

DATA SCIENCE PARA ECONOMÍA Y NEGOCIOS

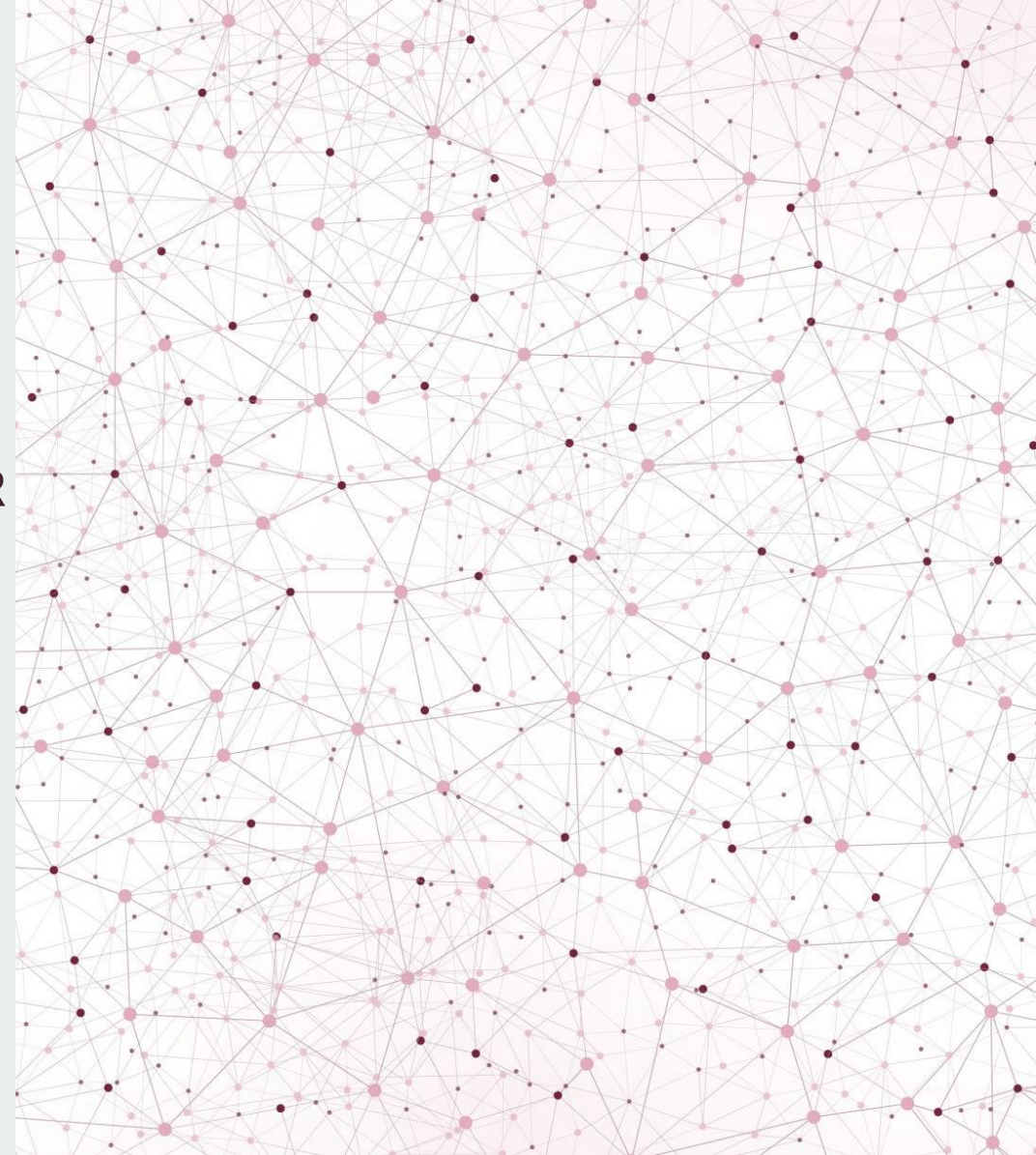
Clase 01 - Introducción al curso y a la programación en R

PROFESOR: MARTÍN SIELFELD

AYUDANTE: CATALINA HIDALGO

PERIODO: SEGUNDO SEMESTRE, 2021

SECCIÓN: 5



Orden de la presentación

Datos previos:

- Profesor: Martín Sielfeld
- Mail: martin.sielfeld.o@edu.uai.cl
- Horario de consulta: agendar vía previa solicitud por mail (siempre habrá buena disposición)
- Ayudante: Catalina Hidalgo
- Mail: cathidalgo@alumnos.uai.cl
- Horario ayudantías: por ver (3 en el semestre)

¿Qué veremos hoy?:

- ¿Qué es la ciencia de datos?
- Normas del curso y la buena convivencia
- Introducción a R y RStudio

¿Qué es y que hace la ciencia de datos?

**Transportes utilizará Big Data
para planificar políticas públicas
en el país**

**La política viral: Cómo el big data
se ha convertido en el arma más
poderosa al alcance de los
políticos**

**Las pymes deben adoptar la Inteligencia Artificial
para más ganancias**

**Big data y optimización de entrenamiento, así ayuda la tecnología
a los atletas**

01-02-2021

PROFESIONALES DE LOS DATOS

**Big Data y Data Science: los profesionales en
ciencias de datos tienen el futuro asegurado**

-
- El volumen de información en empresas y administraciones es tal que se precisan de profesionales especializados, con capacidad para su gestión
-

"The ability to take data - to be able to understand it, to process it, to extract value from, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades"

Hal Varian, Economista Jefe de Google y Académico de UC Berkeley

"El objetivo de la ciencia de datos es mejorar la toma de decisiones, basando esas decisiones en insights extraídos de grandes sets de datos"

Kelleher & Tierney, 2018

"La ciencia de datos es un paso evolutivo en campos interdisciplinarios como el análisis de negocios que incorpora la informática, el modelado, las estadísticas, la analítica y las matemáticas en uno solo proceso"

NYU center for Data Science

Definición

La ciencia de datos estudia la extracción de conocimientos a partir de datos. Este emplea las disciplinas de:

- Estadística
- Matemáticas
- Ciencias de la computación

Por otro lado, un científico de datos es quien utiliza la información disponible para generar información de valor para algún receptor. Hacer ciencia de datos implica:

- Programar (R, Python, Julia, Stata, Tableau, Matlab, entre otros softwares)
- Generar análisis estadísticos
- Comunicar
- Apoyar a la experiencia acumulada



