Exploración de Datos con Tidyverse

Mg. Jesús Salinas Flores

jsalinas@lamolina.edu.pe

Expositor

- Ingeniero Estadístico, egresado de la Universidad Nacional Agraria La Molina
- Mg. en Ingeniería Industrial con especialidad en Gestión Industrial, egresado de la Universidad Nacional Mayor de San Marcos
- Consultor estadístico

Expositor

- Profesor principal del dpto de Estadística e Informática (UNA La Molina)
- Docente en la maestría de Estadística Aplicada (UNA La Molina)
- Miembro del staff de docentes de Data Mining Consulting (DMC)
- Docente en la carrera de Ingeniería de Tecnologías de Información y Sistemas (U ESAN)

Expositor

Áreas de interés

- Machine Learning
- Reconocimiento estadístico de patrones

Facebook

•Grupo: Cursos de Estadística

YouTube

•Estadística para todos

facebook

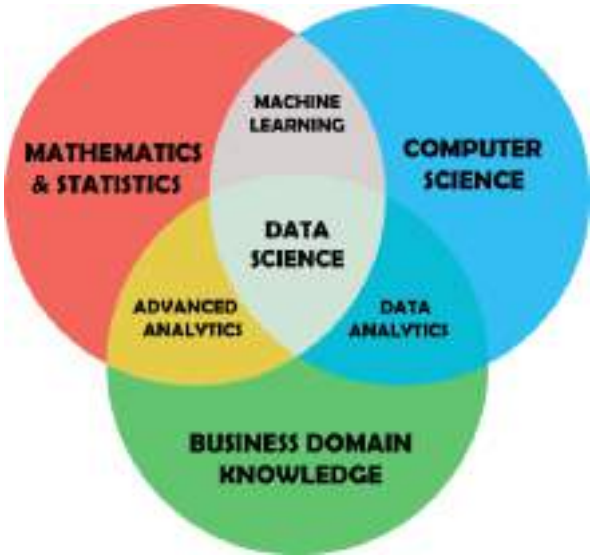
YouTube

El mundo de tidyverse

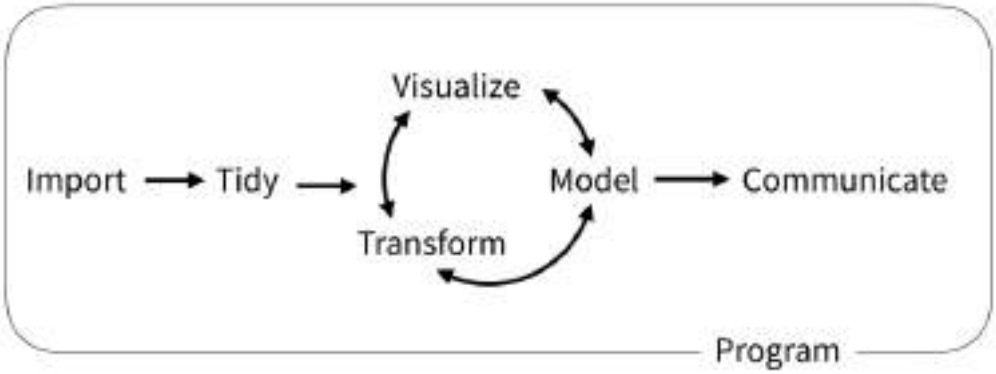
Mg. Jesús Salinas Flores

jsalinas@lamolina.edu.pe

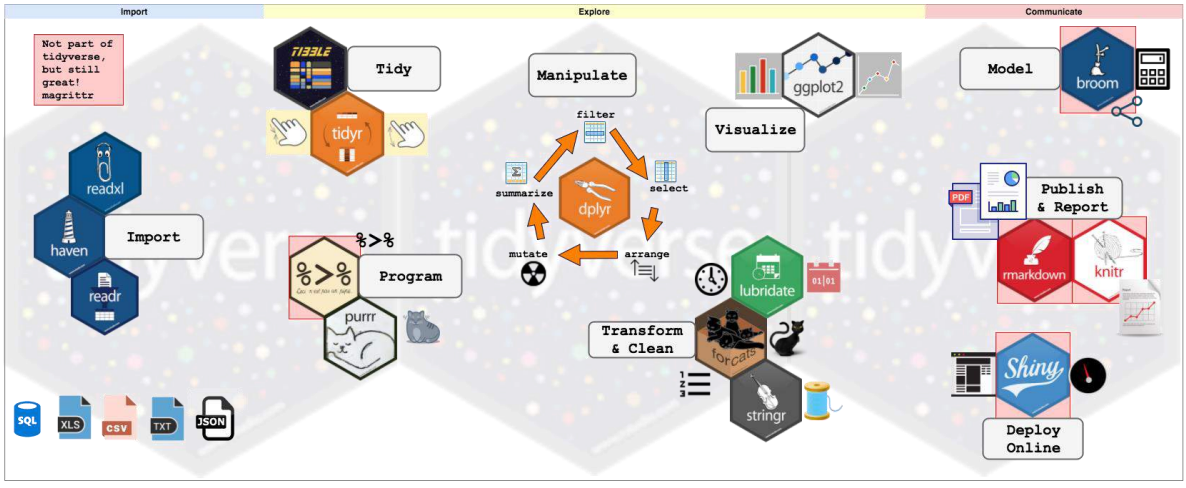
Data Science "Científico de Datos"



