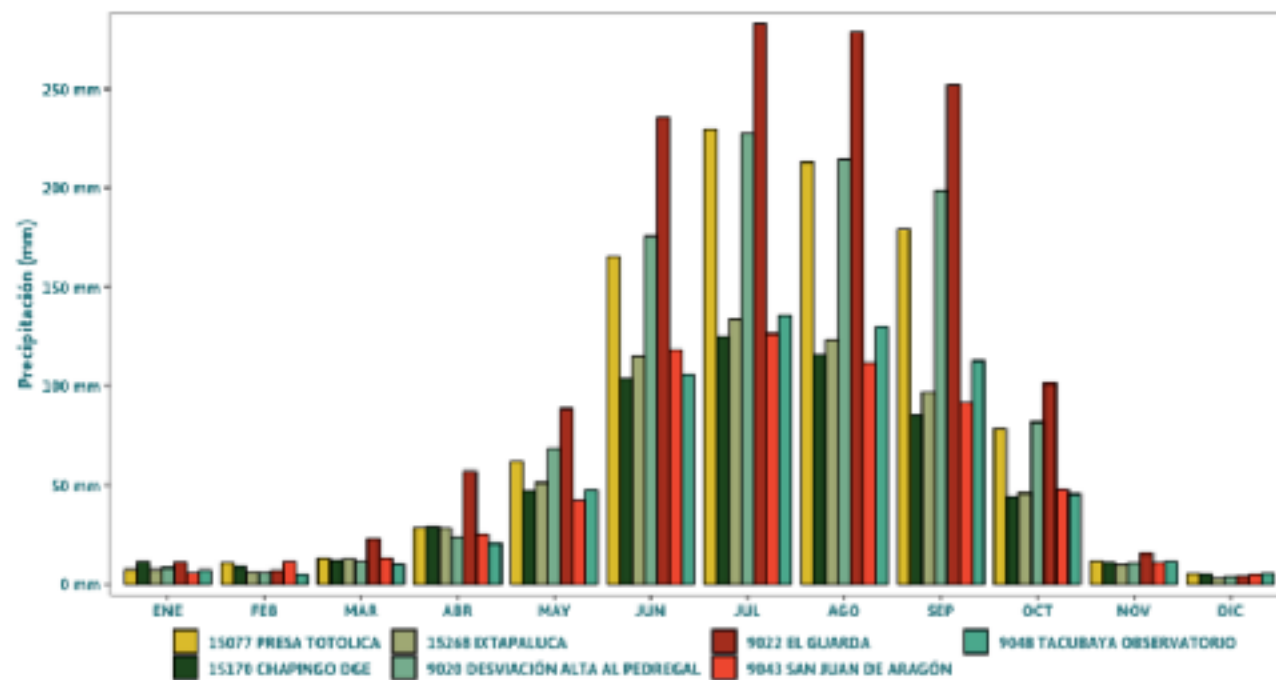
`library(ggplot2)`

`library(plotly)`

`library(leaflet)`

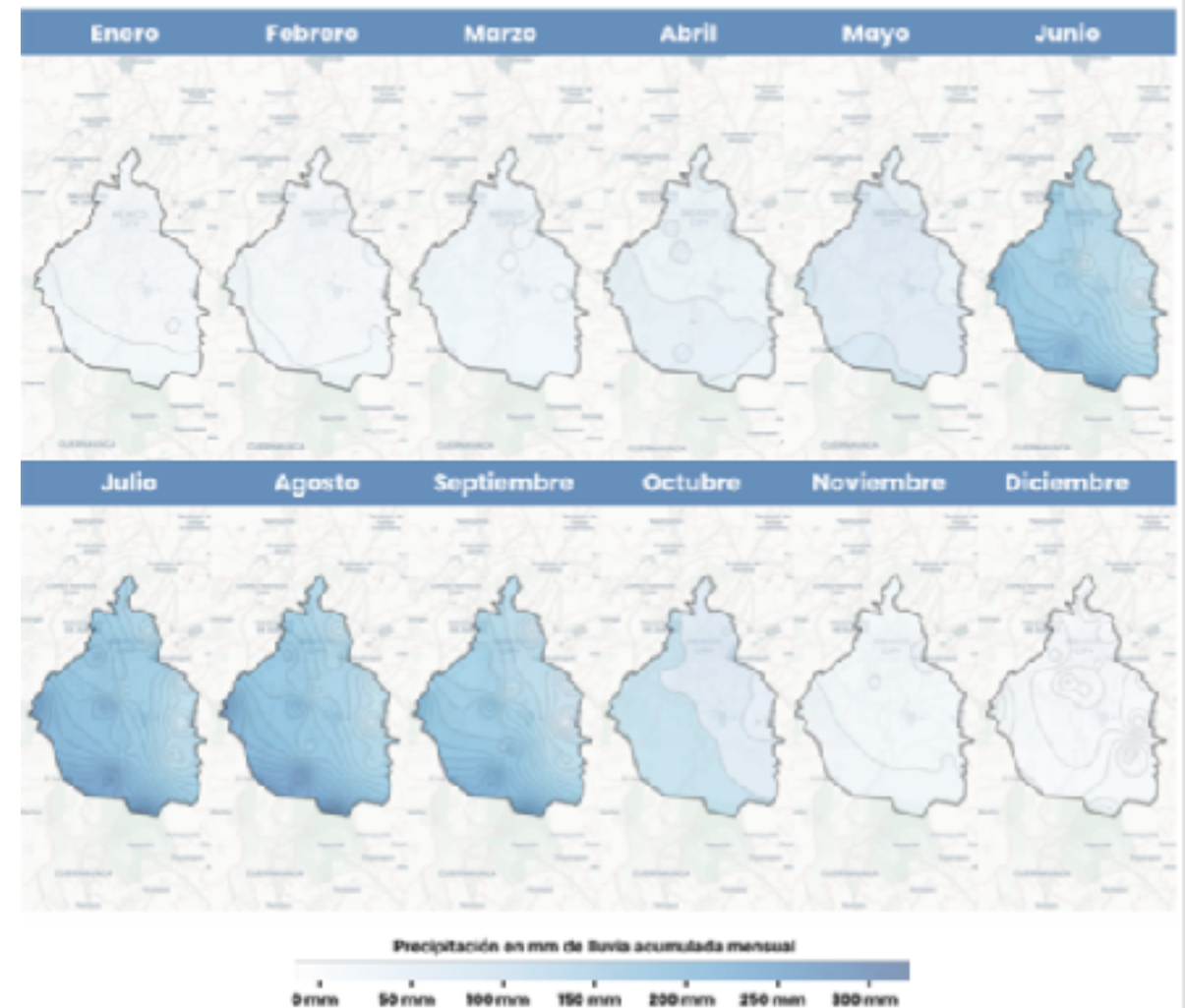
`library(htmlwidgets)`

Figura 24 Precipitación mensual promedio en 7 estaciones seleccionadas de la ZMVM.



Gráfica de barras de la distribución de la lluvia en la CDMX, en el tiempo.

Figura 25 Distribución de la precipitación de la lluvia en la Ciudad de México.



Fuente: Elaboración Propia con datos del Atlas de Riesgo de la Ciudad de México, SGIRPC, 2019.

Mapas de la distribución de la lluvia en la CDMX, en el espacio y tiempo.

¿Qué se puede hacer con R?

