

Analítica de Datos

Introducción a la analítica de datos


Carlos Cardona Andrade

Tabla de contenido

1. Introducción
2. Estructura del Curso
3. Analítica de Datos
4. ¿Por qué R?
5. Fundamentos básicos de R

Introducción

Introducción personal

- Carlos Cardona Andrade
 -  carlos.cardonaa@javeriana.edu.co
- Ustedes:
 - Nombre
 - Hobbie/Algo que les guste mucho hacer
 - Experiencia con lenguajes de programación

Evaluación

Componente	Porcentaje
2 × parciales (15% cada uno)	30%
Tareas y quices	30%
1 × Proyecto Caso Final	40%

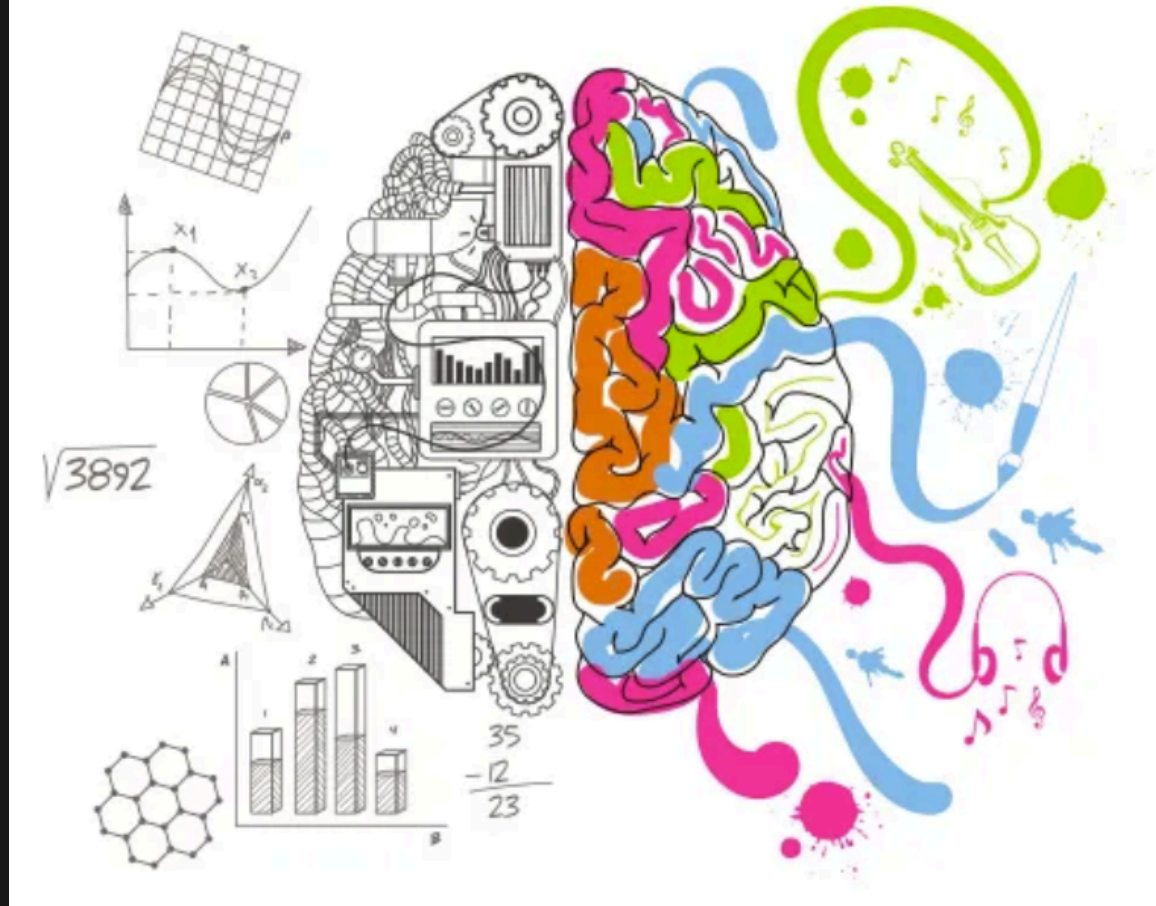
- Parcial 1: 21 de Agosto
- Parcial 2: 9 de Octubre