**¿De donde se
obtienen los
datos?**

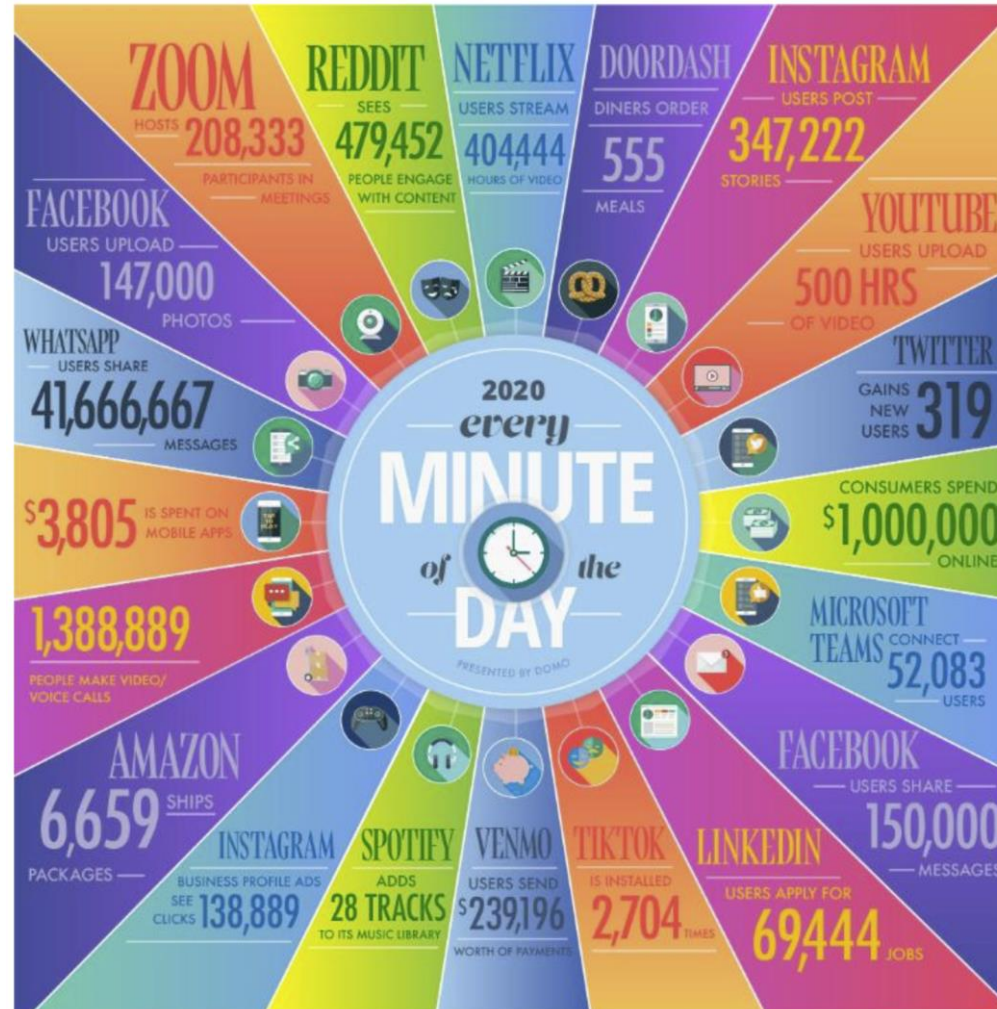


¿Qué es el Big Data?

Datos que contienen una mayor **V**ariiedad de información y que se presentan en **V**olúmenes crecientes y a una **V**elocidad superior. Las 5 "V" del Big Data corresponden a:

- Volumen
- Variedad
- Velocidad
- Veracidad
- Valor

¿Qué puede hacer el Big Data?

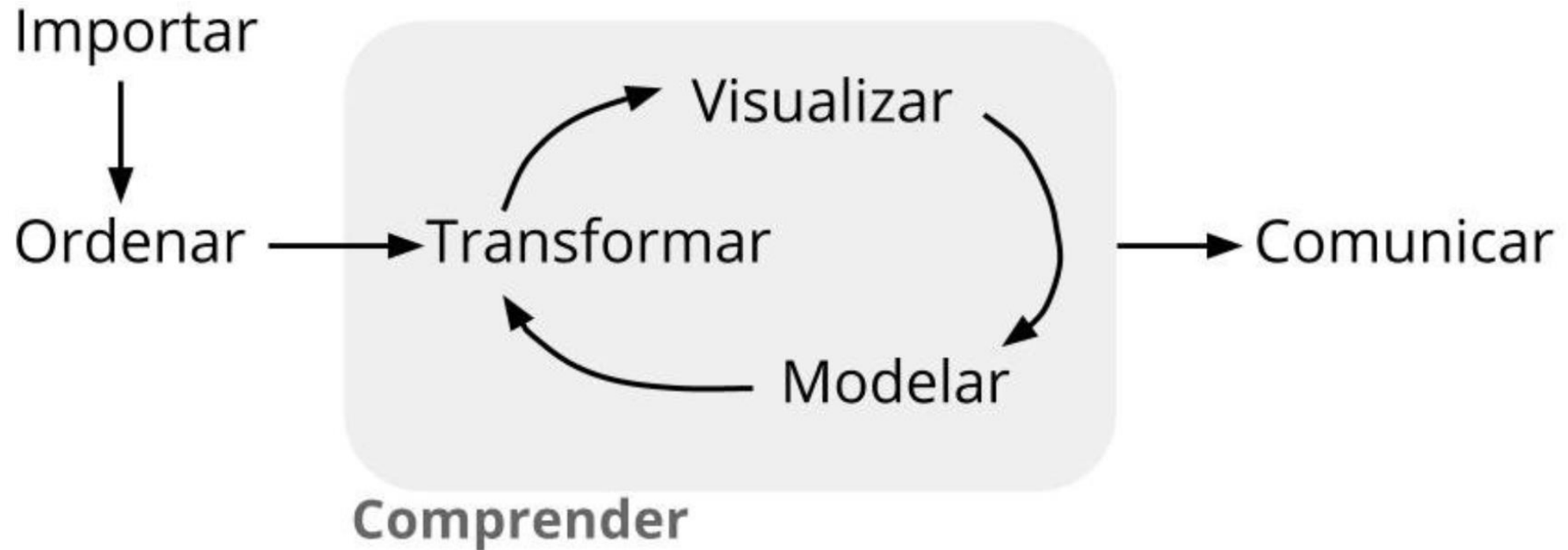


Noticia TreceBits: <https://www.trecebits.com/2020/08/16/que-ocurre-en-un-minuto-en-internet-en-2020-infografia/>

Fuente: Domo (2020). Obtenido de: <https://www.domo.com/learn/infographic/data-never-sleeps-8>

¿Cómo trabaja un científico de datos?

