

The data scientist's toolbox

¿Porque videos automatizados?

Creamos muchos cursos abiertos masivos en línea en el Laboratorio de ciencia de datos de Johns Hopkins. Hemos creado más de 30 cursos en múltiples plataformas durante los últimos 5 años. Nuestro objetivo con estas clases es proporcionar la mejor y más actualizada información a la audiencia más amplia posible. Pero existen importantes desafíos para mantener tanto material en línea. Los paquetes R quedan obsoletos, se inventan nuevos flujos de trabajo y errores tipográficos, joh, los errores tipográficos!

El resultado es que muchas de nuestras conferencias han quedado desactualizadas, incluyen errores o no incluyen las mejores y más recientes versiones de flujos de trabajo y canalizaciones

Entonces, cuando nos sentamos a desarrollar nuestro nuevo proceso para crear y mantener nuestros cursos, queríamos ver si podíamos descubrir cómo hacer una clase hecha completamente de documentos de texto sin formato. Desglosamos un curso abierto en línea masivo en sus elementos básicos:

- Tutoriales: podemos escribirlos fácilmente en formatos de texto sin formato como Markdown o R Markdown
- Diapositivas: son bastante fáciles de mantener y compartir si las hacemos con algo como Google Slides
- Evaluaciones: aquí podemos usar un lenguaje de marcado para crear cuestionarios y otras evaluaciones
- Videos: este era el punto conflictivo, ¿cómo íbamos a hacer videos a partir de documentos de texto sin formato?

Por una feliz coincidencia, las comunidades de ciencia de datos e inteligencia artificial nos estaban resolviendo una gran parte de este problema, ¡mejorando la síntesis de texto a voz! ¡Así que ahora podríamos escribir un guión para un video y usar **Amazon Polly** para sintetizar nuestras voces!

Para aprovechar esta nueva tecnología, creamos dos nuevos paquetes R: ari y didactr.

Ari tomará un guion y un conjunto de diapositivas de Google y narrará el guion sobre las diapositivas utilizando Amazon Polly. También generará el archivo de subtítulos necesario para incluir subtítulos y garantizar que los videos sean accesibles para las personas con discapacidad auditiva. Didacta automatiza varios de los pasos desde la creación de los videos con ari hasta su carga en YouTube, de modo que podamos editar rápidamente los guiones o diapositivas, rehacer los videos, volver a cargarlos y reducir nuestra sobrecarga de mantenimiento para mantener nuestro contenido actualizado.

Siempre que cambiemos el archivo de texto o editamos las diapositivas, podemos recrear el video en un par de minutos. Todo se hace en R. Una de las características más interesantes de ir a este nuevo proceso es mostrarle lo poderoso que es el lenguaje de programación R.



¿Que es ciencia de datos?

La ciencia de datos es utilizar datos para contestar preguntas. Es un campo muy amplio. La ciencia de datos puede involucrar:

- Estadística, informática, matemáticas
- Limpieza y formateo de datos
- Visualización de datos

Un informe especial de "The Economist" resume bien las cualidades necesarias. Indica que un científico de datos se define a grandes rasgos como alguien que combina las habilidades del programador de software, el estadístico y el contador de historias más el artista, para extraer fragmentos de oro ocultos bajo montañas de datos.

Una de las razones del auge de la ciencia de datos en los últimos años es la gran cantidad de datos disponibles en la actualidad y los que se están generando. No sólo enormes cantidades de datos que se recogen acerca de muchos aspectos del mundo y de nuestras vidas, sino que al mismo tiempo tenemos el incremento de la computación de bajo costo. Esto ha creado una tormenta perfecta en la que creamos más datos y las herramientas para analizarlos, mejorando la capacidad de memoria de los computadores, mejores procesadores, más software y ahora también más científicos de datos calificados que pueden ponerlo en práctica y responder preguntas usando estos datos.

¿Que es el big data?

Big Data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día. Pero no es la cantidad de datos lo que es importante. Lo que importa con el Big Data es lo que las organizaciones hacen con los datos. Big Data se puede analizar para obtener ideas que conduzcan a mejores decisiones y movimientos de negocios estratégicos.

Ha sido parte integral del auge de la ciencia de datos. Hay algunas cualidades que caracterizan a los macrodatos.

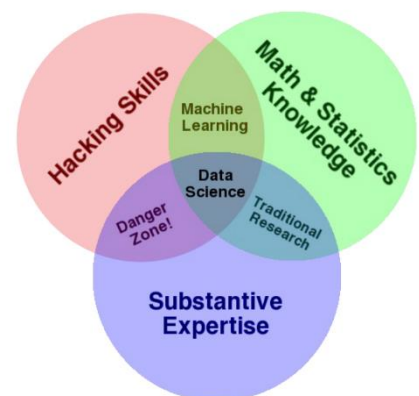
- **Volumen:** Como su nombre lo indica, los macrodatos implican grandes conjuntos de datos, y estos grandes conjuntos de datos se están volviendo cada vez más rutinarios.
- **Velocidad:** Los datos se generan y recopilan más rápido que nunca.
- **Variedad:** Tiene diferentes tipos de datos disponibles.

¿Que es un científico de datos?

Es alguien que utiliza datos para contestar preguntas.

*Que destrezas necesita
un científico de datos?*

Resumen de lo ue se va a enseñar



Que son los datos?

Wikipedia dice que los datos Es un conjunto de valores de variables cualitativas o cuantitativas.

Cambridge English Dictionary establece que los datos son información, especialmente los hechos o números recogidos para ser examinados y considerados y utilizados para ayudar a la toma de decisiones.

Los datos se examinan recopilan y lo más importante se usa para informar las decisiones.

Tipos más comunes de datos confusos

Estas son solo algunas de las fuentes de datos que puede encontrar y veremos brevemente cómo se ven a menudo algunos de estos conjuntos de datos o cómo se pueden interpretar, pero una cosa que tienen en común es el desorden de los datos: usted Tiene que trabajar para extraer la información que necesitas para responder a tu pregunta.

- Secuencia de datos
- Datos del censo de población
- Registros médicos electrónicos (EMR), otras grandes bases de datos
- Datos del sistema de información geográfica (GIS) (cartografía)
- Análisis de imágenes y extrapolación de imágenes
- Idiomas y traducciones
- Tráfico del sitio web
- Datos personales / publicitarios (por ejemplo: Facebook, predicciones de Netflix, etc.)

Un buen científico de datos hace preguntas primero y busca datos relevantes en 2do lugar.

En primer lugar, analizamos dos definiciones de datos, una que se centra en las acciones que rodean a los datos y otra en lo que comprenden los datos. La segunda definición incorpora los conceptos de poblaciones, variables y analiza las diferencias entre datos cuantitativos y cualitativos. En segundo lugar, examinamos diferentes



fuentes de datos que puede encontrar y enfatizamos la falta de conjuntos de datos ordenados. Los ejemplos de conjuntos de datos desordenados, donde los datos sin procesar deben agruparse en una forma interpretable, pueden incluir datos de secuenciación, datos del censo, registros médicos electrónicos, etc. Y finalmente, volvemos a nuestras creencias sobre la relación entre los datos y su pregunta y enfatizamos la importancia de las estrategias de pregunta primero. Puede tener todos los datos que pueda desear, pero si no tiene una pregunta para comenzar, los datos son inútiles.

Getting Help

- Stack Overflow
- Cross Validated
- Foro de coursera



Detalles a incluir en preguntas en foros:
la pregunta que estás tratando de responder

Asking questions on forums - details to include:

- The question you are trying to answer
- How you approached the problem, what steps you took to answer the question
- What steps will reproduce the problem (including sample data for troubleshooters to work from!)
- What was the expected output
- What you saw instead (including any error messages you received!)
- What troubleshooting steps you have already tried
- Details about your set-up, eg: what operating system you are using, what version of the product you have installed (eg: R, Rpackages)

Titling forum posts

Bad:

- HELP! Can't fit linear model!
- HELP! Don't understand PCA!

Better:

- R 3.4.3 lm() function produces seg fault with large data frame (Windows 10)
- Applied PCA to a matrix - what are U, D, and Vt?