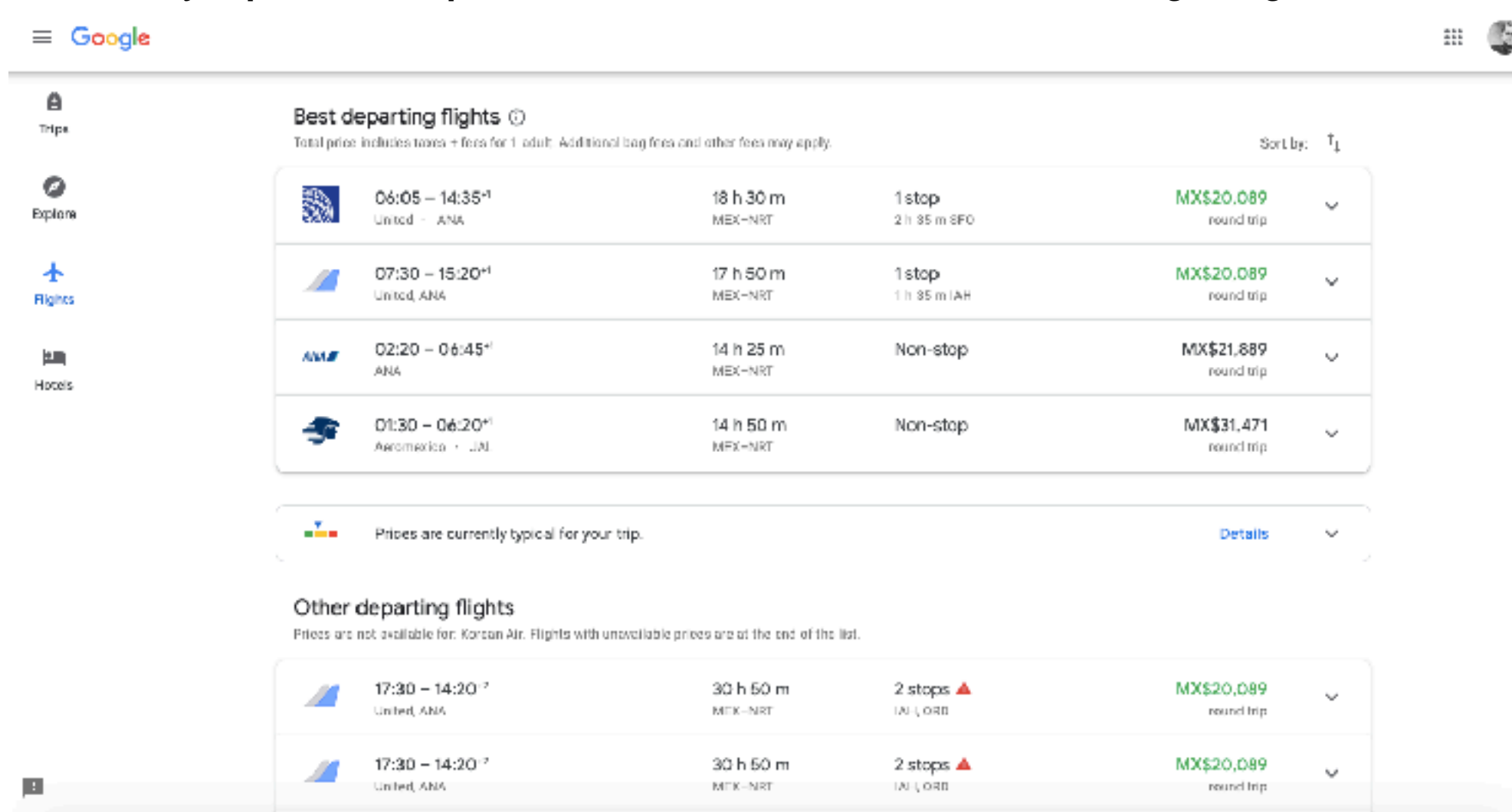
Esto no lo vamos a ver en la clase 😊

7. Recolección automática de información (Web Scrapping, Data Crawling).

`library(rvest)`

`library(xml)`

Ejemplo: Extraer precios e información de vuelos desde Google Flights



The screenshot shows the Google Flights interface for a search from Mexico City (MEX) to Tokyo (NRT). The 'Best departing flights' section lists four options. The first two are round trips for MX\$20,089, while the last two are one-way flights for MX\$31,471. The 'Other departing flights' section shows two more round trip options for MX\$20,089.