Resultados de Aprendizaje Esperados

Después de completar el curso, el estudiante será capaz de:

1. Interpretar correctamente los estadísticos básicos, gráficos más comunes, así como los resultados de una prueba de hipótesis y una regresión.
2. Aplicar los principios básicos de visualizaciones efectivas.
3. Distinguir entre causalidad y correlación.
4. Construir un modelo de regresión lineal para el apoyo de la toma decisiones.
5. Evaluar la aplicación e interpretación de técnicas de la analítica de datos a situaciones de negocio específicas.
6. Diseñar un experimento sencillo.

Data vs Gut



Why Insights from Data Analytics are Underestimated in Companies?

Las personas encargadas de tomar decisiones deben evolucionar junto con un entorno que se caracteriza por un flujo continuo de información.

Estructura del curso

Estadística Descriptiva

- Comenzamos suponiendo que ya tenemos los datos.
- Exploraremos formas de resumir datos usando R:
 1. Visualización de datos.
 2. Construcción de estadísticas descriptivas.

Estadística inferencial

- Aprenderemos a construir **intervalos de confianza**, que nos dicen con qué tanta precisión estamos calculando un promedio.
- Aprenderemos a elaborar **pruebas de hipótesis**, que nos dicen si dos grupos son distintos en cuanto a alguna característica.
- Usaremos **regresiones**, que explotan las relaciones entre dos o más variables para predecir y para prescribir.

Storytelling with Data

When you combine the right visuals and narrative with the right data, you have a data story that can influence and drive change

Uso de Casos Harvard

- Utilizaremos casos de la vida real, con datos reales, para elaborar nuestro análisis.
- Es importante hacer el curso llamado Case Companion (Brightspace).
- Hay que leerlos previo a cada sesión (ver syllabus).
- Discutiremos el caso, definiremos una estrategia para resolverlo, y ustedes harán el análisis en R.

Analítica de Datos



Las decisiones las tomaban HiPPOs (Highest Paid Person in Organization) basados en intuición.

¿Qué es la analítica de datos para los negocios?

Es la aplicación de tecnologías informáticas y herramientas estadísticas que permiten analizar datos relevantes para la toma de decisiones dentro una empresa/organización.

Preguntas a responder con Analítica de Negocios

- ¿Cuáles son las características demográficas de mis clientes? ¿Tengo distintos grupos de clientes con distintos perfiles?
- ¿Cuál será el costo esperado de atender a un paciente con ciertas características?
- ¿Cuál es la disposición a pagar de un cliente por mi producto?
- ¿Cuál sería el efecto en ventas de introducir un torneo entre vendedores?
- ¿Cuál es el efecto de un año más de antigüedad laboral sobre el desempeño de los empleados?
- ¿Cuál diseño de mi página web es más efectivo?
- ¿Cuál campaña online es mejor?

Ejemplos desde la Universidad Javeriana

- Evaluar el desempeño de profesores, basándose en evaluaciones de estudiantes.
- Predecir, basándose en características demográficas y desempeño, si un alumno caerá en prueba académica o no, con el objetivo de intervenir a tiempo.
- Predecir demanda por un curso en particular, basándose en demanda en el semestre anterior y distintas características del curso.
- Comparar el efecto de distintas modalidades (presencial, virtual, combinada, alternancia) sobre los objetivos de aprendizaje.
- Evaluar la pertinencia de un nuevo método de enseñanza.

Aspectos de la analítica

Analítica descriptiva: visualizar y tabular datos que ya se tienen para entender cambios o la situación actual de un negocio (básicamente describir la información que se tiene).

- ¿Han crecido las ventas después de la introducción de un nuevo plan de mercadeo?
- ¿Qué regiones son las más débiles en ventas?
- ¿Cuáles son las características de mis clientes?
- ¿Cuánto ha variado el precio de las acciones de Coca Cola en los últimos 5 años?

Aspectos de la analítica

Analítica predictiva: predecir qué pasará, explotando relaciones entre variables. Para predecir, es suficiente que las variables que estamos estudiando estén correlacionadas, no es necesario que una cause otra.