Even better:

- R 3.4.3 lm() function on Windows 10 – seg fault on large dataframe
- Using principal components to discover common variation in rows of a matrix, should I use U, D or Vt?

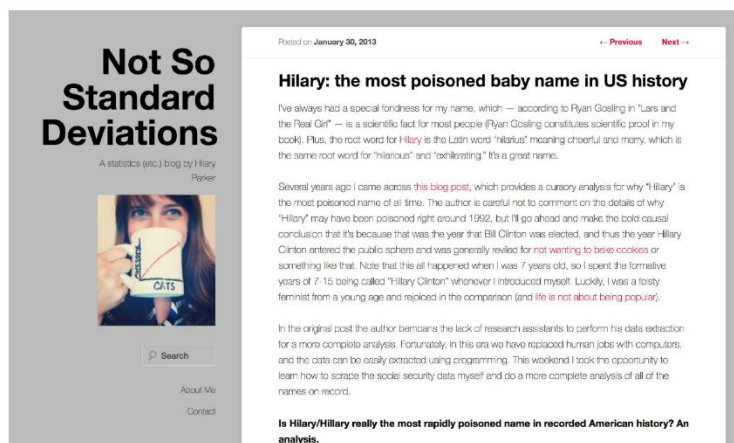
Pregunta 1: Which of these might be a good title for a forum post?
Removing rows with NAs in dataframe using subset(), R 3.4.3

Proceso de la ciencia de datos

Cada proyecto de ciencia de datos comienza con una pregunta que debe responderse con datos.

- 1er paso: Formular la pregunta
- 2do paso: encontrar o generar los datos que vamos a utilizar para responder la pregunta
- 3er paso: Analizar los datos, explorándolos y modelándolos (usar algunas técnicas estadísticas o de aprendizaje automático para analizar los datos)
- 4to paso: Sacar conclusiones del analisis
- 5to paso: Comunicarlo con otros colegas o equipo de trabajo, un proyecto de ciencia de datos casi siempre implica alguna forma de comunicación de los hallazgos del proyecto

Ejemplo: <https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/>



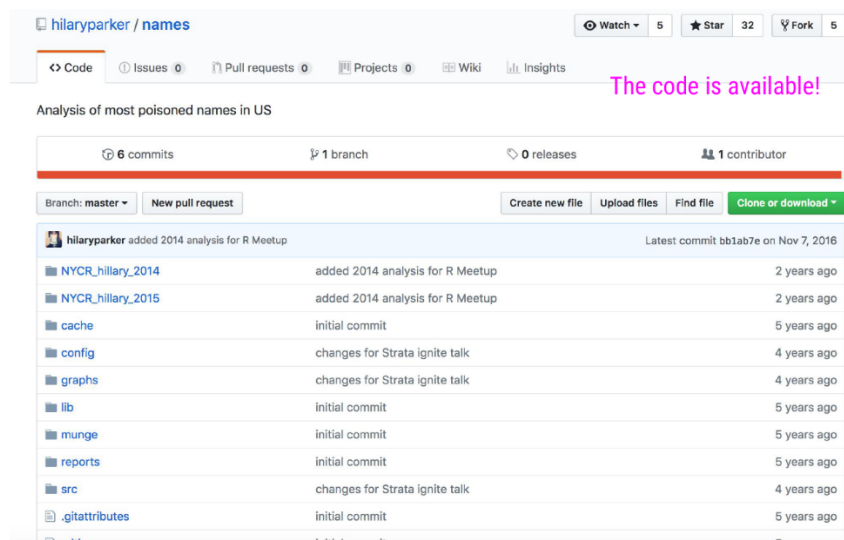
Hilary's post blog

<https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/>

1: Pregunta: Is Hilary/Hillary really the most rapidly poisoned name in recorded American history?
¿Es Hilary / Hillary realmente el nombre envenenado más rápidamente en la historia estadounidense registrada?

2. Los datos: Para responder a esta pregunta, Hilary recopiló datos del sitio web del Seguro Social. Este conjunto de datos incluyó los 1000 nombres de bebés más populares desde 1880 hasta 2011.

3. Análisis de los datos: Hilary estaba interesada en calcular el riesgo relativo para cada uno de los 4110 nombres diferentes en su conjunto de datos de un año al siguiente desde 1880 hasta 2011. A mano, esto sería una pesadilla. Afortunadamente, al escribir código en R, todo lo cual está disponible en GitHub, Hilary pudo generar estos valores para todos estos nombres durante todos estos años. Pero es importante saber que después de reunir los datos, el siguiente paso es averiguar qué necesita hacer con esos datos para responder a su pregunta. Para la pregunta de Hilary, calcular el riesgo relativo de cada nombre de un año al siguiente de 1880 a 2011 y observar el porcentaje de bebés nombrados con cada nombre en un año en particular sería lo que tenía que hacer para responder a su pregunta.



<https://github.com/hilaryparker/names>

Análisis exploratorio de datos: Averiguar cómo hacer lo que quiere hacer para responder a su pregunta de interés es parte del proceso, no siempre aparece en su proyecto final y puede llevar mucho tiempo.

Resultados del análisis de datos: Dicho esto, dado que Hilary ya tenía calculados los valores necesarios, comenzó a analizar los datos. Lo primero que hizo fue mirar los nombres con la mayor caída en porcentaje de un año al siguiente. Según este análisis preliminar, Hilary ocupó el sexto lugar en la lista, lo que significa que había otros cinco nombres que habían tenido una caída de popularidad en un solo año más grande que el que experimentó el nombre "Hilary" de 1992 a 1993.

Name	Loss (%)	Year
Farrah	78	1978
Dewey	74	1899
Catina	74	1974
Deneen	72	1965
Khadijah	72	1995
Hilary	70	1993
Clementine	69	1881
Katina	69	1974
Renata	69	1981
Iesha	69	1992
Minna	68	1883
Ashanti	68	2003
Celestine	67	1881
Infant	67	1991

Otros proyectos de ciencia de datos:

- [Text analysis of Trump's tweets confirms he writes only the \(angrier\) Android half](#), by [David Robinson](#)
- [Where to Live in the US](#), by [Maelle Salmon](#)
- [Sexual Health Clinics in Toronto](#), by [Sharla Gelfand](#)

R

R es un lenguaje de programación en un entorno centrado principalmente en el análisis estadístico y gráfico.

R se descarga en la red de archivo integral R o CRAN

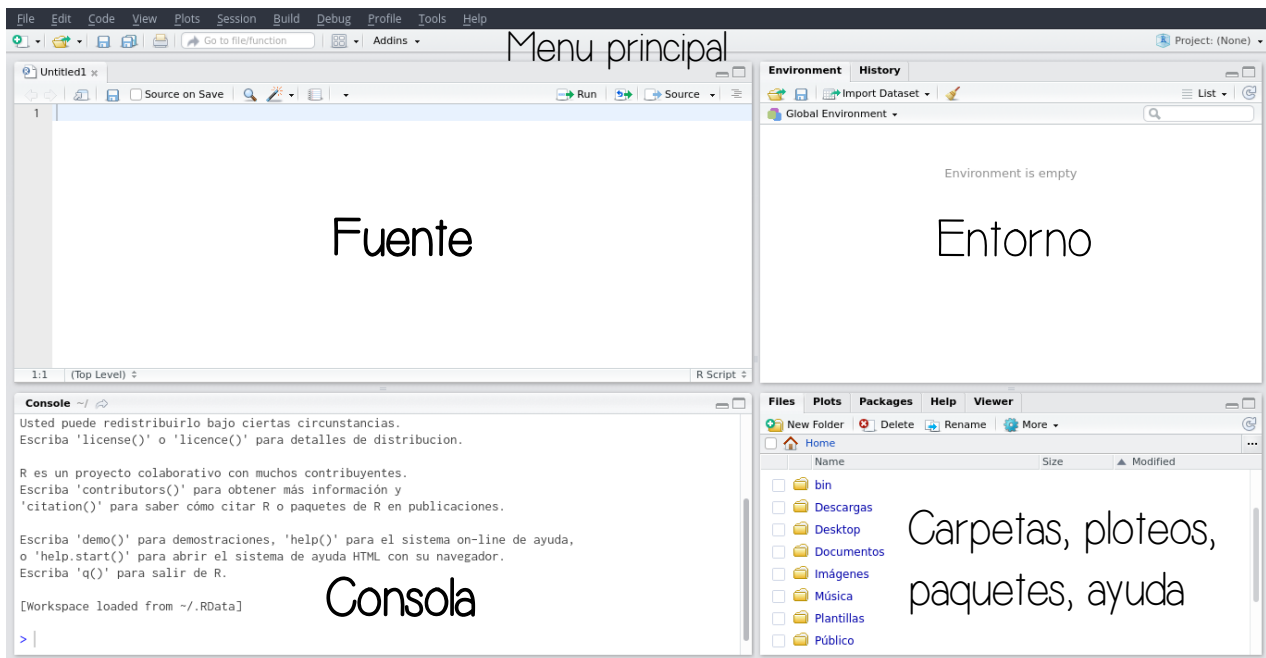
- R es más popular
- R se está convirtiendo rápidamente en el lenguaje estándar para el análisis estadístico. Esto hace que R sea un gran idioma para aprender ya que el software más populares, cuanto más rápido se desarrolla la nueva funcionalidad, cuanto más potente se vuelve y mejor es este soporte.
- R es uno de los 5 primeros idiomas que se piden en las publicaciones de trabajo de los científicos de datos
- R es libre
- Muy versátil: tmb se usa para hacer sitios web y mapas, usar datos SIG, analizar lenguaje e incluso hacer conferencias y videos