Flight	Time	Duration	Stops	Price
United - ANA	06:05 - 14:35 ⁺	18 h 30 m	1 stop 2 h 35 m SFO	MX\$20,089 round trip
United, ANA	07:30 - 15:20 ⁺	17 h 50 m	1 stop 1 h 35 m IAH	MX\$20,089 round trip
ANA	02:20 - 06:45 ⁺	14 h 25 m	Non-stop	MX\$21,889 round trip
Aeromexico - JAL	01:30 - 06:20 ⁺	14 h 50 m	Non-stop	MX\$31,471 round trip
Prices are currently typical for your trip.				
Other departing flights				
Prices are not available for: Korean Air. Flights with unavailable prices are at the end of the list.				
United, ANA	17:30 - 14:20 ⁺	30 h 50 m	2 stops 1 h 10 m ORD	MX\$20,089 round trip
United, ANA	17:30 - 14:20 ⁺	30 h 50 m	2 stops 1 h 10 m ORD	MX\$20,089 round trip

https://www.google.com/flights?lite=0#flt=MEX.NRT.2019-10-21*NRT.MEX.2019-11-05;c:MXN;e:1;sd:1;t:f

¿Qué se puede hacer con R?

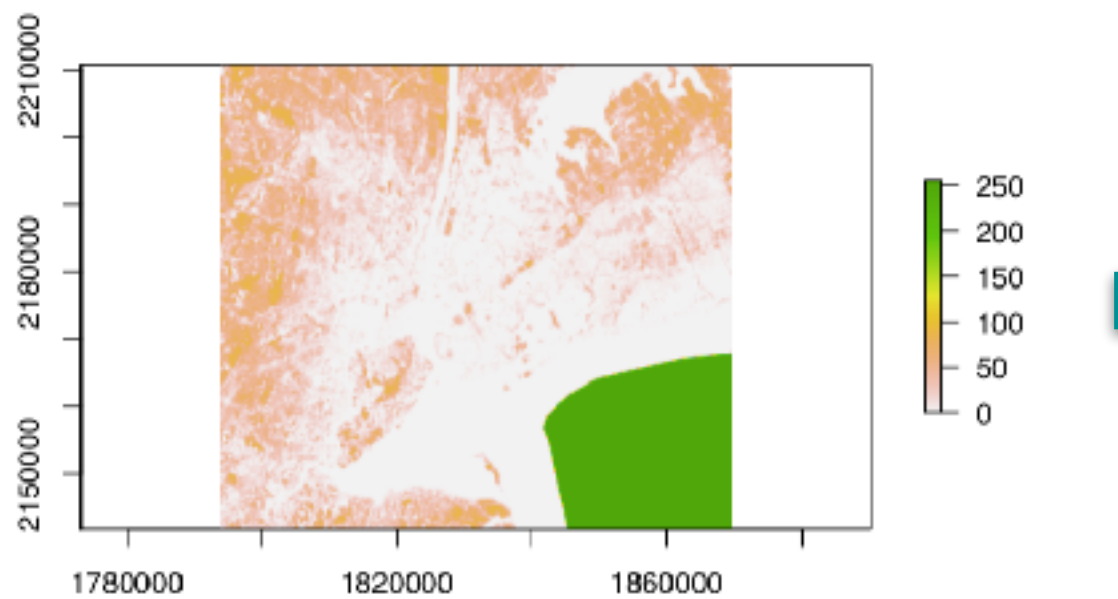
Esto lo vamos a ver en la clase 😊



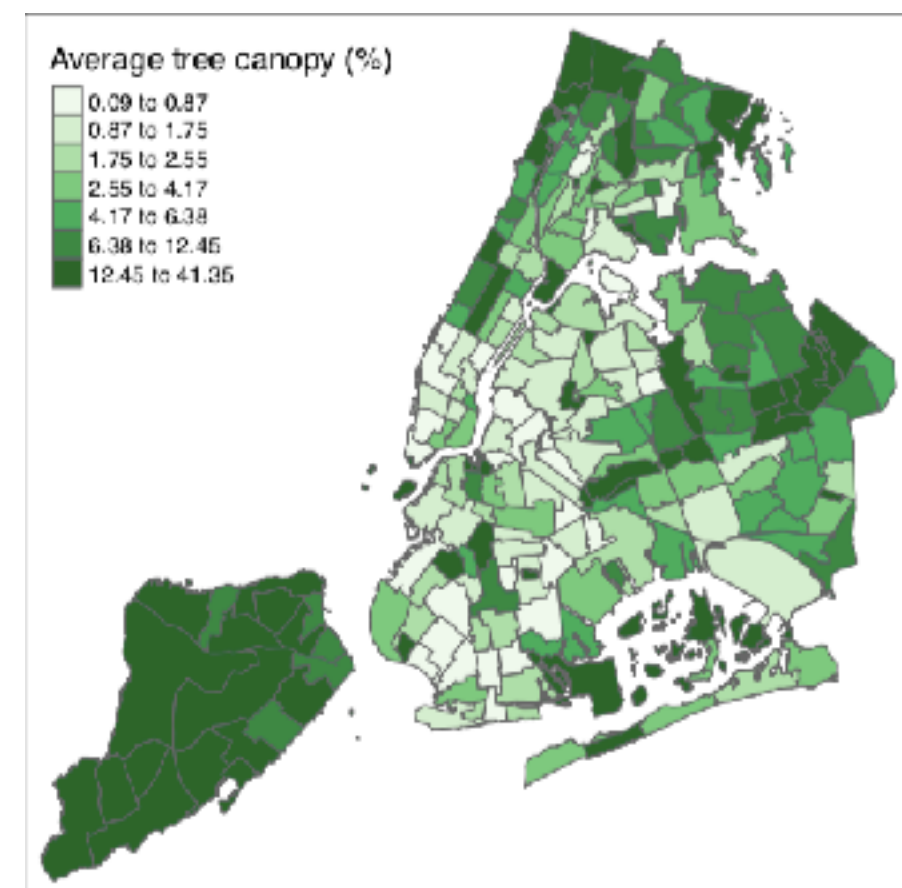
8. Análisis Geoespacial.

`library(sf)`

Abrir información geográfica, modificarla y visualizarla, así como realizar análisis a partir de esta.



**Datos Crudos
(Raw Data)**



Datos Procesados que permiten llegar a conclusiones

¿Qué se puede hacer con R?