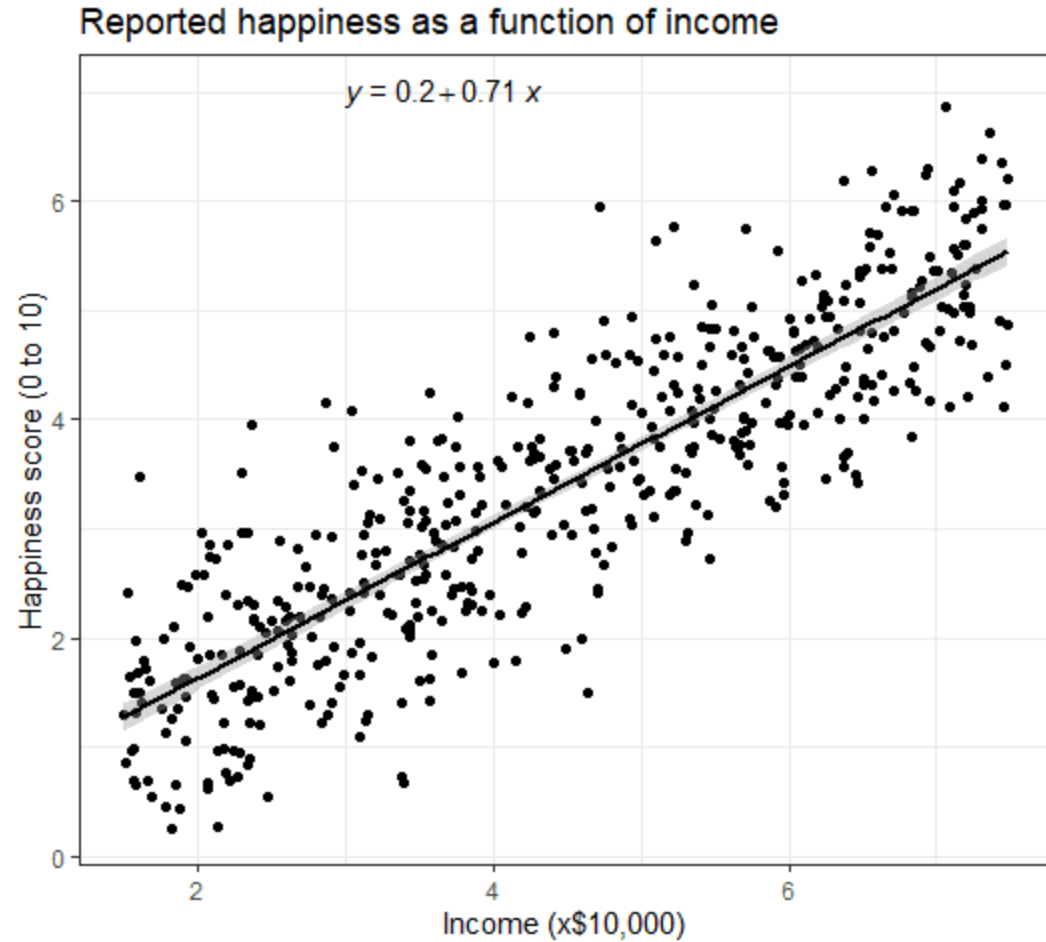
El proceso del análisis de datos



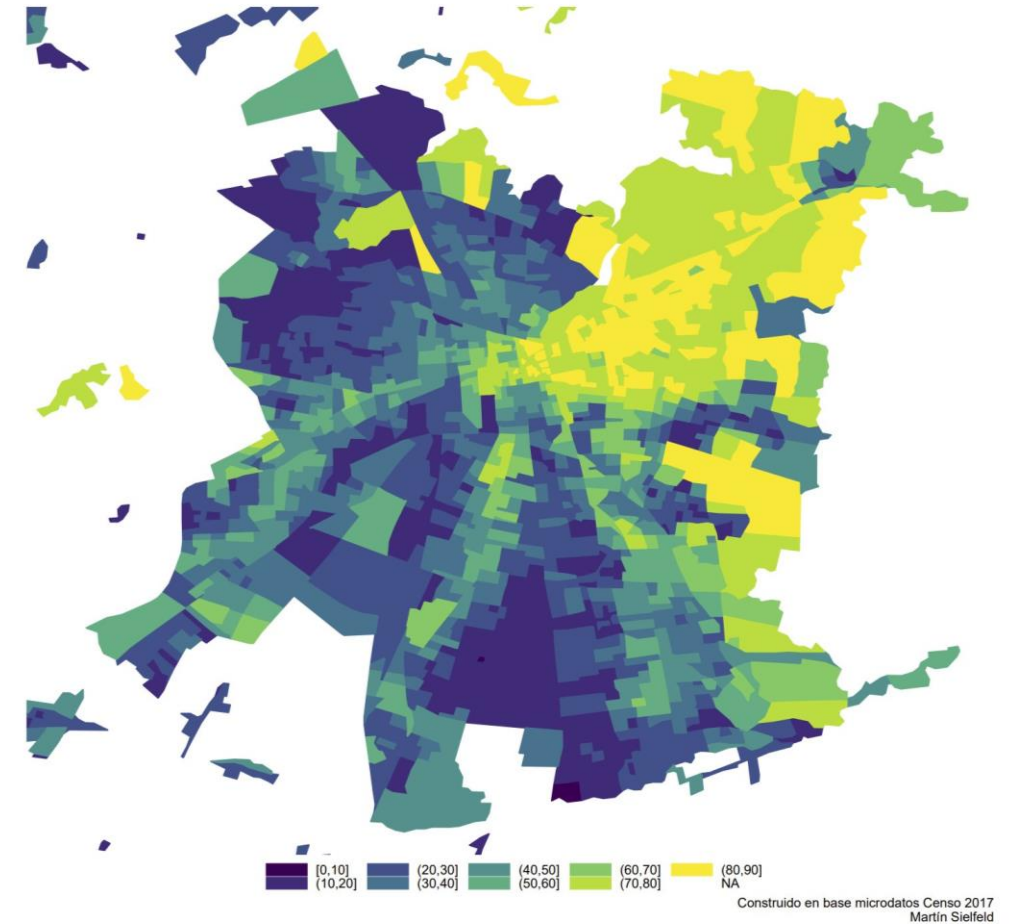
¿Qué puede hacer un científico de datos?

1. Análisis descriptivo (resúmenes estadísticos, regresiones, entre otros)
2. Análisis gráfico (imágenes uni o multidimensionales, mapeo de datos)
3. Análisis predictivo (modelos autorregresivos, machine learning, deep learning)
4. Creación de herramientas (automatización de reportes, aplicaciones shiny)
5. Y más...

¿Qué puede hacer un científico de datos?