R Studio

Es una interfaz grafica de usuario para R que permite escribir editar y almacenar código, generar, ver y almacena trazados, gestionar archivos, objetos y dataframes e integrar con sistemas de control de versiones por nombrar algunas de sus funciones.

Se descarga en la pagina de RStudio



R Packages

Paquete: Un paquete es una colección de funciones, datos, y código convenientemente proporcionado en un buen formato completo para usted Hay aprox 14300 paquetes disponibles para descargar cada uno con sus propios funciones especializadas y código todos para algún propósito distinto.

Biblioteca (library): Es el lugar donde se encuentra el paquete en su computadora y un paquete es un libro dentro de la biblioteca. La biblioteca es donde se encuentran el libro/paquete.

Cada paquete es desarrollado y publicado por la comunidad R en general y depositado en repositorios

Repositorio: Es una ubicación central donde muchos paquetes desarrollados se encuentran y están disponibles para su descarga. Hay 3 grandes repositorios:

- La red integral de archivos R o CRAN (es el más grande repositorio de R con mas de 12 100 paquetes disponibles
- Bioconductor Repository: es principalmente para paquetes de enfoque Bioninformatic
- GitHub: un repositorio de código abierto muy popular que no es específico de R

Como encontrar el paquete que queremos?

CRAN agrupa todos sus paquetes por su funcionalidad/tema en 35 temas

RDocumentation es un motor de búsqueda de paquetes y funciones de CRAN y Bioconductor y GitHub

Como instalar los paquetes?

- Desde cran:
Tools > Install Packages
- Desde Bioconductor: Instalar Bioconductor (<https://bioconductor.org/biocLiteR>) y usar el comando

```
biocLite("Package")
```

- Desde GitHub:

```
install.packages("devtools")  
library(devtools)  
install_github("autor/package")
```

Luego llamamos/cargamos el paquete con el comando: **library(Package)**

Que paquetes están instalados?

```
installed.packages() o library()
```

Actualizar paquetes:

```
old.packages()  
update.packages()  
install.packages("packagename")
```

algo

algo

bggbv

gfvfvb

<https://frhik.github.io/DiplomadoR/introduccion-a-r.html#manejo-de-datos>

Help:

```
help(package=" ")
```

Viñetas son archivos de ayuda ampliados

```
browseVignettes("ggplot2")
```

Proyectos en R

Cuando haces un Proyecto crea una carpeta donde se guardarán todos los archivos que es útil para organizarte y mantener varios proyectos separados uno del otro al volver a abrir un nuevo proyecto Rstudio recuerda que archivos están abiertos y restaurara el entorno de trabajo como si nunca te hubieras ido



Funcionalmente, crea un proyecto en R creara una nueva carpeta y asignara ese como directorio de trabajo para que todos los archivos generados se asignen al mismo directorio. El principal beneficio de usar proyectos es que inicia el proceso de organización de forma correcta. Crea una carpeta para ti y ahora tienes un lugar para almacenar todos tus datos de entrada, tu código y salida de Código

Control de versiones

Es un sistema que registra cambios que se realizan en un archivo o conjunto de archivos a lo largo del tiempo. A medida que realiza ediciones, el sistema de control de versiones toma instantáneas de sus archivos y los cambios y luego guarda esas instantáneas para que se pueda hacer referencia volver a versiones anteriores mas adelante si es necesario.

Sistemas de control de versiones como Git son como un cambio de pista más sofisticado en que son mucho mas potentes y son capaces de rastrear meticulosamente cambios sucesivos en muchos archivos donde potencialmente muchas personas han trabajado simultáneamente en los mismos grupos de archivos. El control de versiones trabaja manteniendo una sola versión actualizada de cada archivo con un registro de todas las versiones anteriores y un registro de exactamente lo que cambio entre las versiones.

Esto puede ser útil cuando estamos colaborando con mucha gente en los mismos archivos, el software de control de versiones realiza un seguimiento de quien cuando y por qué se realizaron estos cabios específicos.

Git es un sistema de control de versiones libre y de código abierto. Fue desarrollado en 2005 y desde ahí es el sistema de control de versiones más utilizado alrededor. Git es un software utilizado localmente en tu ordenador para registrar cambios



Beneficios:

- Git mantiene una copia local de su trabajo y revisiones que luego puede verse fuera de línea. Luego una vez que regrese al servicio de internet puede sincronizar su copia del trabajo con todas sus nuevas ediciones y realizar un seguimiento de los cambios en el repositorio principal en línea.
- Facilidad interacción entre Git y Rstudio

Github es una interfaz en línea para git

Github es un host para sus archivos y los registros de los cambios realizados

Repositorio: Es equivalente a una carpeta o directorio de proyectos. Los repositorios están alojados en github y a través de esta interfaz puedes mantener tus repositorios privados y compartirlos con seleccionados colaboradores o puedes hacerlos públicos

Commit: Es el sitio donde guardas tus ediciones y cambios realizados. Es como una instantánea de tus archivos (like a Snapchat of your files)

Push: Es la actualización del repositorio con sus ediciones. Es enviar los cambios confirmados a ese repositorio por lo que ahora todos tienen acceso a sus ediciones

Pull: Es la actualización de la versión local del repositorio a la versión actual ya que todos pueden haber editado mientras tanto

Staging: (puesta en escena) : El acto de preparar un archivo para una confirmación. La puesta en escena le permite separar los cambios de archivos en confirmaciones separadas.

Resumen: Los archivos están alojados en un repositorio (**repository**) que se comparte en línea con los colaboradores. Extraiga (you **pull**) el contenido del repositorio para que tenga una copia local de los archivos que puede editar. Una vez que esté satisfecho con sus cambios en un archivo, lo escenificará (**stage**) y lo confirmará (**commit**). Empuja (you **push**) esta confirmación en el repositorio compartido. Esto carga tu nuevo archivo y todos los cambios y va acompañado de un mensaje explicando lo que cambió, por qué, y por quién



Rama (branch): Cuando un mismo archivo tiene dos copias simultáneas.

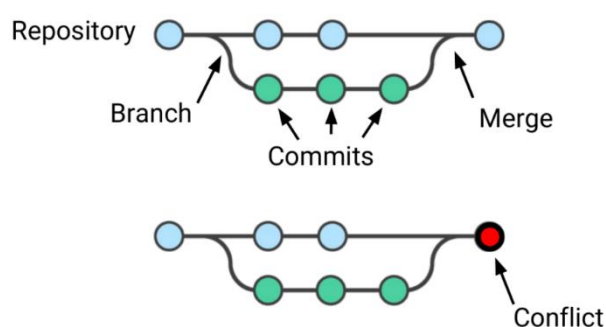
Fusionar (merge): las ediciones independientes del mismo archivo se incorporan en un solo archivo unificado. Git identifica las ediciones independientes y las reúne en un solo archivo, con ambos conjuntos de ediciones incorporados.