El mundo de tidyverse



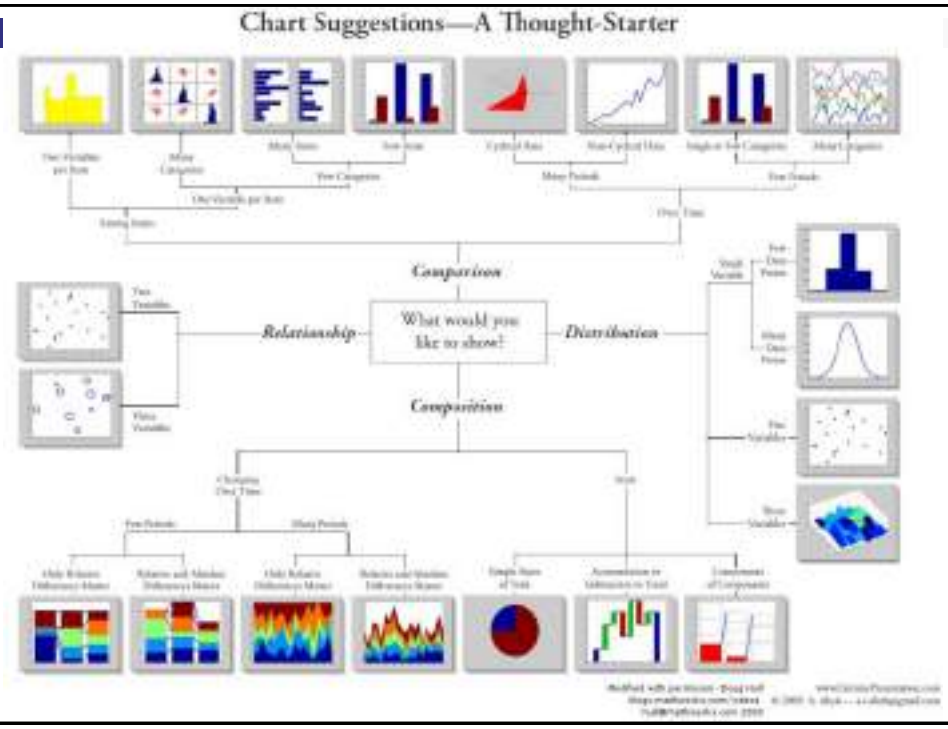
El mundo de tidyverse



La gramática de los gráficos

Mg. Jesús Salinas Flores

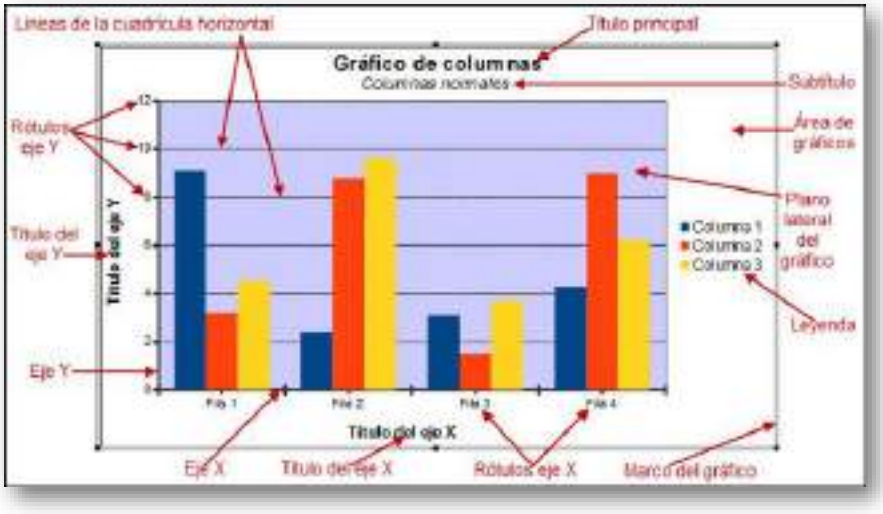
jsalinas@lamolina.edu.pe



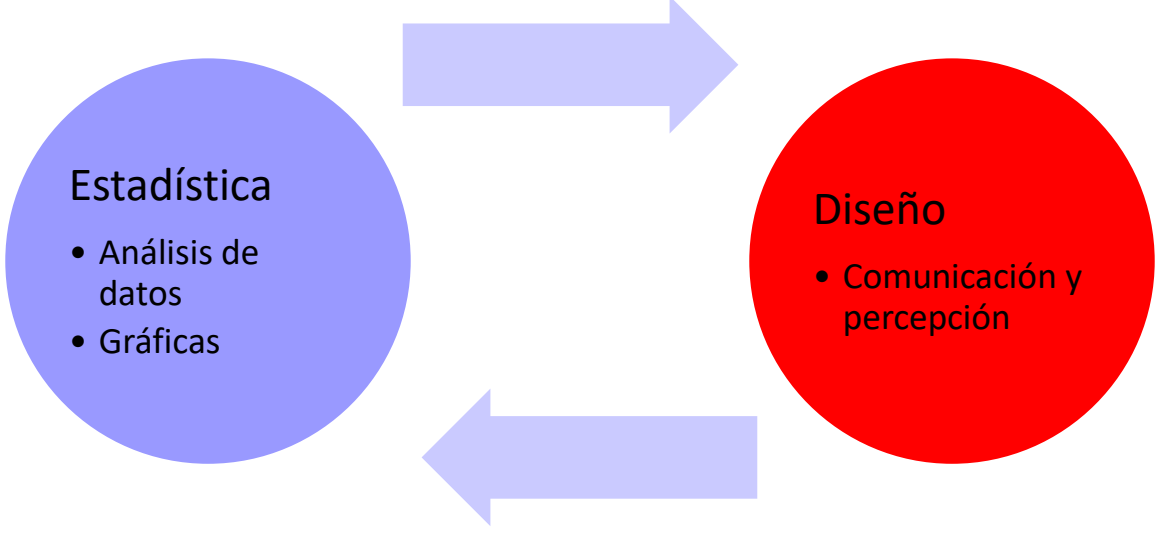
Tipos de gráficas en R

1. Gráficas para una variable cualitativa
2. Gráficas para dos variables cualitativas
3. Gráficas para una variable cuantitativa
4. Gráficas para dos variables cuantitativas
5. Gráficas para más de dos variables cuantitativas
6. Gráficas para una variable cuantitativa y una cualitativa
7. Gráficas para dos variables cuantitativas y una cualitativa
8. Gráficas para más de dos variables cuantitativas y una cualitativa