Tasa de personas que cursaron a lo menos un año de educación superior
Zonas censales, Santiago 2017



¿Qué puede hacer un científico de datos?

Shiny from R Studio

[Back to Gallery](#) [Get Code](#)

Movie explorer

Filter

Minimum number of reviews on Rotten Tomatoes



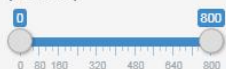
Year released



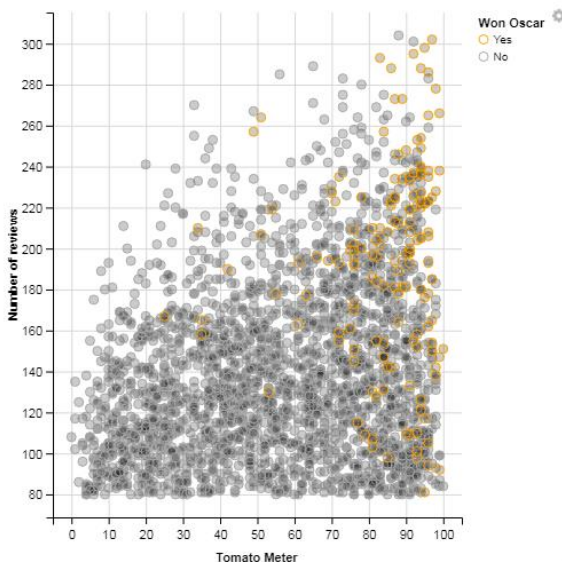
Minimum number of Oscar wins (all categories)



Dollars at Box Office (millions)

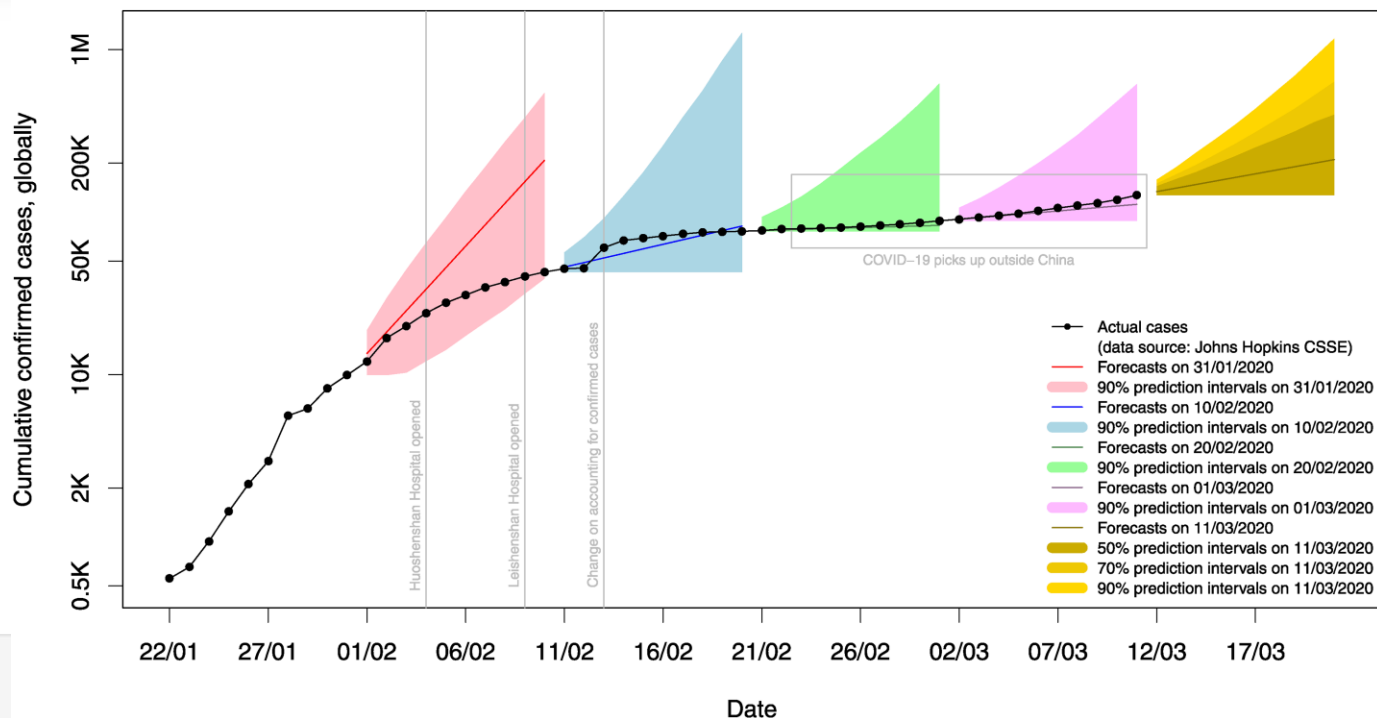


Genre (a movie can have multiple genres)



Number of movies selected:
2557

Forecasting confirmed COVID-19 cases



¿Es fácil?

Google Flu Trends:

En el año 2009 surgió la gripe A (H1N1), lo cual generó un problema de salud pública a nivel mundial. El monitoreo en EE.UU estaba a cargo del “Centro para el control de enfermedades y Prevención de enfermedades” (CDC). Este proceso tomaba alrededor de **10 días**.