- Puntaje de crédito (probabilidad de morosidad)
- Retención de clientes (probabilidad de perderlo, intervención temprana)
- ¿Cuánto me costará atender a un paciente de acuerdo a sus características?
- Detección de fraudes (probabilidad de que sea fraudulenta)
- Protección de infantes (probabilidad de que el niño sea maltratado)
- Predicción de inventarios

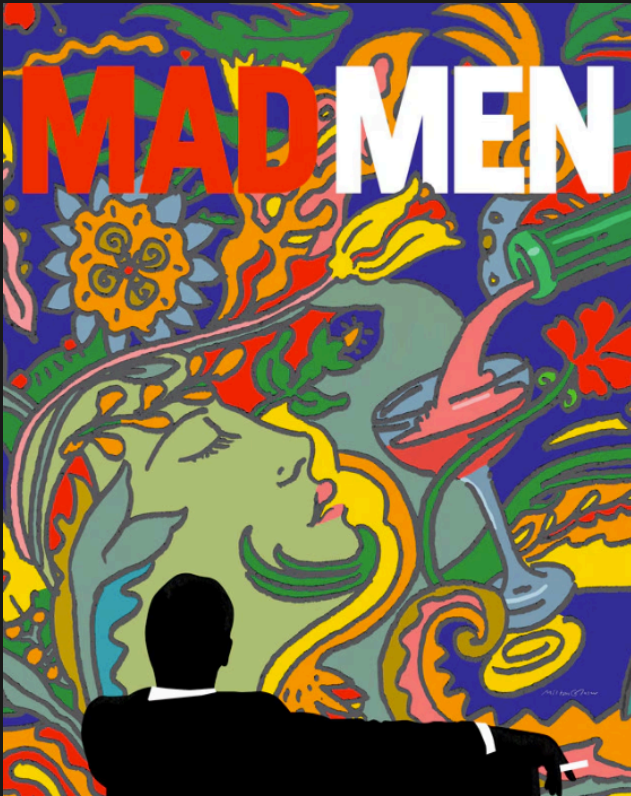
Aspectos de la analítica

Analítica prescriptiva: se enfoca en hacer uso de la analítica para determinar cuál es la mejor decisión que se podría tomar. Aquí es importante determinar causalidad. Muchas veces se hace con experimentos.

- ¿Cuál de los diseños de página deberíamos implementar?
- ¿Si subimos el precio aumentarían las ganancias?
- ¿Sería efectivo implementar un sistema de compensación basado en desempeño?