Partes de un gráfico



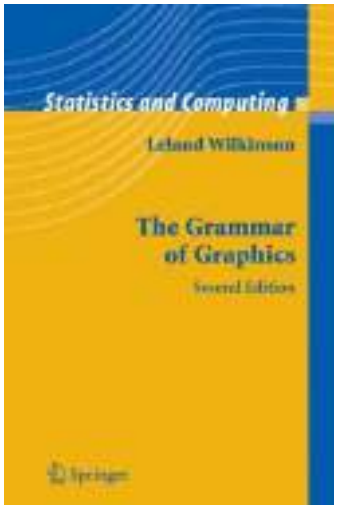
Visualización de datos y Data Science



Exploratorio vs Explicatorio

Fuente: www.datacamp.com

La gramática de los gráficos



- Si la gramática ofrece la habilidad de construir oraciones en español combinando y organizando distintos elementos del lenguaje, la *gramática de gráficos* ofrece elementos básicos para crear gráficos.
- Una "gramática de gráficos" es un marco que permite describir concisamente los componentes de cualquier gráfico.

La gramática de los gráficos

La primera proposición de una gramática de gráficos se le atribuye Leland Wilkinson quien escribió el libro “The Grammar of Graphics” en 1999.

Hadley Wickham, de RStudio y el tidyverse, propuso una grámatica de gráficos en **capas** basado en el trabajo original de Wilkinson e incluyó una adaptación de esta grámatica en R en su paquete ggplot2.

Gráficos con ggplot2
“La gramática de los gráficos”

- **ggplot2** es un paquete gráfico potente y flexible del R, implementado por Hadley Wickham.
- Un gráfico en ggplot2 consiste en tres partes fundamentales:
  - **Plot** = data + Aesthetics + Geometry

Gráficos con ggplot2
"La gramática de los gráficos"

Plot = data + Aesthetics + Geometry

- **Data** es un dataframe
- **Aesthetics** es usado para indicar las variables x e y. Se puede usar para controlar el color, el tamaño o la forma de los puntos, la altura de las barras, etc.
- **Geometry** define el tipo de gráfico (histogram, box plot, line plot, density plot, dot plot, ….)

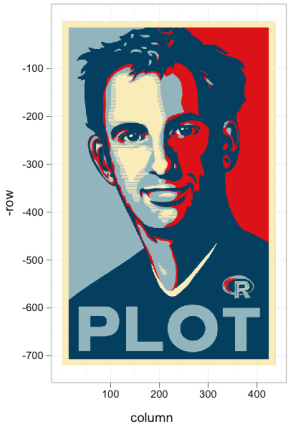
Gráficos con ggplot2

- Hay dos funciones principales en el paquete ggplot2: qplot() and ggplot()
- **qplot()** para gráficos rápidos y simples.
- **ggplot()** es más flexible y robusta que qplot para construir un gráfico paso a paso.

qplot(wt, mpg, data=mtcars)

Es equivalente a:

ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point()



Gráficos con ggplot2
“La gramática de los gráficos”

Gráficos con ggplot2

Caso de Estudio

Mg. Jesús Salinas Flores jsalinas@lamolina.edu.pe

Caso de estudio: Identificación de morosos en un producto crediticio

La información a analizar proviene de un producto crediticio de una institución financiera.

El objetivo es predecir si un **nuevo cliente** de la institución financiera podría ser clasificado como moroso o no moroso.

Se recolectaron datos respecto a 11 atributos registrados de un cliente al momento que este se afilia a dicha institución financiera.