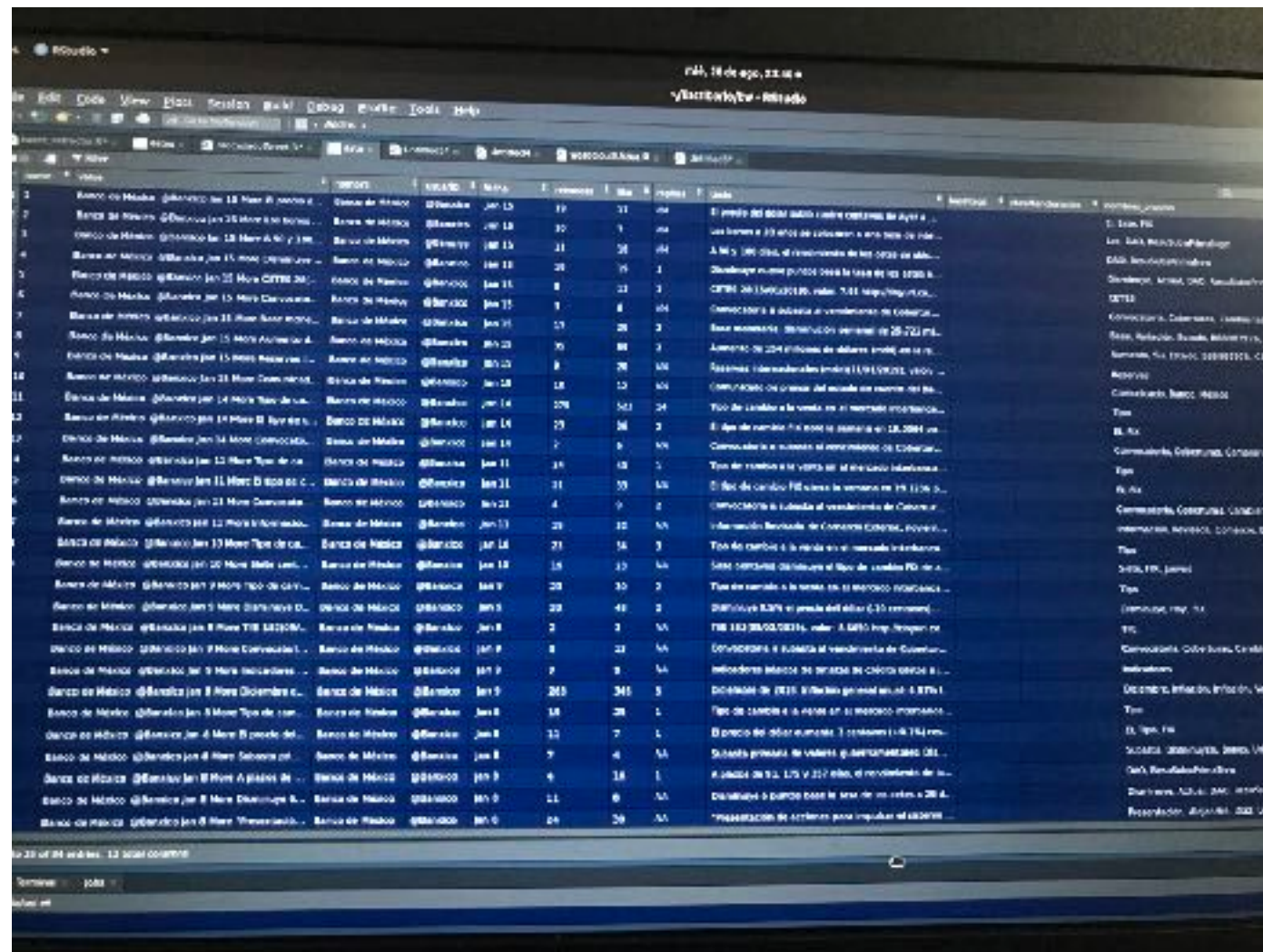


Esto no lo vamos a ver en la clase 😊

9. Automatización de tareas.

library(Rselenium)

La librería RSelenium permite programar el navegador para que replique cosas que nosotros podríamos hacer manualmente (p. ej. Descargar archivos, revisar Twitter, mandar correos, etc.).



¿Qué se puede hacer con R?