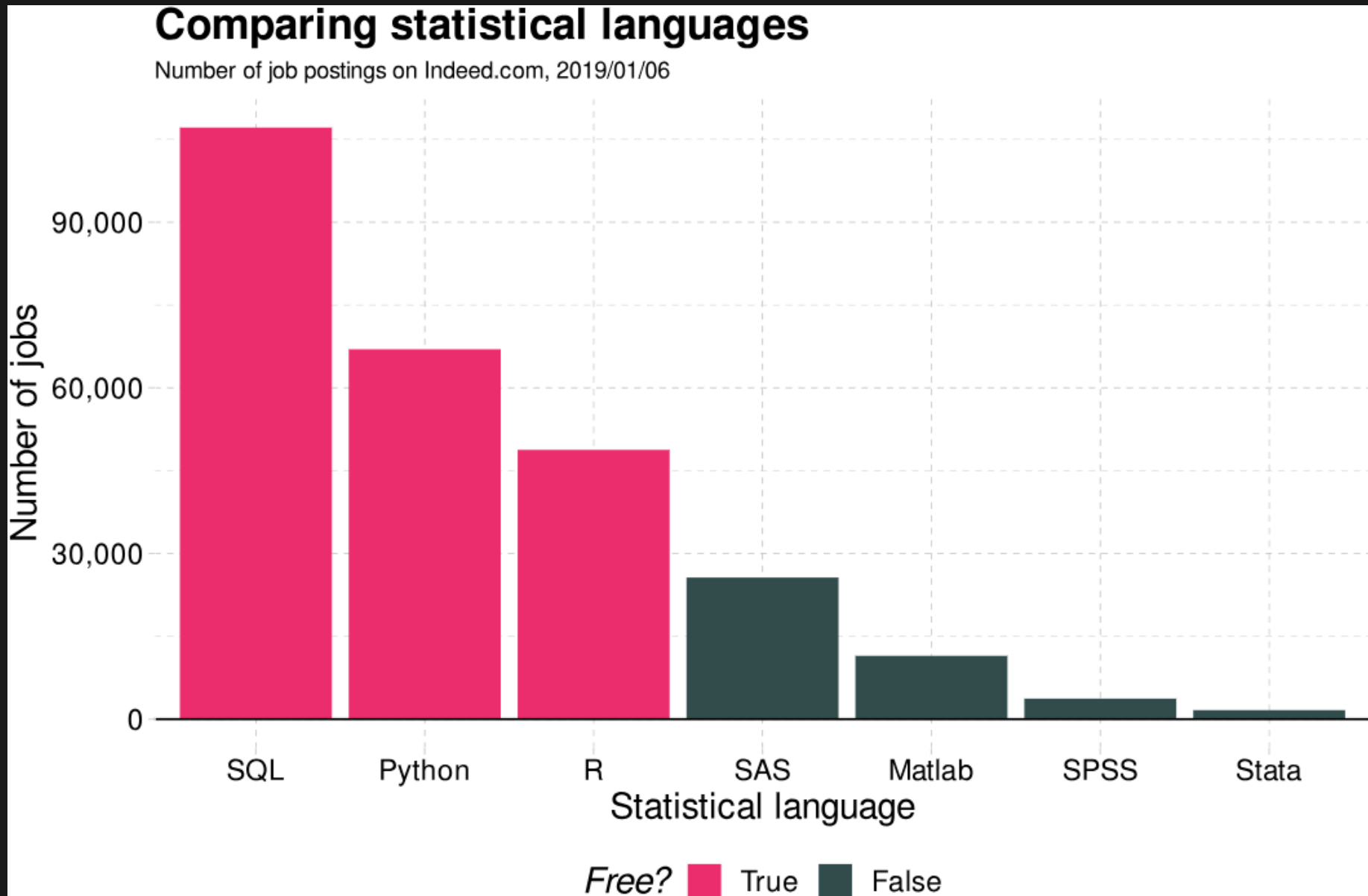
Ejemplos con Big Data

- How to use data to make a hit TV show?