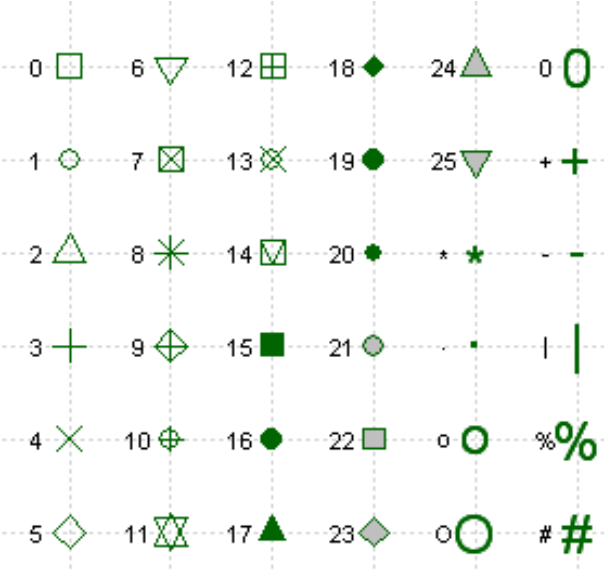
Riesgo_morosidad.sav

edad	sexo	nrodepen	fonopart	fonolab	autovalue	esaval	tieneaval	antiguedad	tiporenta	dpto	morosidad
44	Femenino	3	Si	Si	No	No	No	63	Fijo	Lima	No Moroso
77	Femenino	4	Si	Si	No	No	Si	62	Fijo	Lima	Moroso
59	Femenino	5	No	No	No	Si	No	59	Fijo	Lima	Moroso
35	Femenino	5	No	Si	No	No	No	58	Fijo	Lima	Moroso
65	Femenino	0	Si	No	No	Si	No	56	Fijo	Lima	No Moroso
66	Masculino	1	Si	Si	Si	Si	Si	54	Fijo	Lima	No Moroso
73	Masculino	5	Si	No	No	No	No	53	Fijo	Lima	Moroso
73	Femenino	0	No	No	No	Si	Si	55	Fijo	Lima	No Moroso
74	Femenino	0	Si	No	No	Si	No	53	Fijo	Lima	No Moroso
75	Femenino	1	No	No	No	No	No	53	Fijo	Lima	Moroso
52	Masculino	0	Si	No	No	No	No	52	Variable	Lima	No Moroso
76	Masculino	0	Si	Si	Si	No	Si	52	Fijo	Lima	No Moroso

Variables Independientes / Predictoras

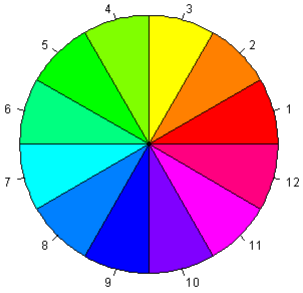
Variable Dependiente / a Predecir / Target

Opciones de shape

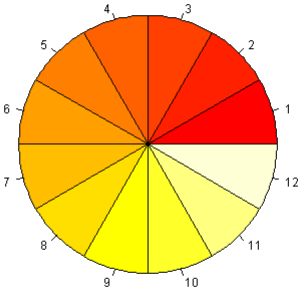


Colores en R

col=rainbow(n)

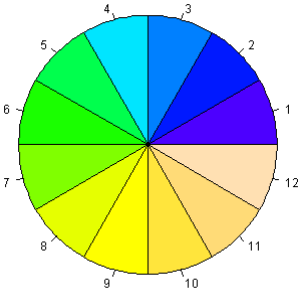


col=heat.colors(n)

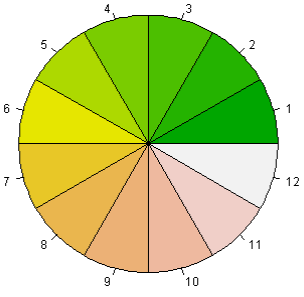


Colores en R

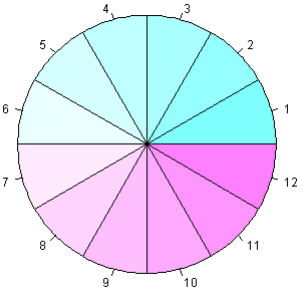
col=topo.colors(n)



col=terrain.colors(n)

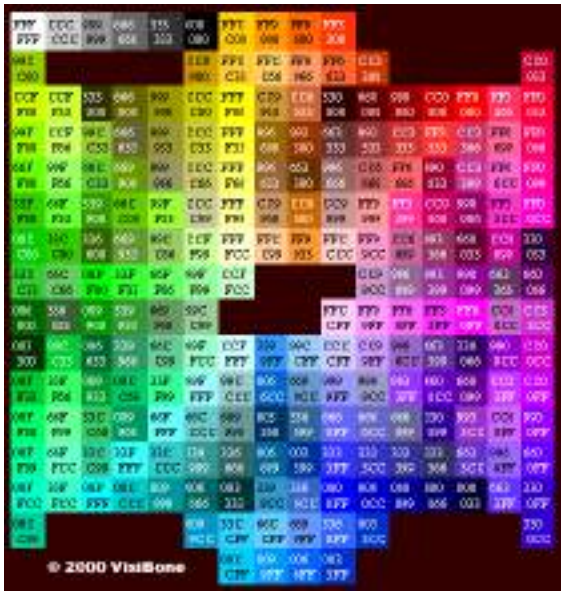


col=cm.colors(n)

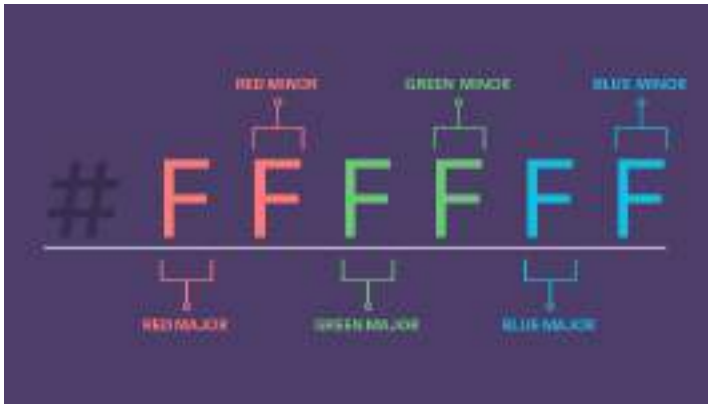


Colores con código hexadecimal

col = “#FFCC00”