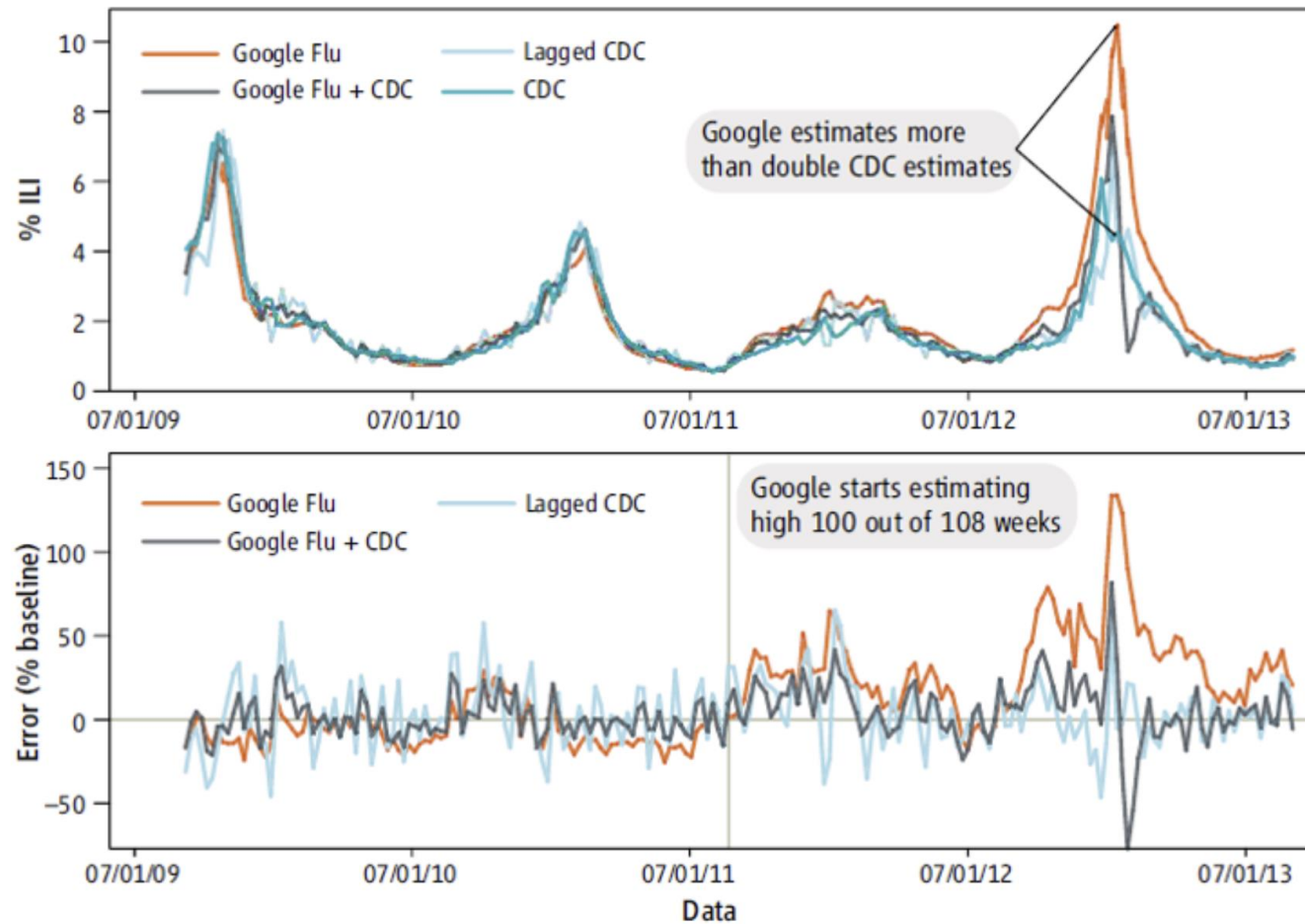
Google ofreció un método para predecir la intensidad de la gripe en tiempo real y con posición geográfica. Entre otros aspectos, dicho modelo de predicción contemplaba:

- Datos históricos
- Datos de la intensidad de búsqueda

$$Intendencia_gripe = f(intendencia_busqueda, datos_históricos, otros)$$

¿Es fácil?

La decadencia de Google Flu Trends:



¿Es fácil?

Aprendizajes:

Ambiciones estratégicas y modelamiento de Big Data no siempre solucionarán todo:

- Más no siempre es preferido a menos.
- Es importante encontrar un correlato de los datos con la realidad.
- Tener cuidado con incluir “relaciones espurias” (relacionar sin elación real).

Pero, Google Flu Trends si resolvió un problema en un comienzo.

¿Qué veremos en el curso?

Introducción a la programación en R:

- Aprender a programar en R (mediante Rstudio).
- Introducción a paquetes para el análisis de datos (data.table y lenguaje base).
- Introducción a la manipulación de datos (limpieza, manipulación y análisis estadísticos de bases de datos)

Visualización de datos:

- Análisis en forma visual para generar contexto, detectar problemas y generar hipótesis de estudio.
- Generar visualizaciones potentes que permitan captar la atención y mostrar dinámicas interesantes.

Técnicas de análisis de datos:

- Introducción al Machine Learning.
- Análisis de regresión.
- Modelos de clasificación (análisis de clusters, arboles de decisión).

Normas del curso

Orden del curso

Unidad 01: Introducción a la codificación y al manejo de bases de datos

- Aprender a programar con lenguaje data.table y limpieza de datos.
- Análisis estadístico básico.
- Publicación de trabajos en Rpubs mediante archivos Rmarkdown.

Unidad 02: Visualización de datos:

- Introducción a los datos estáticos mediante paquete ggplot2.
- Introducción a los gráficos dinámicos y mapeo mediante paquete plotly y leaflet.

Unidad 03: Técnicas de análisis de datos:

- Introducción al Machine Learning.
- Análisis de regresión.
- Modelos de clasificación (análisis de clusters, arboles de decisión).
- Presentaciones de científicos de datos (economía, negocios, finanzas, entre otros posibles invitados).

Evaluaciones

Evaluaciones	Ponderación	Fechas Tentativas
Control 1	15%	10 de Septiembre
Control 2	10%	08 de Octubre
Control 3	15%	19 de Noviembre
Entrega de Productos (5 a 6)	20%	No Aplica
Proyecto Final - Primera Entrega	10%	Por Definir
Proyecto Final - Informe y Presentación	30%	29 de Noviembre a 13 de Diciembre

- Los controles son en horario de clase e individuales.
- Los productos son en grupos de **máximo** 3 integrantes.
- El alumno solo puede justificar un control o bien dos productos. Justificar un control implica que la ponderación de este se sume a la del proyecto final (40%). Justificar un producto le quita el derecho a borrar la peor nota de los productos. Justificar dos productos implica no poder borrar la peor nota de los productos y que a la ponderación del proyecto final se le sume la ponderación de uno de estos productos.

Proyecto final

Objetivos:

1. Entender la estructura de las bases de datos y el cómo el paradigma de la ciencia de datos permite ayudarnos a responder preguntas aplicadas a fenómenos de la economía y negocios.
2. Desarrollar la capacidad del estudiante a entender la información contenida dentro de una base de datos, comprendiendo las limitantes que pueden tener estos para entender un problema.
3. Entregarle al estudiante múltiples herramientas para presentar información relevante para comprender un fenómeno, tanto a nivel de visualizaciones como de métodos estadísticos clásicos.
4. Preparar un informe escrito y una presentación oral.

Procedimiento:

La presentación final y el reporte consistirán en los siguientes ítems:

1. Introducción y descripción general de la base de datos elegida por el grupo utilizando estadísticos y visualizaciones enseñadas durante el curso.
2. Empleo de estrategia enseñada en el curso para responder a la problemática tratada en el Proyecto.
3. Desarrollo de una sección donde se describan los principales resultados encontrados en el análisis de datos, explicando los mecanismos y los limitantes de estos.

Otros

Ayudantías:

- El curso cuenta con 3 ayudantías, las cuales serán coordinadas entre la disponibilidad de la ayudante y las de los alumnos.
- Se efectuará una ayudantía previo a cada control, con el objetivo de repasar **parte** de la materia que será evaluada en el control.