- The Social Dilemma, documental en Netflix



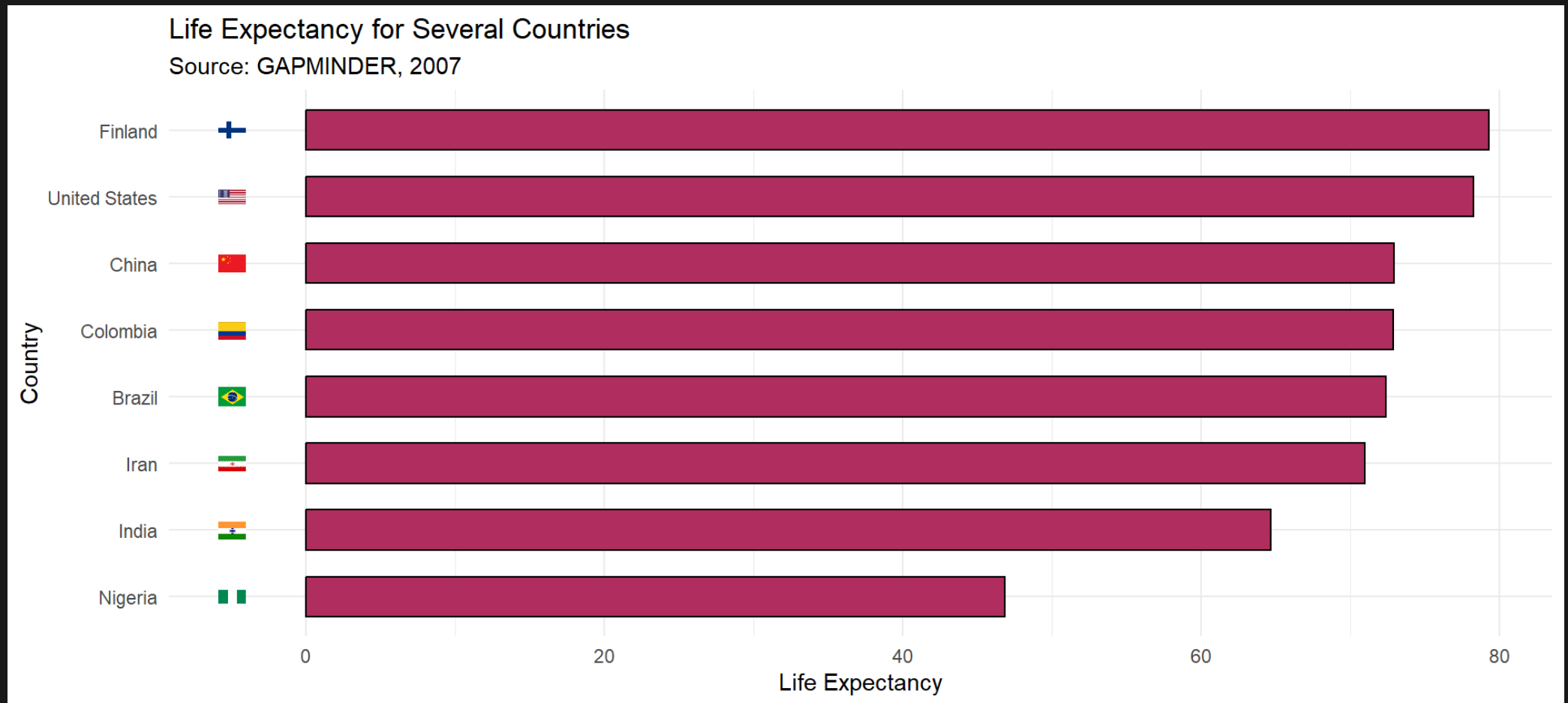
¿Por qué R?