Conflicto: cuando varias personas realizan cambios en el mismo archivo y Git no puede fusionar las ediciones.

Clonar: Hacer una copia de un repositorio de Git

existente. Si acaba de ingresar a un proyecto que ha sido rastreado con control de versiones, debe clonar el repositorio para obtener acceso y crear una versión local de todos los archivos del repositorio y todos los cambios registrados.

Bifurcación (fork): una copia personal de un repositorio que ha tomado de otra persona



Mejores practicas:

- Realizar compromisos intencionados. Cada confirmación debe abordar solo un problema. De esta manera, si necesita identificar cuándo cambió una determinada línea de código, solo hay un lugar donde buscar para identificar el cambio y puede ver fácilmente cómo revertir el código.
- Escribir mensajes informativos en cada confirmación es un hábito útil. Si cada mensaje es preciso en lo que se estaba cambiando, cualquiera puede examinar el archivo comprometido e identificar el propósito de su cambio.
- Tenga en cuenta la versión de los archivos en los que está trabajando. Compruebe con frecuencia que está actualizado con el repositorio actual tirando con frecuencia

- ☒ **Purposeful, single issue commits**
- ☒ **Informative commit messages**
- ☒ **Pull and push often**

Preguntas:

I'm done editing a file, I need to _____ those changes then _____ them, and _____ it to the _____.

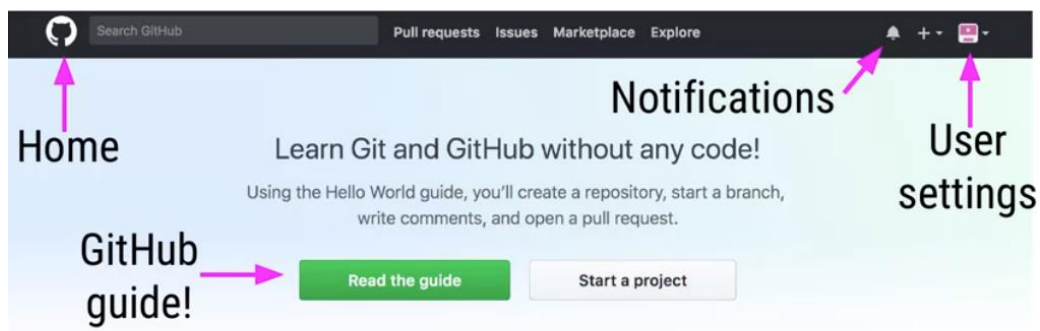
Stage, commit, push, repository

Version control minimizes the need to save different versions of the same file on your computer

Version control helps make sure that you do not lose work that you have done

Github and Git

GitHub es un sistema de gestión basado en la nube para sus archivos controlados por versiones. Al igual que Dropbox, tus archivos están a la vez localmente en tu ordenador y alojados en línea y de fácil acceso. Su interfaz le permite administrar el control de versiones y proporciona a los usuarios una interfaz basada en la web para crear proyectos, compartirlos, actualizar código, etc.



© 2018 GitHub, Inc. [Terms](#) [Privacy](#) [Security](#) [Status](#) [Help](#) [Contact GitHub](#) [API](#) [Training](#) [Shop](#) [Blog](#) [About](#)

Help files are
your friends!

Instalación de Git: FALTA LINK

Configuración de Git

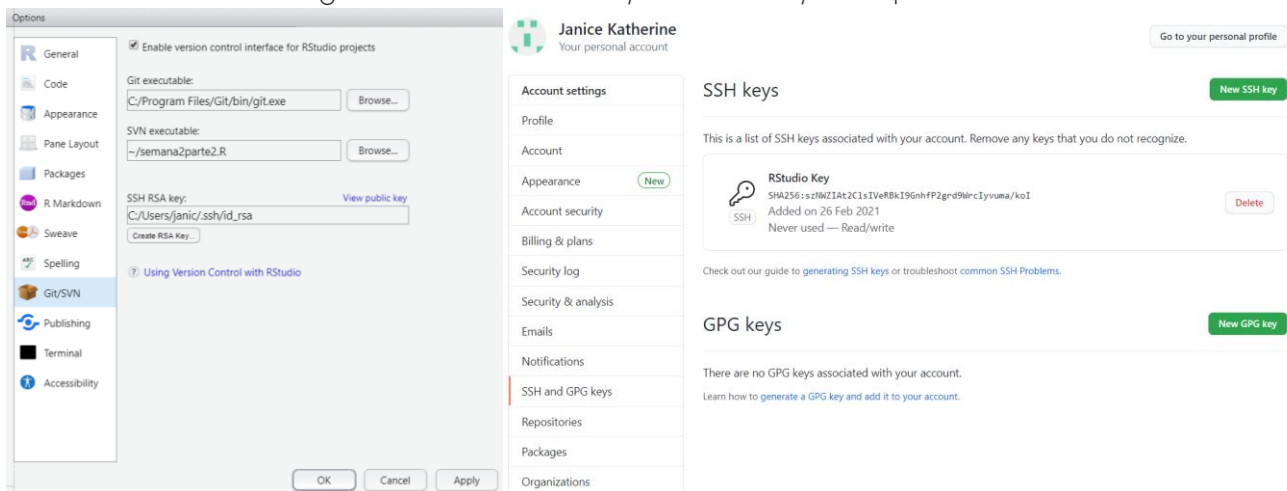
```
MINGW64:/c:/Users/janic
janic@LAPTOP-MP8T33KT MINGW64 ~
$ git config --global user.name "Janice Escobedo"
janic@LAPTOP-MP8T33KT MINGW64 ~
$ git config --global user.email janice.ev224@gmail.com
janic@LAPTOP-MP8T33KT MINGW64 ~
$ git config --list
diff.astextplain.textconv=astextplain
filter.lfs.clean=git-lfs clean -- %f
filter.lfs.smudge=git-lfs smudge -- %f
filter.lfs.process=git-lfs filter-process
filter.lfs.required=true
http.sslbackend=openssl
http.sslcainfo=C:/Program Files/Git/mingw64/ssl/certs/ca-bundle.crt
core.autocrlf=true
core.fscache=true
core.symlinks=false
pull.rebase=false
credential.helper=manager-core
credential.https://dev.azure.com.usehttppath=true
init.defaultbranch=master
user.name=Janice Escobedo
user.email=janice.ev224@gmail.com
```