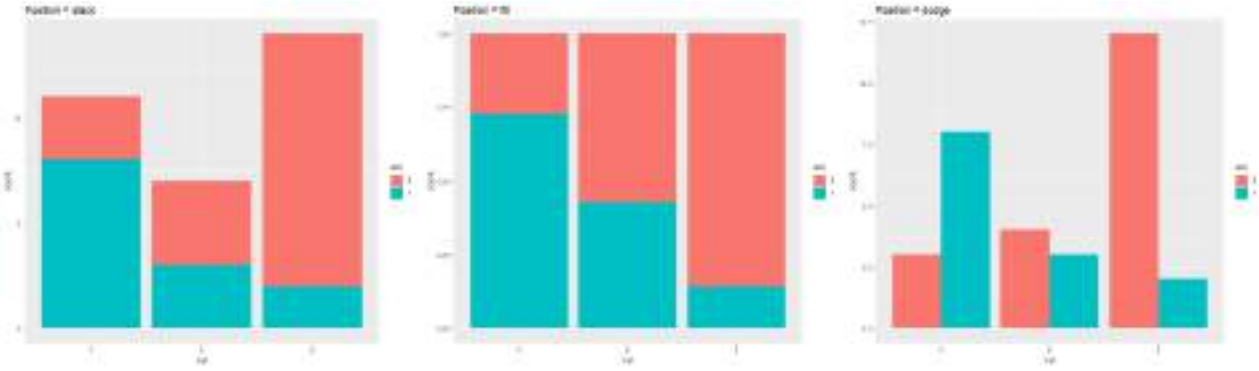


Códigos de color en formato hexadecimal



El formato hexadecimal está en base 16. Se dispone del 0 al 9, y de la A representando al 10 hasta la F representando al 15.

Gráficos con ggplot2



stack fill dodge

Cuando ves a tus alumnos usando el paquete graphics de R en vez de ggplot2

Data StoryTelling

Mg. Jesús Salinas Flores

jsalinas@lamolina.edu.pe

¿Qué es el Storytelling y el DataStorytelling?

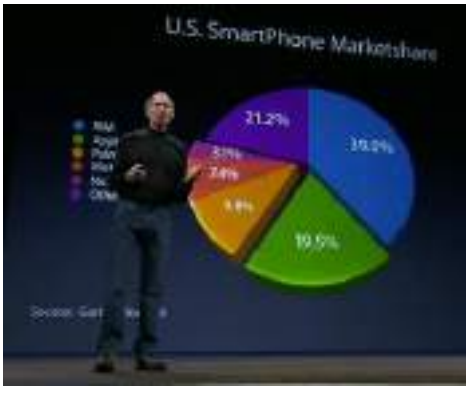
"El «Storytelling» es el arte y la manera de contar una historia dándoles un toque humano y el «Data Storytelling» es el arte y la manera de contar una historia apoyándose en los datos, cifras, o hechos, ya que si tomamos una gráfica o una simple curva ésta no cuenta ninguna historia. Entonces, el «Data Storytelling» ayuda a crear una historia que permitirá explicar las cifras y los datos".

Fuente: https://www.decideo.com/Aprenda-a-narrar-historias-a-partir-de-los-datos-con-el-Data-Storytelling_a679.html

37

Data StoryTelling

Representantes del Data StoryTelling



Manipulación de datos con dplyr

Mg. Jesús Salinas Flores

jsalinas@lamolina.edu.pe

Funciones del dplyr

- **select():** selecciona columnas de los datos
- **filter():** filtra filas que cumplen con el criterio
- **count():** cuenta observaciones
- **group_by ():** agrupa diferentes observaciones
- **summarise():** resume cualquiera de las funciones anteriores
- **arrange():** ordena los datos por columna columna en orden ascendente o descendente
- **mutate():** crea nuevas columnas conservando las variables existentes
- **join():** realiza left, right, full y inner join en R

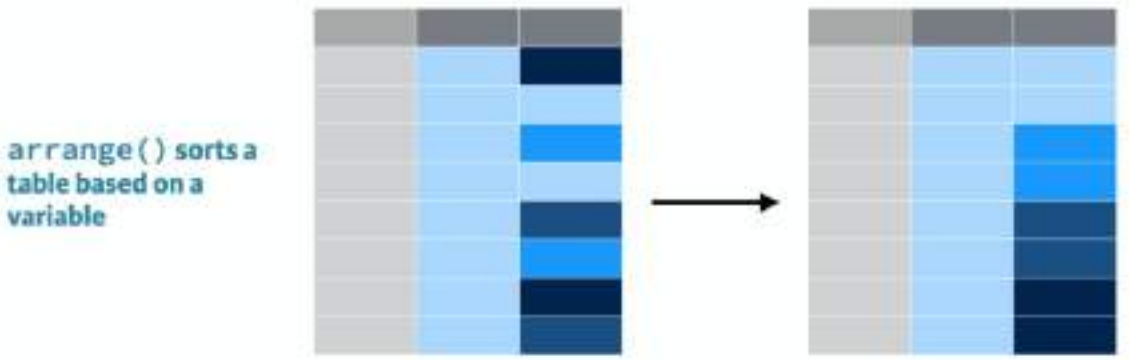
41

filter()

Tomado de: https://www.datacamp.com

42

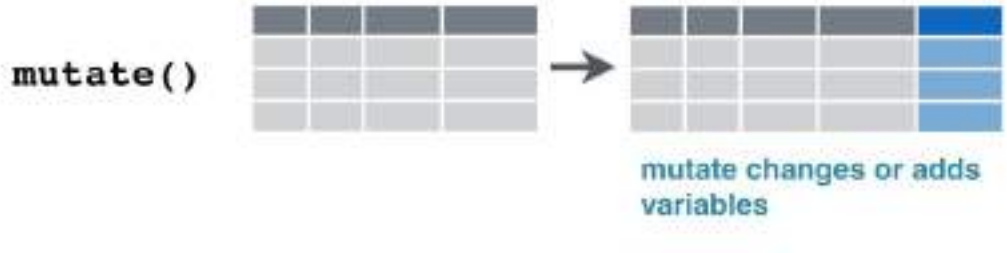
arrange()