Número de Empleos



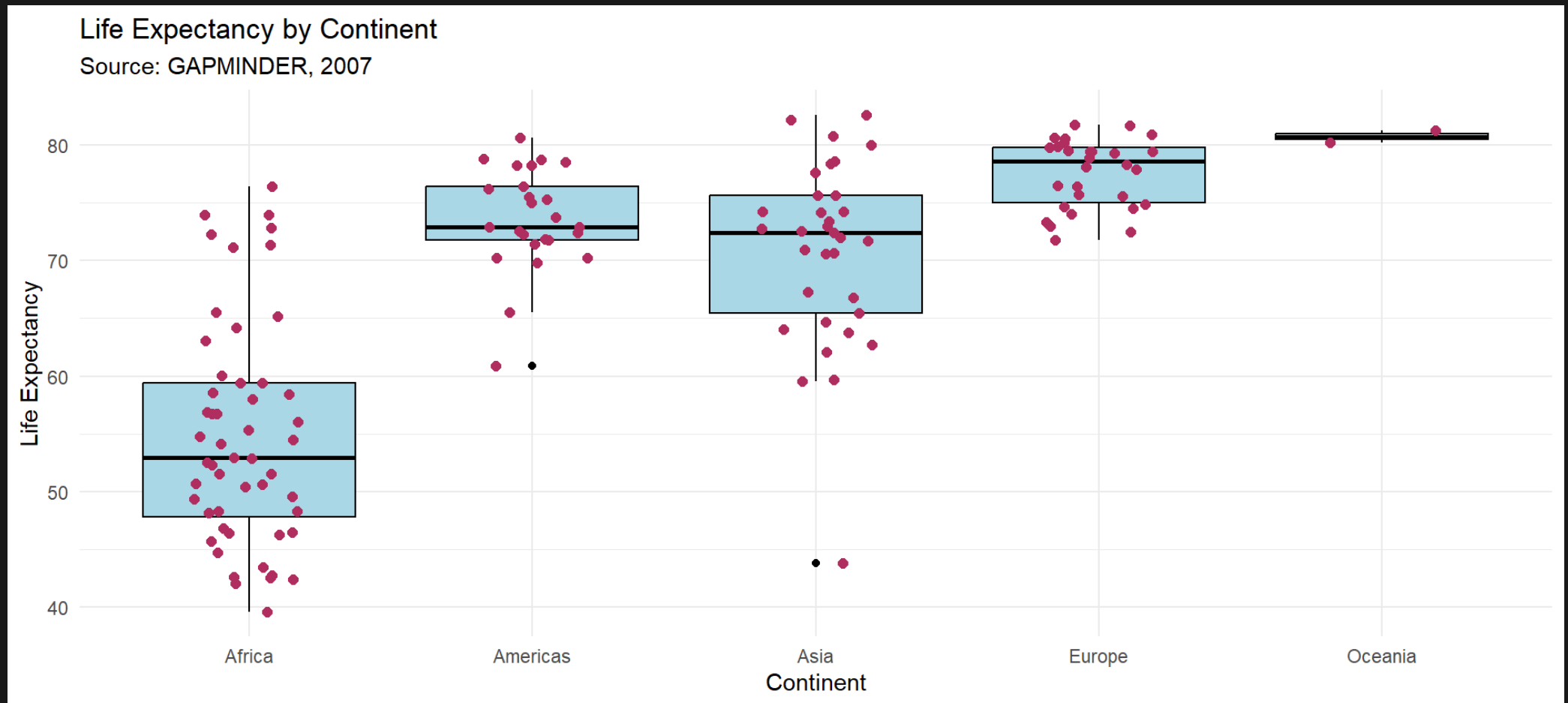
Visualización de Datos en R

► Código de la gráfica



Visualización de Datos en R

► Código de la gráfica

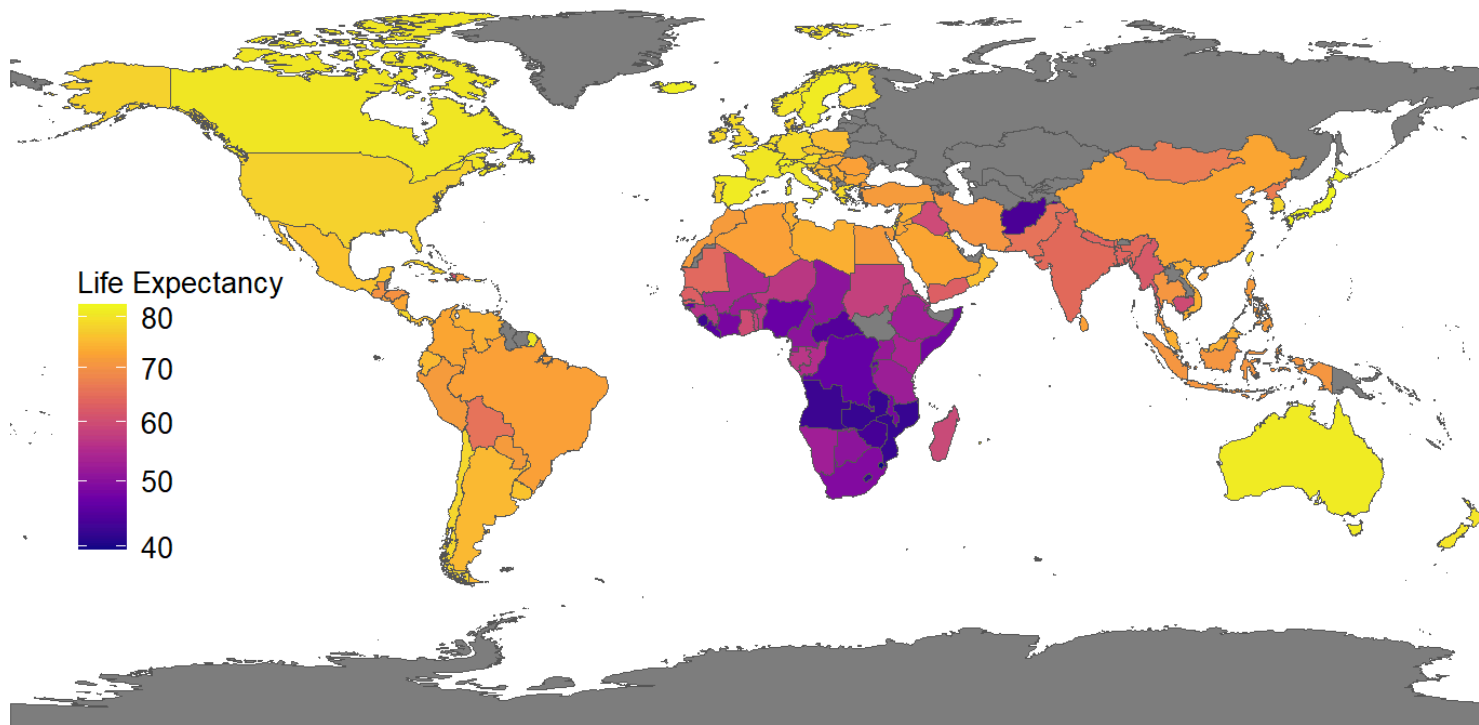


Visualización de Datos en R

► Código del mapa

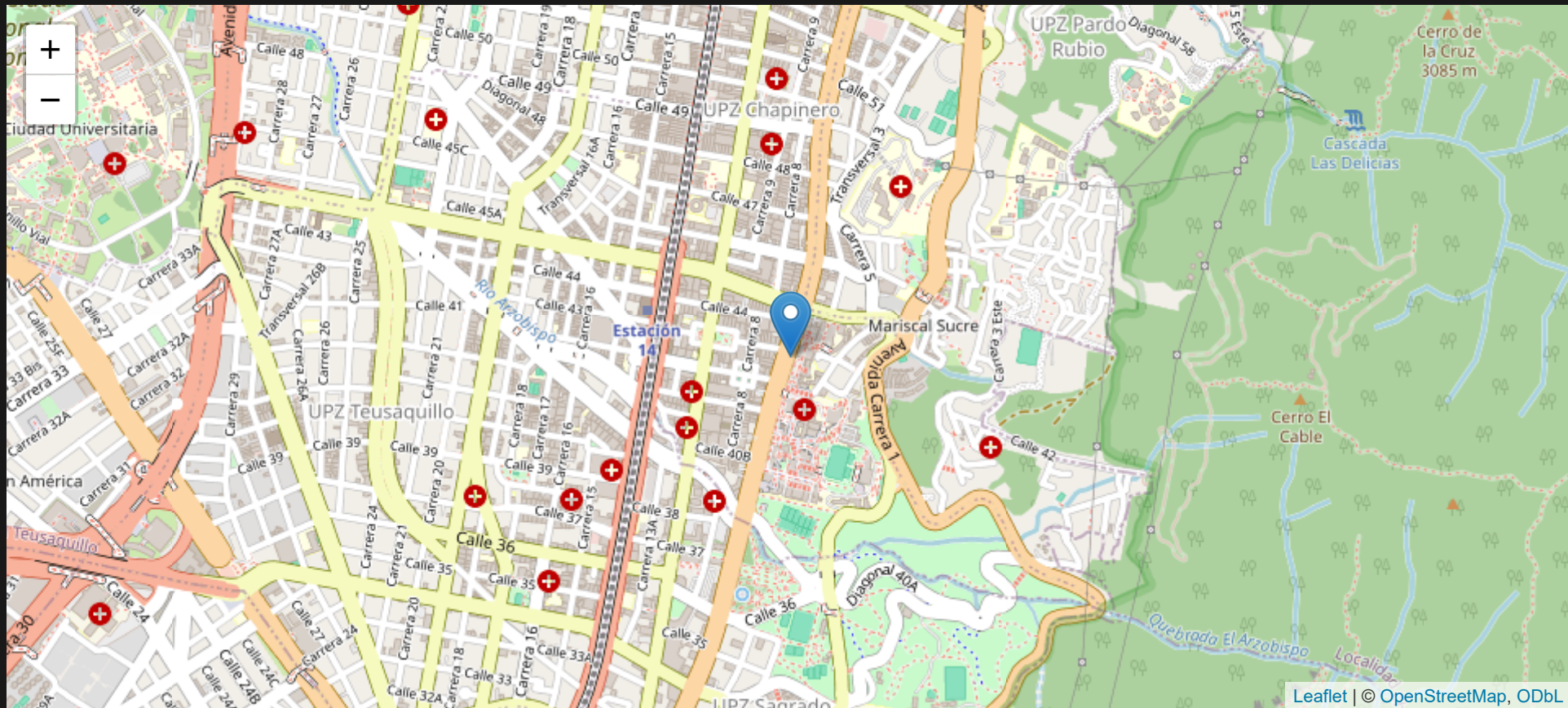
Life Expectancy by Country

Source: GAPMINDER, 2007



Visualización de Datos en R

► Código del mapa



Algunos recursos útiles

- *R for Data Science* de Hadley Wickham and Garrett Grolemund.
- *Data Visualization: A Practical Guide* de Kieran Healy.
- Swirl Website
- Data Science Course de Grant McDermott
- Introducción a la ciencia de datos de Rafael Irizarry