Problemas e inasistencias:

- Las asistencias deben ser justificadas en su debida forma ante la secretaria de pregrado.
- Ante cualquier eventualidad, puede contactarse con el profesor para indicarle su problema. Cualquier problema es solucionable, en la medida que se indique con la debida anticipación. Se espera que sea el alumno quien contacte al profesor (me temo que no soy adivino).

Introducción a R y RStudio



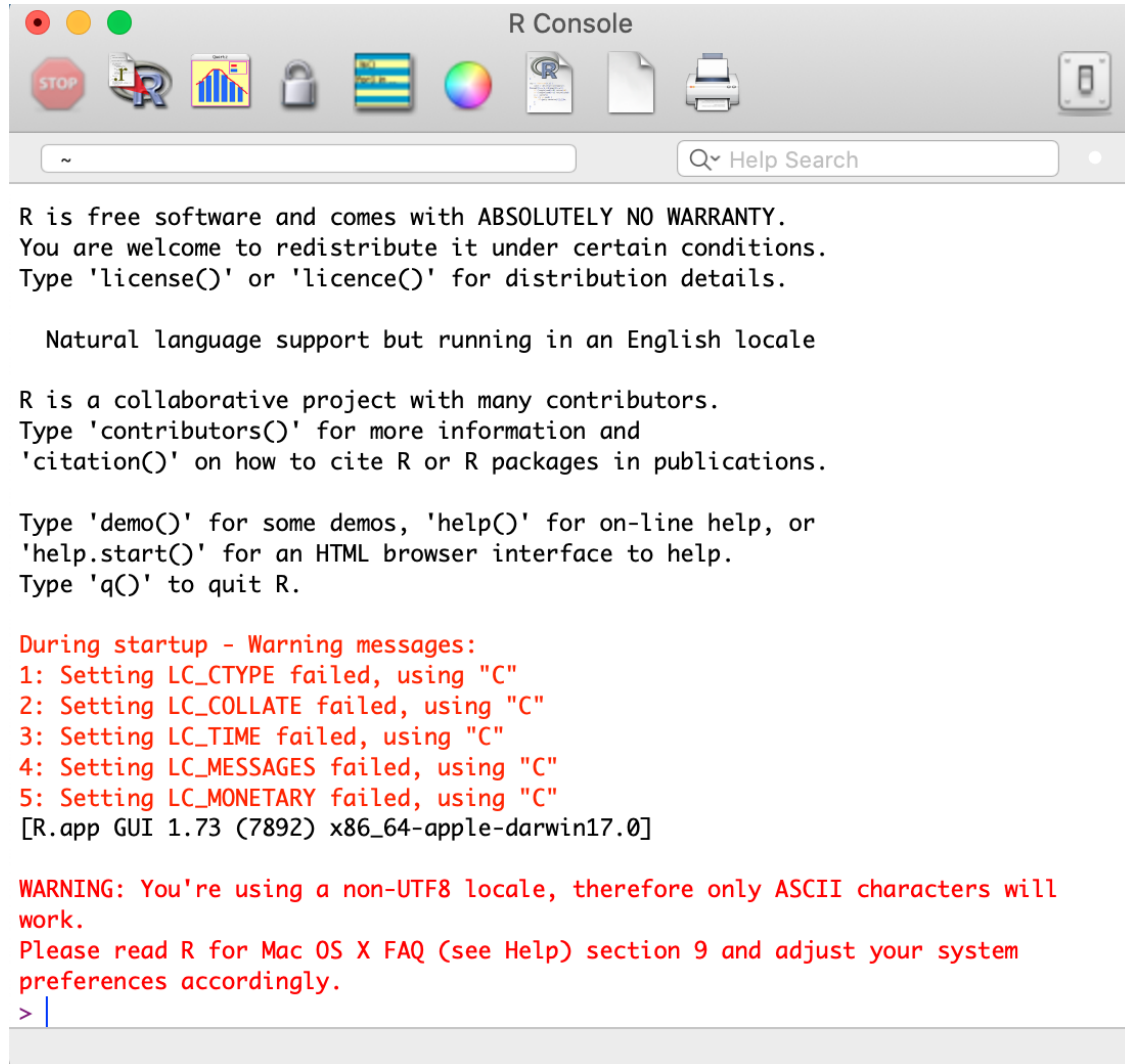
¿Qué es R?

R es un software libre de programación, el cual sirve para la computación estadística y la generación de información gráfica. Se caracteriza por ser:

- Abierto
- Gratuito
- Independiente
- Reproducible

Aprender R es como aprender un nuevo idioma, por lo que hay que partir por lo básico.