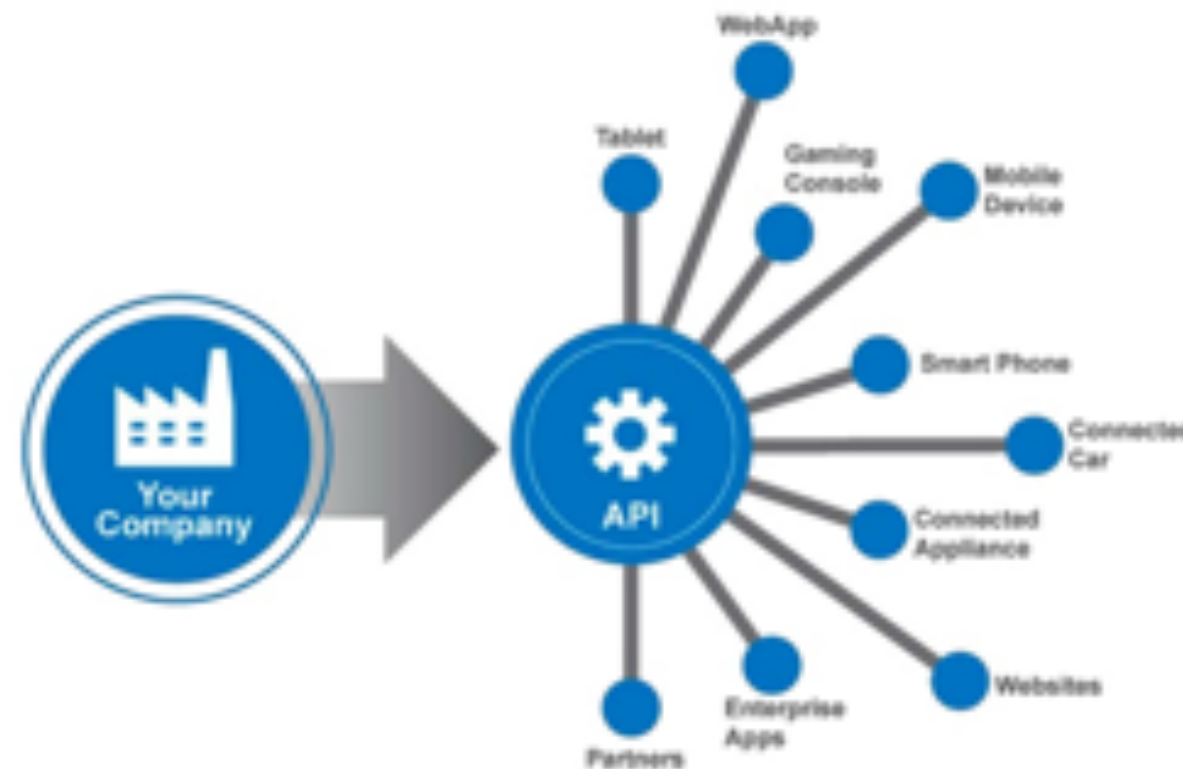
Esto no lo vamos a ver en la clase 😊

10. Creación de APIs.

library(plumber)

Definición: API (Interfaz de Programación de Aplicaciones).

Un API es un programa que permite la comunicación entre componentes de