Tomado de: https://www.datacamp.com

43

mutate()



Tomado de: https://www.datacamp.com

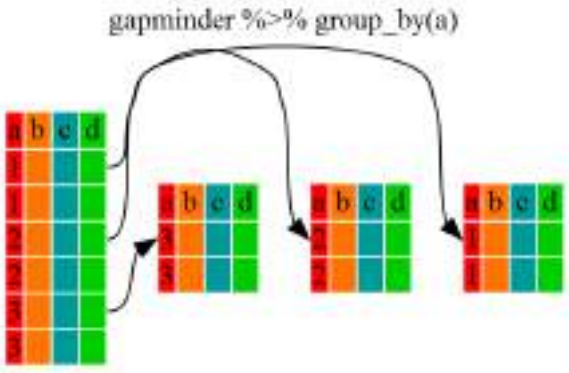
44

summarize()

Tomado de: https://www.datacamp.com

45

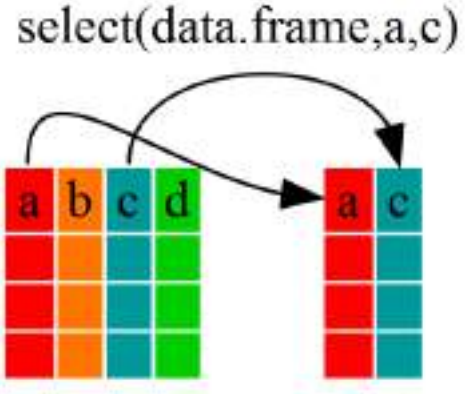
group_by()



Tomado de: https://swcarpentry.github.io/r-novice-gapminder-es/13-dplyr/

46

select()



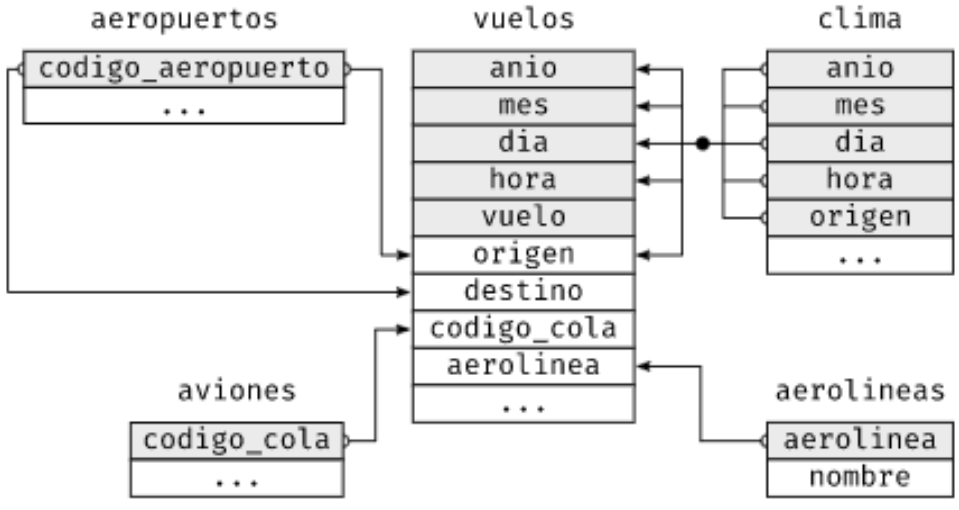
Tomado de: https://swcarpentry.github.io/r-novice-gapminder-es/13-dplyr/

47

Datos relacionales con dplyr

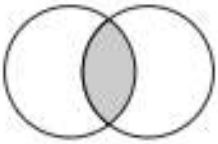
Mg. Jesús Salinas Flores

jsalinas@lamolina.edu.pe

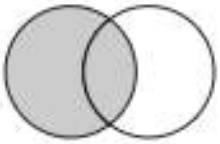


Fuente: Versión en español del libro R para Ciencia de Datos. Grolemund, G. & Wickham, H. 2017. O'Reilly Media

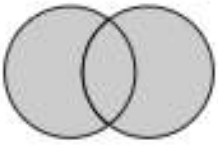
49



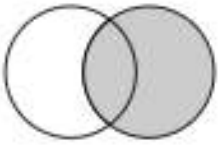
Unión interior
`inner_join(x,y)`



Unión por la izquierda
`left_join(x,y)`



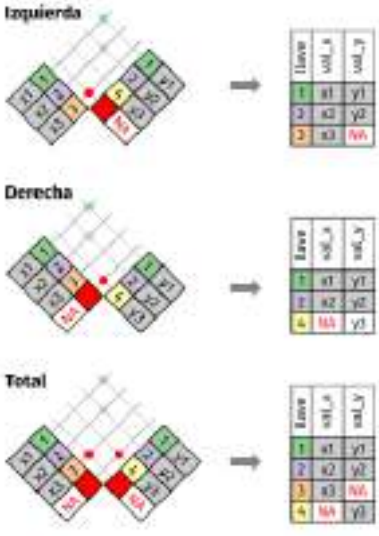
Unión total
`full_join(x,y)`



Unión por la derecha
`right_join(x,y)`

Fuente: Versión en español del libro R para Ciencia de Datos. Grolemund, G. & Wickham, H. 2017. O'Reilly Media

50



Fuente: Versión en español del libro R para Ciencia de Datos. Grolemund, G. & Wickham, H. 2017. O'Reilly Media