Linking GitHub and RStudio

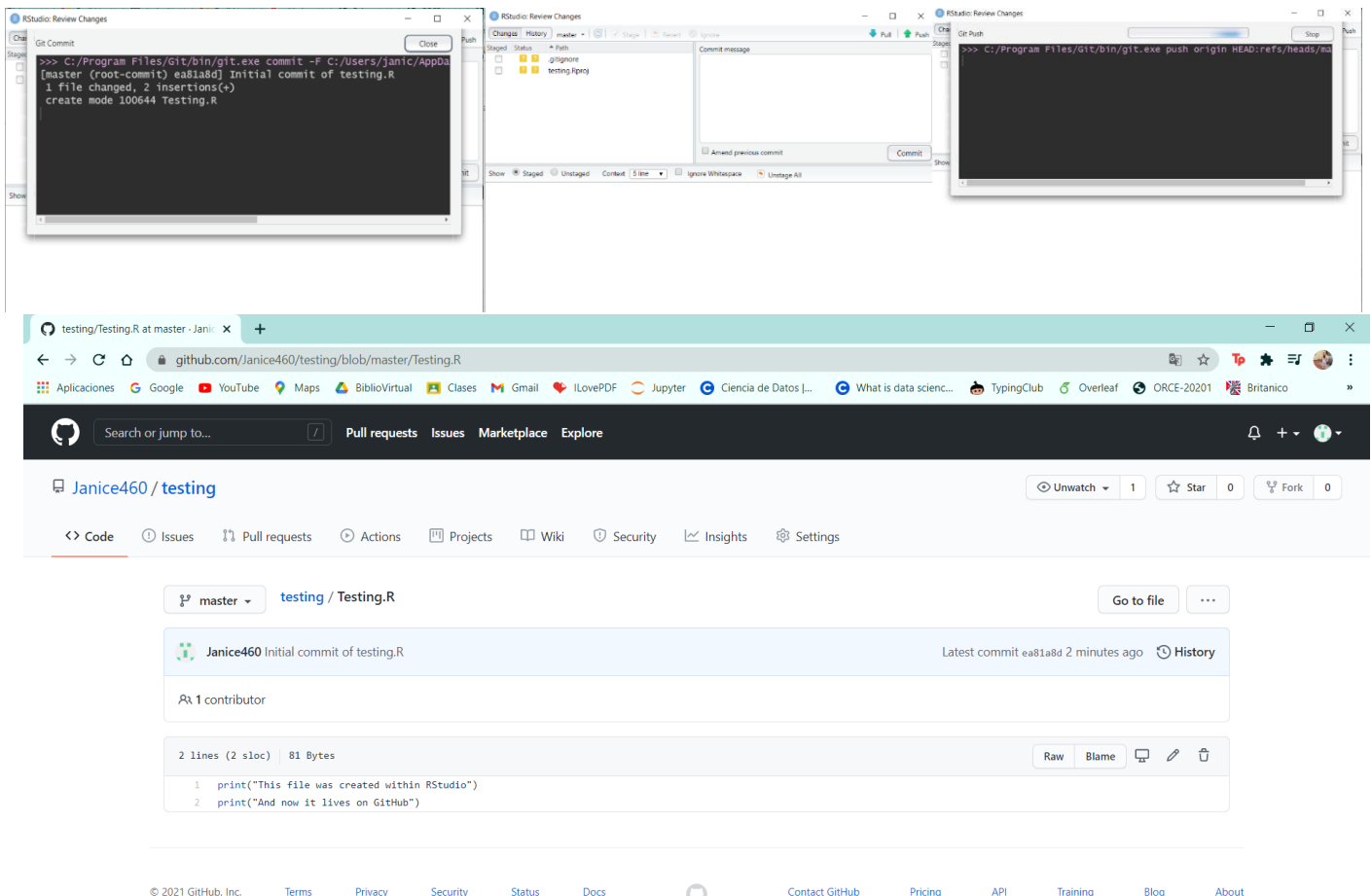
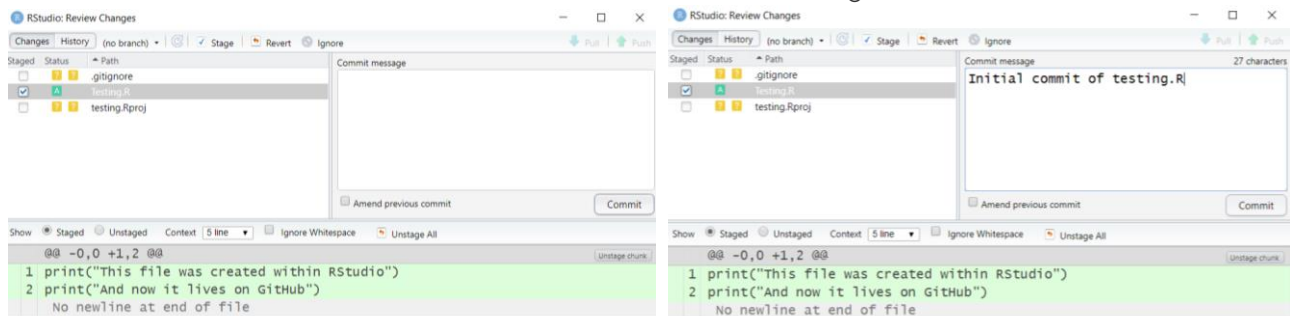
Crear nuevo Proyecto (testingRproj)

Tools > Global options > Git/SVN > Generar llave > View public key

Settings > SSH and GPG keys > new key > Copiar la llave



Crear un Rscript (TestingR) > En environment ir a la ventana Git (abrirá una nueva ventana) > Seleccionar el archivo creado > Escribir un commit message > Darle a Commit > Push



Hemos enlazado git con rstudio para que Rstudio reconozca lo que estas usando como tu software de control de versiones

Dsp de esto vinculamos rstudio a github para que pueda empujar y extraer repositorios desde RStudio

Preguntas:

1. In what quadrant of RStudio will you find the Git tab? Environment
2. What is the order of commands to send a file to GitHub from within RStudio
Stage > Commit message > commit > push
3. Which can you do from the Commit window of RStudio?
 - See the differences between your original file and your updated file
 - Stage files
 - Pull and push content from the repository
 - Write a commit message

Proyectos bajo control de versiones

Linking an existing Project with Git:

Crear un nuevo Proyecto (temporally_add_to_control_version)

Crear un repositorio con el mismo nombre del proyecto > usar las líneas de código adecuadas para vincularlo con git > actualizar R y ver disponibilidad de la ventana Git

```
janic@LAPTOP-MP8T33KT MINGW64 /c
$ cd Coursera

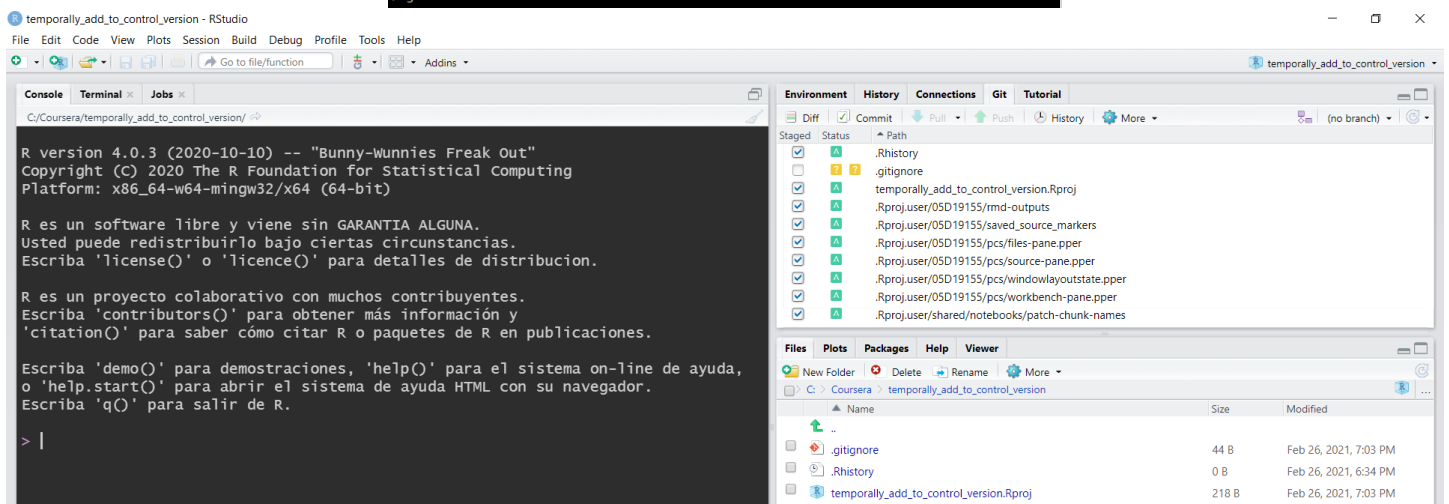
janic@LAPTOP-MP8T33KT MINGW64 /c/Coursera
$ cd temporally_add_to_control_version

janic@LAPTOP-MP8T33KT MINGW64 /c/Coursera/temporally_add_to_control_version
$ git init
Initialized empty Git repository in C:/Coursera/temporally_add_to_control_version/.git/

janic@LAPTOP-MP8T33KT MINGW64 /c/Coursera/temporally_add_to_control_version (master)
$ git add .
warning: LF will be replaced by CRLF in .Rproj.user/05D19155/pcs/files-pane.pper
The file will have its original line endings in your working directory
warning: LF will be replaced by CRLF in .Rproj.user/05D19155/pcs/source-pane.pper
The file will have its original line endings in your working directory
warning: LF will be replaced by CRLF in .Rproj.user/05D19155/pcs/windowlayoutstate.pper
The file will have its original line endings in your working directory
warning: LF will be replaced by CRLF in .Rproj.user/05D19155/pcs/workbench-pane.pper
The file will have its original line endings in your working directory
warning: LF will be replaced by CRLF in .Rproj.user/05D19155/rmd-outputs.
The file will have its original line endings in your working directory

janic@LAPTOP-MP8T33KT MINGW64 /c/Coursera/temporally_add_to_control_version (master)
$ git remote add origin https://github.com/Janice460/temporally_add_to_control_version.git

janic@LAPTOP-MP8T33KT MINGW64 /c/Coursera/temporally_add_to_control_version (master)
$ git branch -M main
```