software (por ejemplo, la App de Facebook del celular se comunica con los servidores mediante la API de FB).



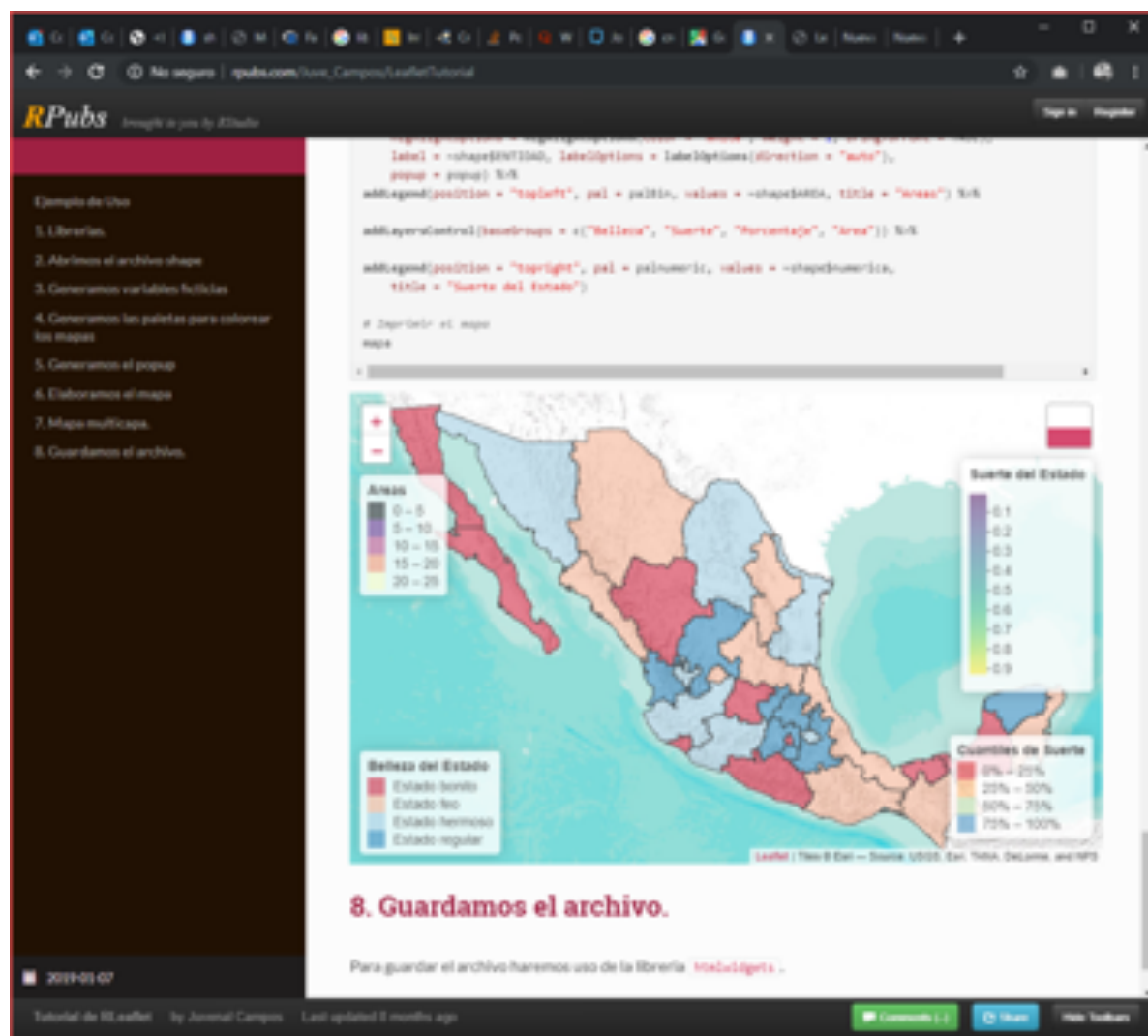
Si un día necesitamos desarrollar una API, podemos hacer esto a partir de la librería *plumber*.