R Markdown

Mg. Jesús Salinas Flores

jsalinas@lamolina.edu.pe

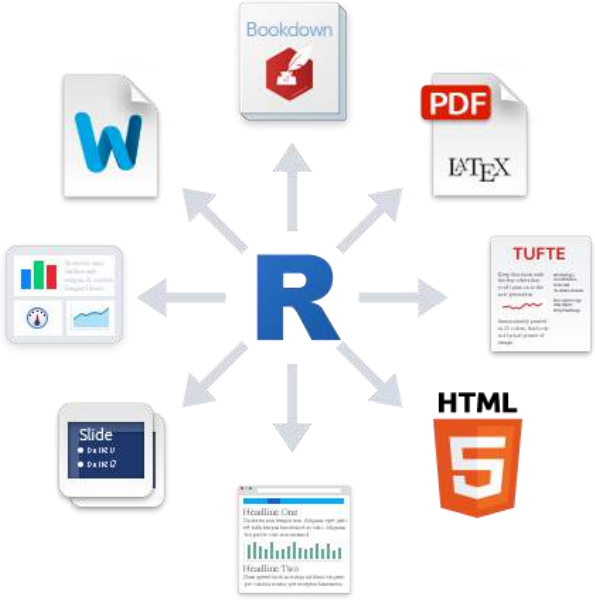
R Markdown

- R Markdown es un formato que permite de manera fácil crear de documentos, presentaciones dinámicas e informes de R.
- Markdown es un formato de sintaxis simple para crear documentos en HTML, PDF, y Word.

R Markdown

Markdown fue creado en 2004 por John Gruber, con el objetivo de permitir que las personas "escriban usando un formato de texto en lenguaje sencillo y fácil de leer" y luego poder convertirlo a un formato HTML

66



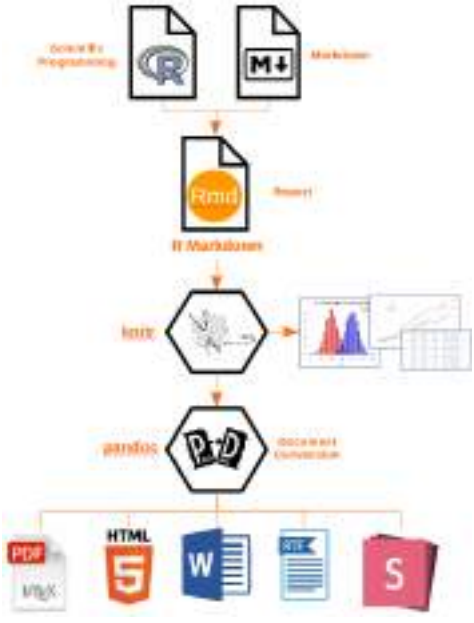
Fuente: rmarkdown.rstudio.com (http://rmarkdown.rstudio.com/images/RMarkdownOutputFormats.png)

67

Conjunto de datos ordenado

- Cuando se hace click en *knit* el documento (knit en español significa tejer), R Markdown envía el .Rmd a *knitr*, http://yihui.name/knitr/, que ejecuta todos los bloques de código y crea un nuevo documento markdown (.md) que incluye el código y su output.
- El archivo markdown generado por *knitr* es procesado entonces por pandoc, http://pandoc.org/, que es el responsable de crear el archivo terminado.
- La ventaja de este flujo de trabajo en dos pasos es que se puede crear un muy amplio rango de formatos de salida.

68



Fuente: http://applied-r.com/project-reporting-template/

69

Opciones en los chunk

Opción	Efecto
include = F	no incluye los resultados, aunque si eval=TRUE si los evalúa
echo = F	solo muestra los resultados, no el código
message = F	No muestra los resultados de salida
warning = F	dirige lo warnings a consola y no al documento
eval = F	No evaluará las expresiones

70

Opciones en las figuras

Opción	Posibles valores	Efecto
fig.height.	Numérico, pulgadas	La altura de la imagen en pulgadas
fig.width	Numérico, pulgadas	El ancho de la imagen en pulgadas
fig.align	"left", "right" o "center"	La alineación de la imagen en el reporte

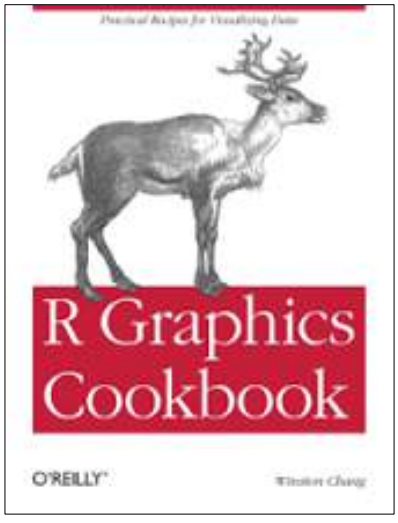
Más opciones en http://yihui.name/knitr/options