Trabajar en un repositorio de GitHub existente:

Crear proyecto en versión control

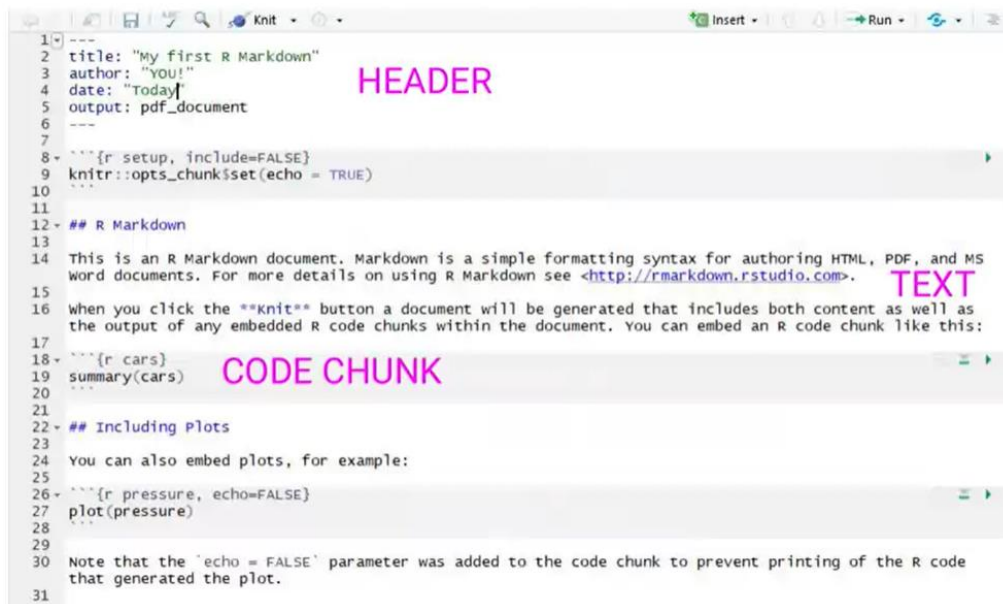
Preguntas:

1. What do you call it when you create a local copy of a repository that you will work on collaboratively with the original repository owner? Clone
2. What is the command to initialize git in a directory? **git init**
3. How do you add all of the contents of a directory to version control? **git add .**
4. How do you make a commit from within the command line? **git commit -m "Message"**

R Markdown

R Markdown es una forma de crear documentos totalmente reproducibles en los que tanto texto como código se pueden combinar. A pesar de que todos los comienzan sin texto se pueden renderizar en HTML o PDF o Word documentos o diapositivas

- Por su facilidad de combinar texto y código puede integrar fácilmente introducciones hipótesis, resultados y conclusiones, todo en un solo documento.
- Por ser texto plano funciona muy bien con los sistemas de control de versiones. Es fácil rastrear que cambios de caracteres ocurren entre confirmaciones a diferencia de otros formatos que estén en texto sin formato



The screenshot shows an RStudio window with an R Markdown document. The document is divided into sections: a header (lines 1-6), a code chunk (lines 8-9), a text block (lines 14-16), another code chunk (lines 18-19), and a final text block (lines 24-30). The sections are labeled with pink text: 'HEADER', 'TEXT', and 'CODE CHUNK'. The code chunks contain R code for setting up knitr options and for summarizing cars and plotting pressure data.

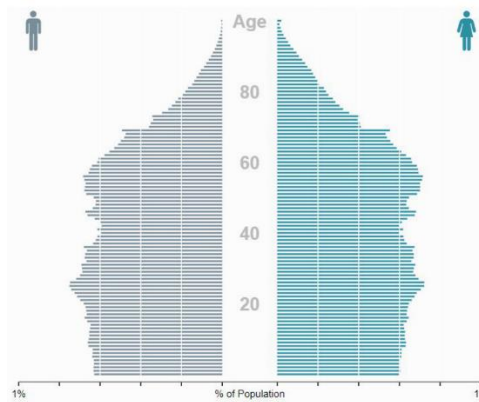
```
1 ---
2 title: "My first R Markdown"
3 author: "YOU!"
4 date: "Today"
5 output: pdf_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS
15 word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 when you click the **knit** button a document will be generated that includes both content as well as
18 the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
19
20 ```{r cars}
21 summary(cars)
22 ```
23
24 ## Including Plots
25
26 You can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30 ```
31
32 Note that the 'echo = FALSE' parameter was added to the code chunk to prevent printing of the R code
33 that generated the plot.
```

Más sobre R Markdown: <https://rmarkdownrstudio.com/>

Tipos de preguntas sobre ciencia de datos

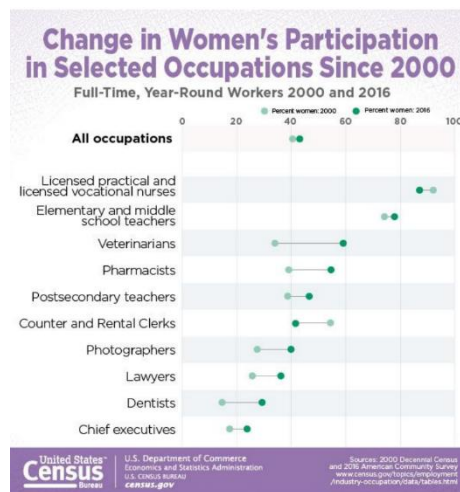
Tipos de análisis de datos según orden aproximado de dificultad:

1. Descriptivos: describir o resumir un conjunto de datos
 - a. El primer análisis que hace cuando recibe los datos
 - b. Genera simples resúmenes sobre las muestras y sus mediciones
 - i. Ejm: medidas de tendencia central (media, mediana, modo), medidas de variabilidad (rango, desviación estándar, varianza)
 - c. NO generaliza los resultados a una población mayor, o tratando de sacar conclusiones



El objetivo es describir la distribución. No hay inferencias sobre lo que esto significa o predicciones sobre cómo podrían tender los datos en el futuro. Es solo para mostrar un resumen de datos recopilados.

2. Exploratorios: Examinar o explorar los datos y encontrar relaciones que no se conocían previamente
 - a. Exploran como las diferentes medidas puedan estar relacionadas entre sí, pero no confirman que la relación sea causal (correlación no implica causalidad)
 - b. Útil para descubrir nuevas conexiones
 - c. Puede permitir formular hipótesis e impulsar el diseño de futuros estudios y la recopilación de datos



Aquí podemos ver como 2 o mas variables están relacionadas entre si, podemos sacar muchas relaciones

3. Inferenciales: usa una muestra relativamente pequeña de datos para inferir algo sobre la población en general. El análisis inferencial es comúnmente el objetivo de la modelización estadística (donde tienes una pequeña población para extrapolar y generalizar esa información a un grupo más grande)
 - a. proporcione su estimación de la variable para la población y proporcione su incertidumbre sobre su estimación
 - b. Su capacidad de inferir con precisión información sobre la población mas grande depende en gran medida de su esquema de muestreo

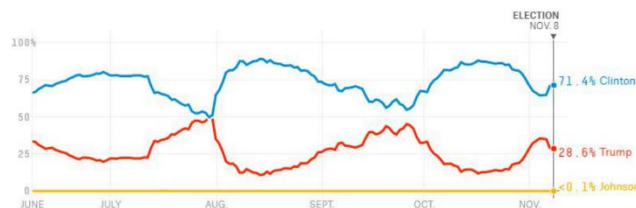
Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 U.S. Counties for the Period from 2000 to 2007