¿Qué es R?



```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

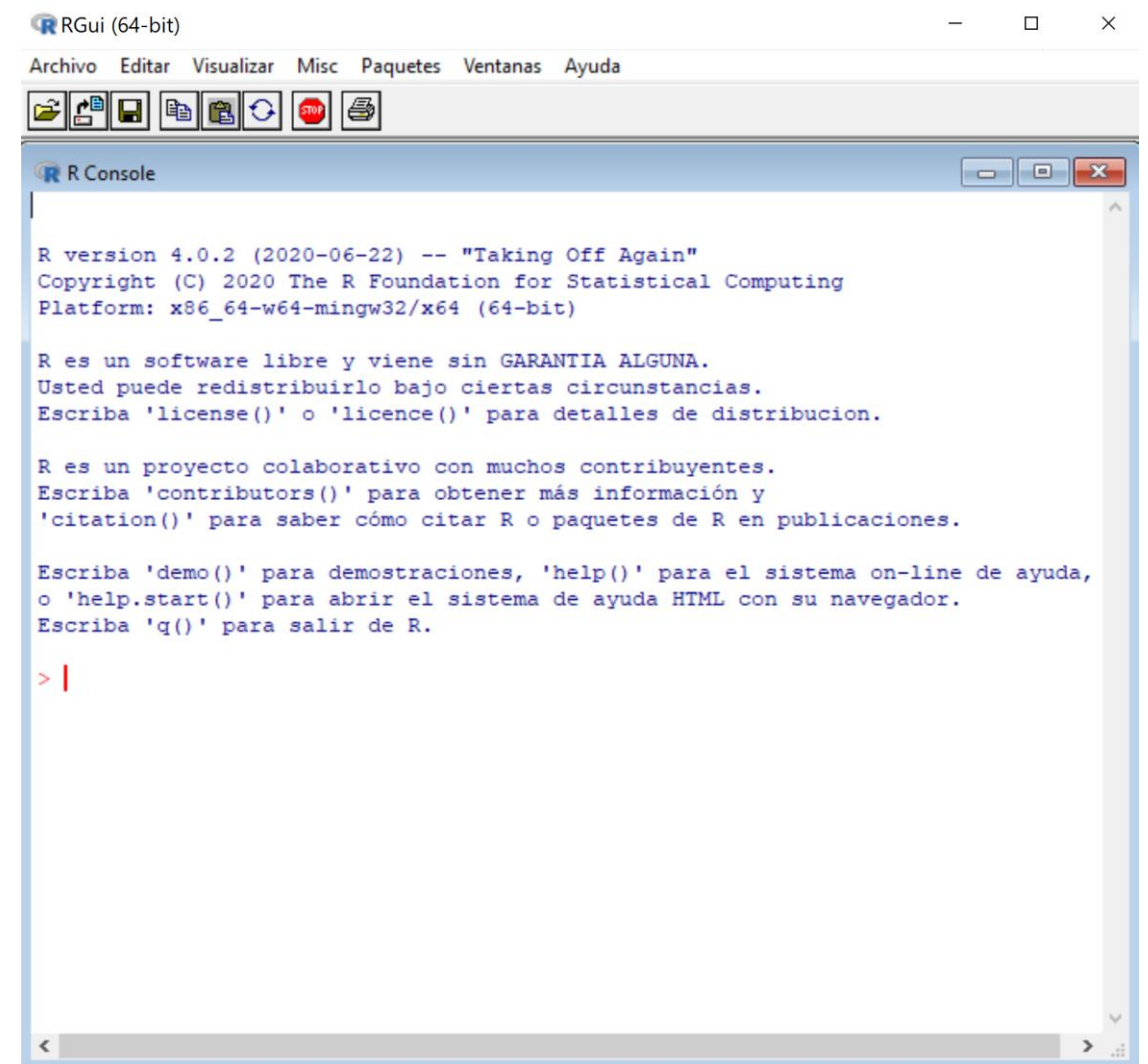
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

During startup - Warning messages:
1: Setting LC_CTYPE failed, using "C"
2: Setting LC_COLLATE failed, using "C"
3: Setting LC_TIME failed, using "C"
4: Setting LC_MESSAGES failed, using "C"
5: Setting LC_MONETARY failed, using "C"
[R.app GUI 1.73 (7892) x86_64-apple-darwin17.0]

WARNING: You're using a non-UTF8 locale, therefore only ASCII characters will
work.
Please read R for Mac OS X FAQ (see Help) section 9 and adjust your system
preferences accordingly.
> |
```



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

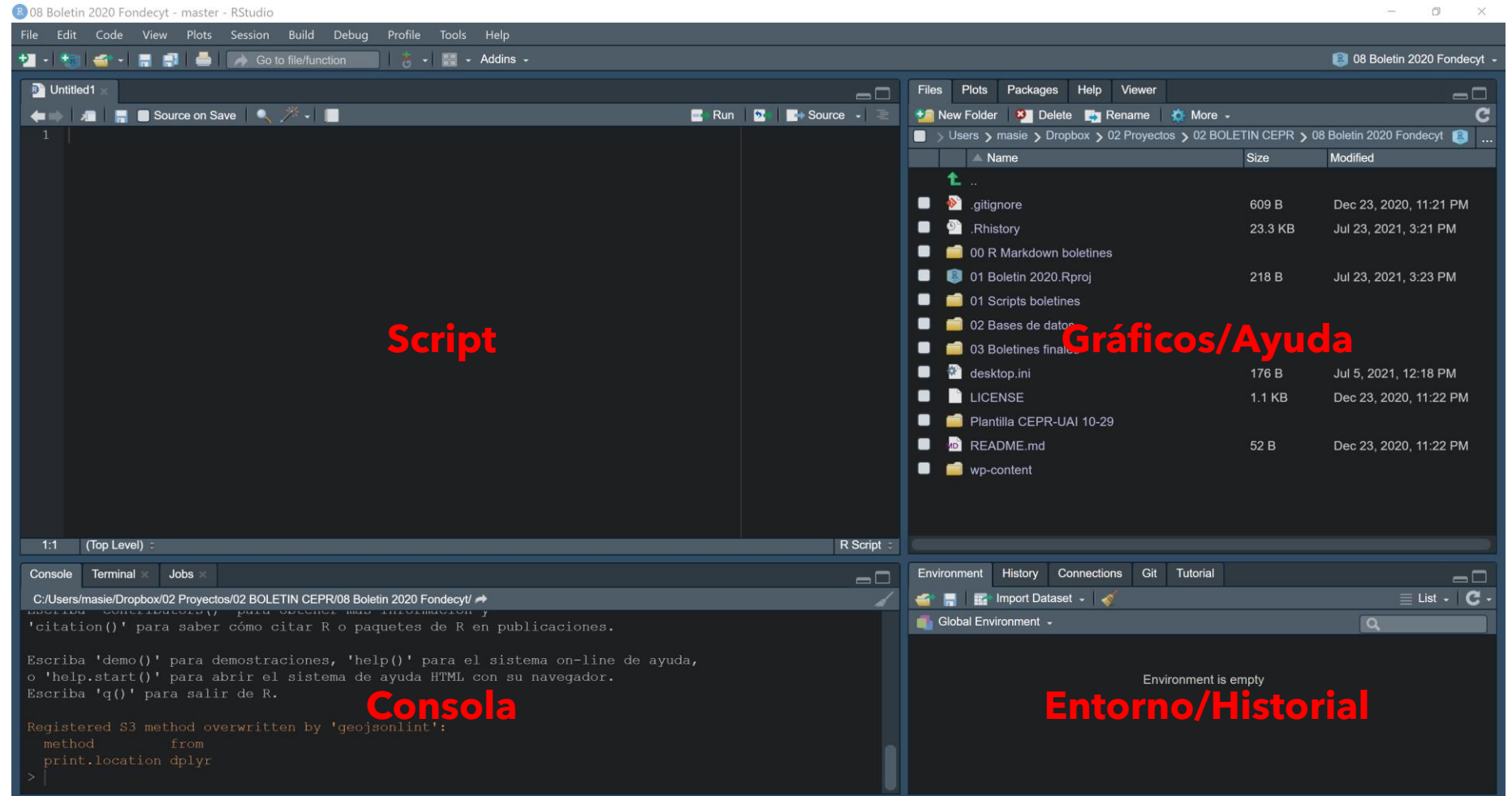
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