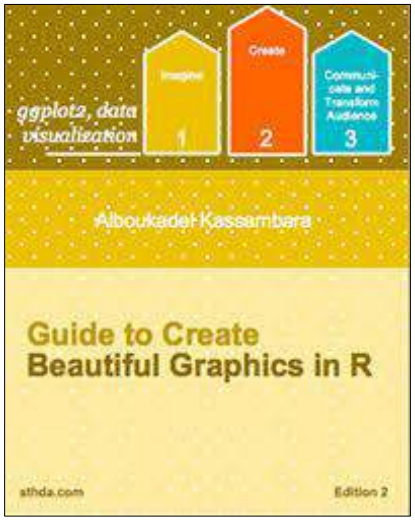
71

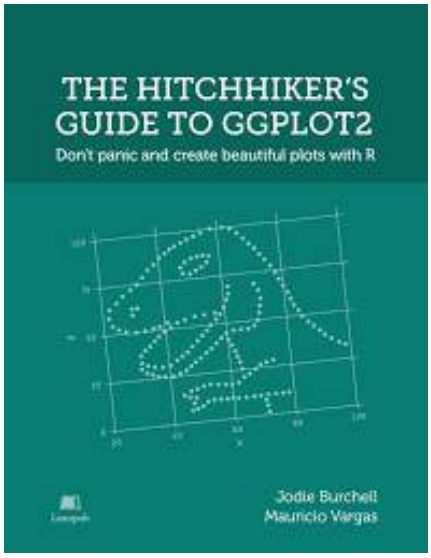
Bibliografía

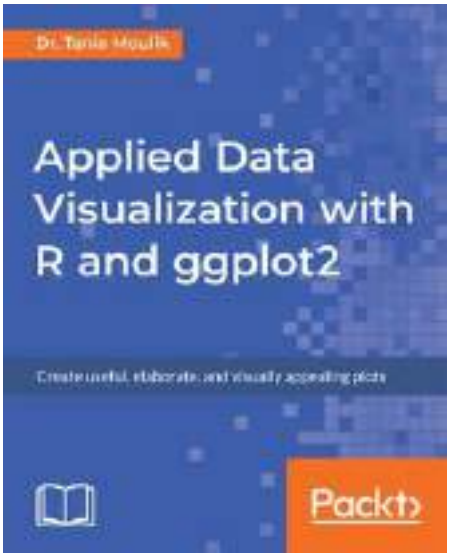
Mg. Jesús Salinas Flores

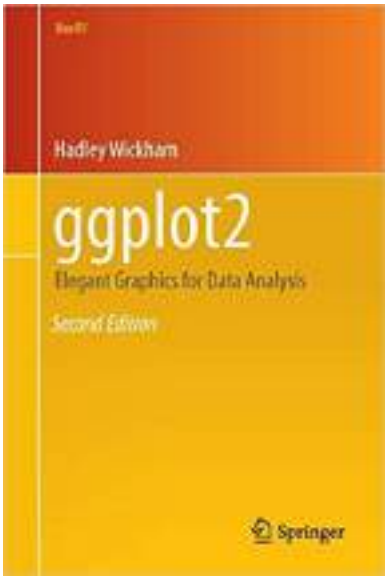
jsalinas@lamolina.edu.pe

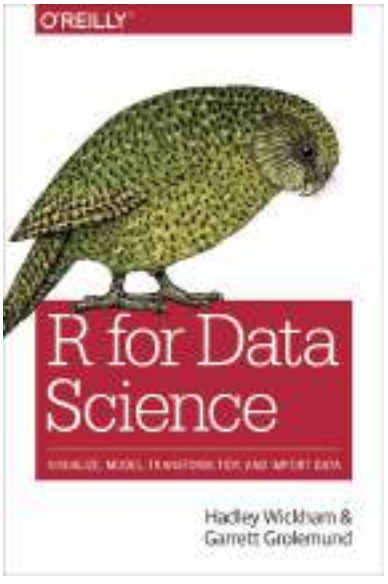


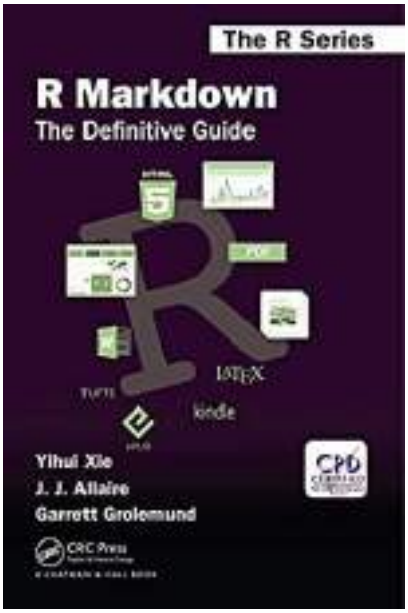


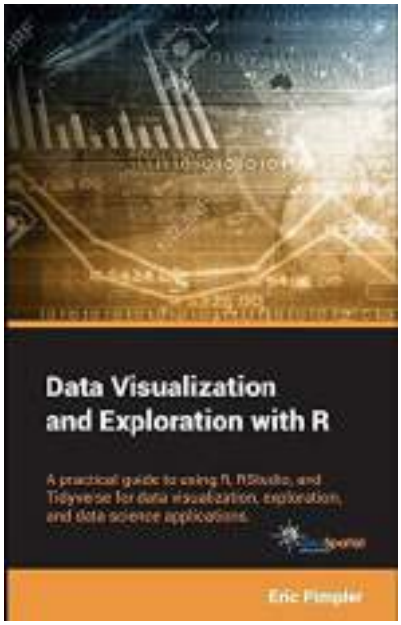

THE HITCHHIKER'S GUIDE TO GGPLOT2
Don't panic and create beautiful plots with R
Jodie Burchell
Mauricio Vargas


Dr. Tania Moulik
Applied Data Visualization with R and ggplot2
Packt


Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Second Edition
Springer


O'REILLY
R for Data Science
Hadley Wickham &
Garrett Grolemund


The R Series
R Markdown
The Definitive Guide
Yihui Xie
J. J. Allaire
Garrett Grolemund
CRC Press


Data Visualization
and Exploration with R
Eric Pimpler