Andrew W. Correia; C. Arden Pope; Douglas W. Dockery; Yun Wang; Majid Ezzati; Francesca Dominici

4. Predictivos: Utiliza datos actuales para hacer predicciones sobre datos futuros, usa datos históricos o actuales para encontrar patrones y predecir la probabilidad de datos futuros
 - a. Su precisión en las predicciones depende de la medición de las variables correctas
 - b. Hay muchas maneras de construir modelos de predicción siendo algunos mejores o peores para cada caso específico
 - i. Teniendo mas datos y un modelo simple generalmente funciona bien en predecir resultados futuros

Solo porque una variable pueda predecir otra no significa que una causa la otra

FiveThirtyEight's predictions of the 2016 US election



No hay maneras fáciles de medir lo bien que vas a predecir un evento hasta que ese evento haya llegado a suceder. Así que evaluar diferentes enfoques o modelos es un reto

5. Casuales: Ve que sucede con una variable cuando manipulamos la otra variable, mirando la causa y el efecto de la relación
 - a. estándar de oro en el análisis de datos: es bastante complicado con los datos observados solamente. Siempre habrá pregunta sobre si estas correlaciones están impulsando sus conclusiones, o si las suposiciones subyacentes a su análisis son validas
 - i. se ve con frecuencia en el estudio de científicos donde los científicos están buscando la causa de un fenómeno
 - b. A menudo se aplica a los resultados de estudios aleatorizados que fueron diseñados para identificar la causalidad
 - c. Los datos generalmente se analizan en agregado y las soluciones observadas suelen ser efectos promedio

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy

R.S. Finkel, E. Mercuri, B.T. Darras, A.M. Connolly, N.L. Kuntz, J. Kirschner, C.A. Chiriboga, K. Saito, L. Servais, E. Tizzano, H. Topaloglu, M. Tulinius, J. Montes, A.M. Glanzman, K. Bishop, Z.J. Zhong, S. Gheuens, C.F. Bennett, E. Schneider, W. Farwell, and D.C. De Vivo, for the ENDEAR Study Group*

Ejemplo: ensayos controlados aleatorizados para medicamentos

6. Mecanistas: Comprender los cambios exactos en las variables que conducen a cambios exactos en otras variables. Son extremadamente difíciles de usar para inferir mucho, excepto en situaciones simples, o aquellas que estén bien modeladas por ecuaciones deterministas
- a. Comúnmente aplicado para ciencias físicas o de ingeniería
 - i. Ciencias biológicas, son demasiado ruidosas para usar análisis mecanicista
 - b. A menudo cuando se aplican estos análisis, el único ruido en los datos es el error de medición que puede ser contabilizado



Polymer Testing
Volume 61, August 2017, Pages 364-372



Compatibilization of toughened polypropylene/biocarbon biocomposites: A full factorial design optimization of mechanical properties

Ehsan Behazin ^{a, b}, Manjusri Misra ^{a, b}, Amar K. Mohanty ^{a, b}

En este estudio se estaba examinando cómo el tamaño de las partículas de biocarbono, el tipo de polímero funcional y la concentración afectaban las propiedades mecánicas del plástico resultante.

Diseño experimental

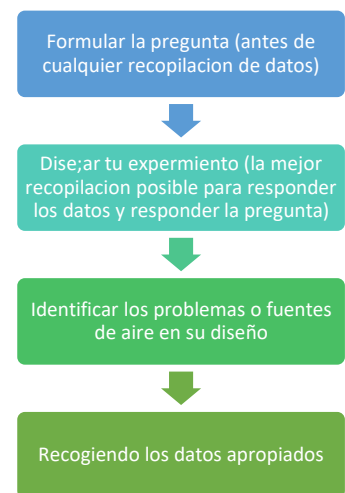
El diseño experimental es la organización y los experimentos. Entonces que usted tiene los datos correctos y suficientes de ellos para responder clara y eficazmente a su pregunta de ciencia de datos. Si haces un análisis erróneo puedes llegar a conclusiones equivocadas. A veces las malas prácticas son resultado de un mal diseño experimental y análisis

Algunos términos inherentes al diseño experimental

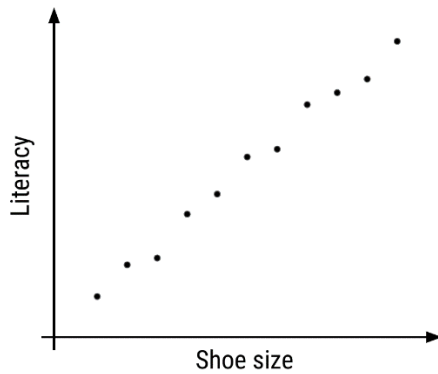
1. **Variable independiente (AKA factor):** Es la variable que es experimentador manipula. No depende de otras variables que se midan. (eje X)
 2. **Variables dependientes:** Son aquellas que se espera que cambie como resultado de cambios en la variable independiente (eje Y)
- Adicionalmente debemos desarrollar una hipótesis, como una conjetura educada en cuanto a la relación entre una variable y el resultado de su experimento.

Ejemplo:

Imaginemos que el tamaño de un zapato aumenta igual que su alfabetización



Hypothesis: As shoe size increases, literacy also increases



El tamaño de zapato afecta la alfabetización?

medir el tamaño de 100 tallas zapatos y probar su nivel de alfabetización

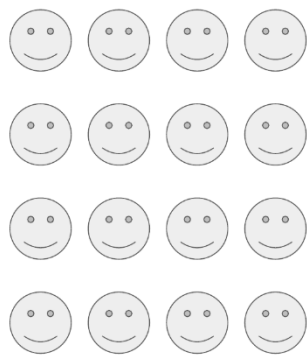
Confundidor: edad
Ajustar o arreglar

Recogemos los datos

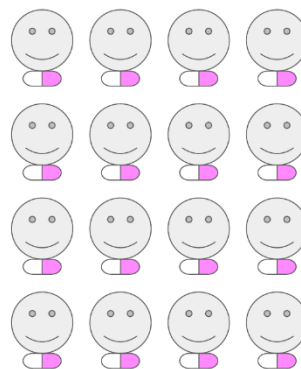
- Fluidez de lectura en el eje Y
- Para responder a la pregunta se diseñará un experimento en el que se medirá este tamaño de zapato y el nivel de alfabetización de 100 individuos
- El tamaño de la muestra es el número de sujetos experimentales que incluirá en su experimento
- Antes de recopilar los datos necesitamos considerar si hay problemas con este experimento que podrían causar un resultado erróneo. En este caso el experimento puede ser fatalmente defectuoso por un confundidor
- Confundidor: es una variable extraña que puede afectar la relación entre las variables dependientes e independientes
- Para esto en el ejemplo podemos tener en cuenta los efectos de la edad en la alfabetización y de otra manera podríamos controlar para las edades el efecto en alfabetización sería fijar la edad de todos los participantes

En otros paradigmas de diseño experimental puede ser apropiado un grupo de control esto es cuando tienes un grupo de sujetos experimentales que no son manipulados