¿Qué se puede hacer con R?

Esto lo vamos a ver en la clase 😊



11. Páginas web y reportes (*.pdf, *.doc, diapositivas, tableros estáticos). library(markdown)



http://rpubs.com/Juve_Campos/LeafletTutorial

L^AT_EX pdf

¿Qué se puede hacer con R?



Esto no lo vamos a ver en la clase 😊

12. Web Apps. library(shiny)



Página web de resultados de la métrica de gobierno abierto, 2019.

<https://lnppmicrositio.shinyapps.io/metricadegobiernoabierto2019/>

Sugerencias

1. Google y StackOverflow son tus aliados.

Si tienes alguna duda o surge algún error desconocido en tu código, o si quieres aprender a hacer algo muy puntual y específico, **googlear debe ser tu primera opción. La probabilidad de que alguien haya tenido la misma duda que tu es muy alta**, y es probable que la solución ya se encuentre en línea. Y si está en línea, lo más probable es que se encuentre en StackOverflow.com.

Sugerencias

2.No sufran en silencio.

Tus compañeros pueden ayudarte, al igual que tus profesores.
No acumules dudas por pena ni durante mucho tiempo, ya que los temas son acumulativos y una duda no resuelta en una clase puede hacer que no entiendas las clases posteriores.



3. Haz un perfil de GitHub

GitHub es una plataforma para programadores.

Sirve como:

- almacén de código y archivos (repositorio),
- como control de versiones,
- como plataforma de colaboración y
- como red social para programadores.

Si tu trabajo es perfectamente reproducible, **un usuario ajeno debe poder descargar el repositorio de GitHub y, siguiendo tus instrucciones, poder correrlo en su computadora personal sin mayor problema.**

Sugerencias

4. Dale estructura a tu código.

Trata de llevar un orden y estructura de tu código, de tal forma que facilite su lectura y reproducción. Recuerda incluir los comentarios necesarios para que sepas que hiciste en cada paso de tu trabajo.

Sugerencia personal:

- Primero:** Configurar el entorno de trabajo (Que se trabaje con caracteres del idioma español, UTF-8, quitar la notación científica, etc.).
- Segundo:** Sección de librerías.
- Tercero:** Sección de funciones propias.
- Cuarto.** Sección para importar las bases de datos.
- Quinto:** Exploración, Limpieza y manejo de las Bases de Datos.
- Sexto:** Análisis y estadísticas.
- Séptimo:** Impresión de resultados y guardado de la información.

Ejemplo: http://rpubs.com/Juve_Campos/estructuraCodigoR

Fin