> |
```

¿Qué es RStudio?

Es un software IDE de R



Programación basada en objetos

Reunir datos e información y guardarlos en un objeto:

- Vectores de valores (numéricos o caracteres)
- Matrices
- Data frames (bases de datos de filas de observaciones y columnas de variables)

Utilizar estos objetos para extraer información mediante funciones:

- `sum()`
- `mean()`
- `summarise()`
- `quantile()`
- `weighted.mean()`
- Entre otras

Lenguaje - Sintaxis 1: Comandos

¿Cómo se escribe en R?: se utilizan comandos, los cuales pueden ser ejecutados desde la consola o mediante un script.

- Para crear o “guardar” un objeto, se utiliza el símbolo “<-” o bien “=”.
- Ctrl + Enter permiten correr la línea de código en donde se encuentra el cursor.
- Ctrl + R permite correr la totalidad del script.

Lenguaje - Sintaxis 2: Scripts

Apartado donde se pueden guardar líneas de códigos, las cuales pueden ser ejecutables. El orden del código dentro del script debe ser cronológico (o secuencial).

En el script pueden haber:

- Objetos
- Funciones
- Paquetes/librerías

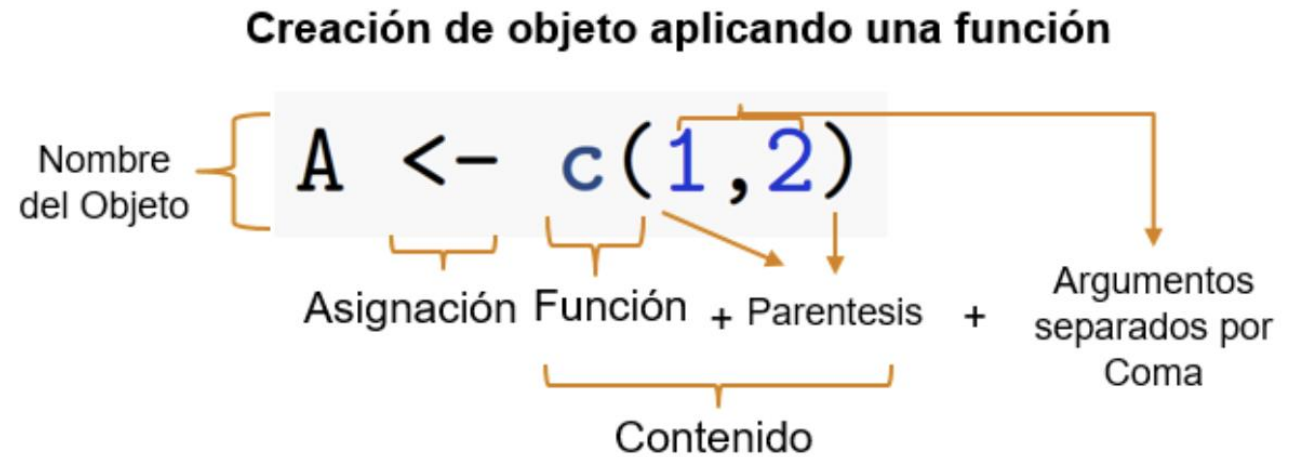
Es difícil ser organizado al codificar, por lo cual se recomienda que en el script se escriba un comando por línea.

Lenguaje - Sintaxis 3: Funciones

Es un conjunto de comandos compactados en un objeto. Suelen estar compuestos por.

- Nombre de la función
- Paréntesis
- Argumentos

Por ejemplo, en el caso de la función "concatenar":



En caso de querer saber los argumentos de una función, puede ejecutar Tab dentro de los paréntesis de la función, o bien ejecutando en la consola ?nombre_de_la_función (ej: "?c").

Lenguaje - Sintaxis 4: Objetos

Existen muchos tipos de objetos:

- Numéricos y Strings (caracteres o factores)
- Listas
- Matrices
- Secuencias
- Data frames

Un objeto puede poseer distintas dimensiones, dependiendo de la forma en la cual estén organizados los elementos dentro del objeto:

- Unidimensionales (vectores)
- Bidimensionales (matrices y data frames)
- Tridimensionales (arrays)
- N-dimensiones (listas)

Lenguaje - Sintaxis 5: Indexing

¿Cómo busco un elemento específico dentro de un objeto?:

- Por posición (manual)
- Por condición (R busca la posición del elemento que cumple la condición)

Indexar requiere el uso de corchetes "[]" justo después del nombre del objeto. Si el objeto tiene más de una dimensión, se añade una "," para separar cada dimensión.

Objeto unidimensional

```
A<- c(3:20)  
A #entrega todos los elementos de A
```

```
## [1] 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```


Lenguaje - Sintaxis 5: Indexing

```
A[4] # El cuarto elemento de A
```

```
## [1] 6
```

```
A[-1] # Todos los elemento menos el primero
```

```
## [1] 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```

```
A[5:10] # Desde el quinto al decimo tercer elemento
```

```
## [1] 7 8 9 10 11 12
```

```
A[length(A)] # Último elemento
```

```
## [1] 20
```

Lenguaje - Sintaxis 5: Indexing

Objeto multidimensional

```
X<-matrix(12:15,2,2)  
X #entrega todos los elementos de X
```

```
##      [,1] [,2]  
## [1,]  12  14  
## [2,]  13  15
```

```
X[1,2] #entrega el elemento que se encuentra en la primera fila, en la segunda col
```



```
## [1] 14
```

Lenguaje - Sintaxis 6: Vectores lógicos

Por condición:

- Ejemplo: ¿De un conjunto de datos, cuales son los menores a 5?
- Indexando objetos unidimensionales usando operaciones lógicas

```
z<-c(10,2.1,1.55,5,9)  
z
```

```
## [1] 10.00  2.10  1.55  5.00  9.00
```

```
z<5 # Cada elemento dentro de z que si cumple la condición
```

```
## [1] FALSE  TRUE  TRUE FALSE FALSE
```

```
z[z<5] # ¿Cuales son?
```

```
## [1] 2.10 1.55
```

Lenguaje - Sintaxis 7: Manipulación de objetos

Funciones para recordar:

- `class(x)`: revela el tipo de objeto e indirectamente dice que se puede hacer con él
- `str(x)`: permite ver las clases de los atributos contenidos en un data frame
- `length(x)`: revela la longitud de un objeto unidimensional
- `dim(X)`: revela la dimensión de un objeto n-dimensional
- `names(x)`: revela los nombres (si existen) de los elementos de un objeto
- `attributes(x)`: revela todos los atributos de un objeto
- `rm(x)`: elimina un objeto de la memoria de R, en una sesión
- `ls(x)`: revela todos los objetos en la memoria de R en una sesión específica

Referencias del curso

Teoría:

- Big Data – Walter Sosa (BDWS). Editorial Siglo Veintiuno.
- Storytelling with Data: A Data visualization Guide for Business Professionals – Cole Nussbaumer

Programación:

- R para Ciencia de Datos (R4DS) – link: <https://es.r4ds.hadley.nz/>