Control group

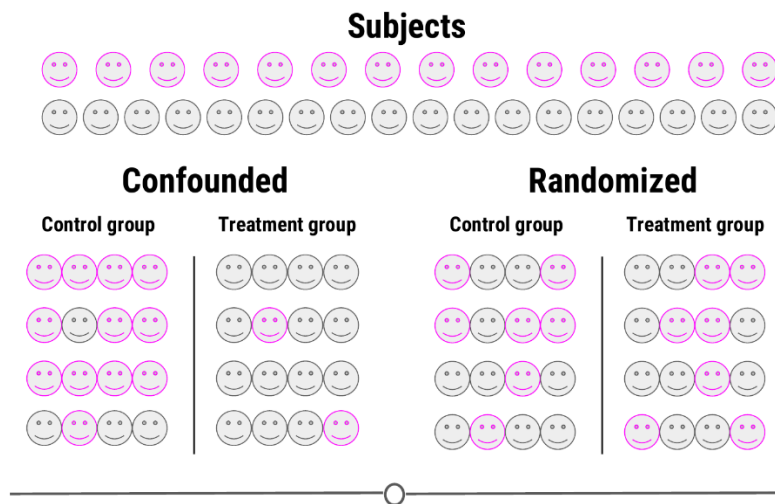


Treatment group



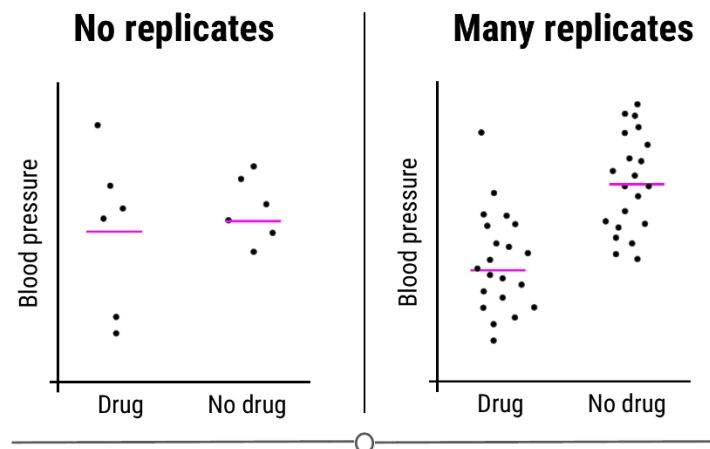
Hay estrategias que se pueden usar para controlar los efectos de confusión

- I Ponemos a ciegas a su grupo de tratamiento asignado: los participantes están cegados ante el grupo de tratamiento que se encuentran. Esto generalmente se logra dando el grupo de control y el tratamiento de bloqueo

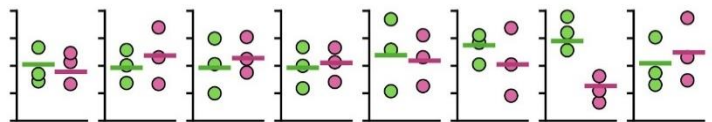


Replicación: repetir un experimento con diferentes sujetos experimentales. Como experimentos individuales los resultados pueden haber ocurrido por casualidad. Un confundidor fue distribuido de manera desigual entre sus grupos. Hubo un error sistemático en la recopilación de datos. Hubo unos valores atípicos etc.

Sin embargo; si usted puede repetir el experimento recoger un nuevo conjunto de datos y todavía llegar a la misma conclusión, su estudio es mucho más fuerte. Tmb en el corazón de la replicación es que le permite medir la variabilidad de sus datos con mayor precisión.



P-hacking: Esto es cuando busca exhaustivamente un conjunto de datos para encontrar patrones y correlaciones que parecen estadísticamente significativos en virtud del gran número de pruebas que ha realizado. Estas correlaciones espurias se pueden reportar como significativas y si realiza suficientes pruebas, puede encontrar un conjunto de datos y análisis que le mostrará lo que quería ver.



Preguntas:

1. En un estudio que midió el efecto de la dieta sobre el IMC, el colesterol, los niveles de lípidos, los niveles de triglicéridos y el índice glucémico, ¿cuál es una variable independiente?
Dieta / niveles de lípidos / no es IMC
2. Which of the following is NOT a method to control your experiments? No es efectos del placebo, si es grupo de control y blinding
3. ¿Cuál podría ser un factor de confusión en un experimento que analiza la relación entre la prevalencia de las canas en una población y las arrugas? Edad
4. De acuerdo con las recomendaciones del grupo Leek, ¿qué datos necesita compartir con un estadístico colaborador?

- Los datos brutos
- Un conjunto de datos ordenado
- Un libro de códigos que describe cada variable y sus valores en el ordenado conjunto de datos
- Una receta explícita y exacta de cómo pasó de los datos sin procesar a los datos ordenados y al libro de códigos

- Si establece su nivel de significancia en un valor $p \leq 0.01$, ¿cuántas pruebas significativas esperaría ver por casualidad si realiza 1000 pruebas? 10
- ¿Qué es una herramienta de diseño experimental que se puede utilizar para abordar variables que pueden ser factores de confusión en la fase de diseño de un experimento? Randomization
- ¿Cuál de los siguientes describe un análisis descriptivo? Genere una tabla que resuma la cantidad de observaciones en su conjunto de datos, así como las tendencias y variaciones centrales de cada variable.

Big Data

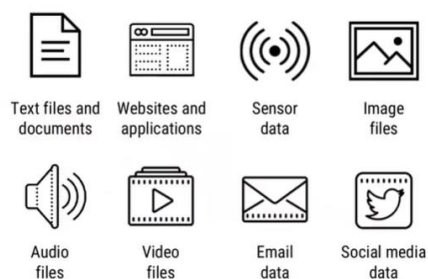
El big data se ha convertido en una palabra clave en la ciencia de datos. Big data son datasets muy grandes

A medida que la tecnología en el almacenamiento de datos ha evolucionado para poder contener conjuntos de datos cada vez más grandes. La definición de <<grande>> también ha evolucionado. Nuestra capacidad de recopilar y registrar datos ha mejorado con el tiempo de manera que la velocidad con la que se recogen los datos es sin precedentes

Uno de los principales cambios en la ciencia de datos ha sido pasar de conjuntos de datos estructurados a abordar datos no estructurados

Name	Country of origin	Sex	Weight (kg)	Height (cm)
A. Bee	Canada	M	75	163
C. Dee	UAE	M	80	180
E. Eff	China	F	72	175
G. Haitch	South Africa	F	68	172
I. Jay	Poland	M	77	168
K. Elle	Japan	N/A	76	173
M. Enn	Chile	M	80	190

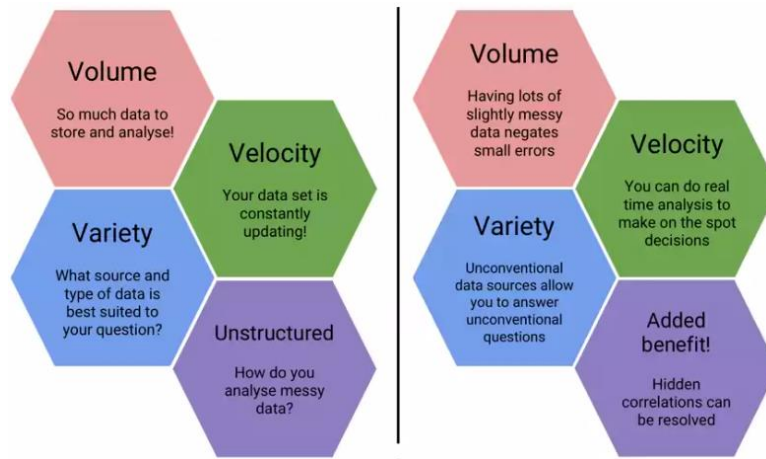
Unstructured Data Types



Los datos estructurados son lo que tradicionalmente se puede pensar de datos, tablas largas, hojas de calculo o base de datos, con columnas y filas de información que puede sumar o promediar o analizar como quiera dentro de esos límites. Pero hoy en día estos datos no se presentan así, los datos que comúnmente se encuentran son desordenados y es nuestro trabajo extraer la información que queremos y acorralarla en algo ordenado y estructurado.

Imagen: Algunas fuentes de datos no estructurados

Algunos desafíos que pueden estar asociados con el Big data



<<La combinación de algunos datos y un deseo doloroso de una respuesta no asegura que una respuesta pueda ser extraída de un determinado cuerpo de datos>>

Esencialmente, cualquier conjunto de datos puede no ser adecuado para su proyecto incluso si realmente lo quería y el big data no arregla esto. Incluso los conjuntos de datos más grandes podrían no ser lo suficientemente grandes como para que pueda responder a su pregunta si no son los datos correctos

Preguntas:

- 1 ¿Cuál es la razón detrás de la explosión del interés en big data?
 - a) El precio y la dificultad de recopilar y almacenar datos se han reducido drásticamente.
(Correcta)
 - b) Ha habido grandes mejoras en los algoritmos de aprendizaje automático.
 - c) Ha habido mejoras masivas en las técnicas de análisis estadístico.
- 2 ¿Cuál es el formato para incluir un enlace que aparece como texto azul en su documento de markdown?
[Enlace a R Studio] (www.rstudio.com)
Las URL desnudas también se resaltarán: <http://www.rstudio.com>