Fundamentos básicos de R

Instalación y pasos a seguir

1. Descargar [R](#).
2. Descargar [RStudio](#).
3. ¿Tienen la versión más reciente de R?

```
1 version$version.string
```

```
[1] "R version 4.1.1 (2021-08-10) "
```

4. ¿Tienen la versión más reciente de RStudio? (La [versión previa](#) también sirve.)

```
1 RStudio.Version()$version
```

```
2 # Requiere la versión interactiva pero debería mostrar algo como "[1] '2023.3.0.386' "
```

Algunos fundamentos básicos de R

1. Todo es un objeto.
2. Todo objeto tiene un nombre.
3. Todo se opera usando funciones.
4. Las funciones existen dentro de paquetes (i.e. “libraries”), aunque ustedes pueden escribir sus propias funciones.

Puntos 1. y 2. pueden ser resumidos como un enfoque de **programación orientada a objetos** (OOP). Esto puede sonar super abstracto ahora, pero veremos *muchos* ejemplos en las próximas semanas que harán todo más claro.

Uso de la consola

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The main editor window displays a presentation slide titled "Working interactively: Using the Console". The slide content includes a list of features and a recommendation to set the language to English. The console window at the bottom, highlighted with a red border, shows the execution of R commands and their output.

Slide Content:

- Working interactively: Using the Console
- usually on the left
- startup message showing R version and license information
- input prompt "> ..." waiting for your commands
- recommendation: set to English language (depends on OS) (<https://stackoverflow.com/questions/13575188/how-to-change-language-settings-in-r#13575413>)

Console Output:

```
~/Development/big_data_tutorium/src/slides/ >  
> head(airquality)  
  Ozone Solar.R Wind Temp Month Day  
1    41     190   7.4   67     5   1  
2    36     118   8.0   72     5   2  
3    12     149  12.6   74     5   3  
4    18     313  11.5   62     5   4  
5    NA      NA  14.3   56     5   5  
6    28      NA  14.9   66     5   6  
> mean(airquality$Ozone)  
[1] NA  
> mean(airquality$Ozone, na.rm = TRUE)  
[1] 42.12931  
> |
```

File Explorer:

Name	Size	Modified
..		
.Rhistory	1.1 KB	Aug 10, 2018, 6:17 PM
01intro-figure		
01intro.html	1.4 MB	Aug 13, 2018, 10:17 AM
01intro.Rpres	7.6 KB	Aug 13, 2018, 10:46 AM
slides.Rproj	205 B	Aug 13, 2018, 10:16 AM
01intro.md	7.6 KB	Aug 13, 2018, 10:46 AM

Aritmética básica

R posee una calculadora poderosa y reconoce todas las operaciones estándar de aritmética:

```
1 1+2 ## Adición
```

```
[1] 3
```

```
1 6-7 ## Resta
```

```
[1] -1
```

```
1 5/2 ## División
```

```
[1] 2.5
```

```
1 2^3 ## Exponencial
```

```
[1] 8
```

```
1 2+4*1^3 ## Order estándar de precedencia (`*` antes de `+`, etc.)
```

```
[1] 6
```

Sintaxis de R

Algunas reglas generales: 1. Cada línea es una sentencia (“comando”), varias sentencias se evalúan de arriba a abajo.

```
1 c<-a+b  
2 d<-sqrt(c)
```

Excepción: Si una expresión no está cerrada (véase la regla de la paréntesis más abajo), puede abarcar varias líneas:

```
1 a*(b  
2 +c  
3 +d)
```

Sintaxis de R

2. Los espacios suelen ignorarse.

Todos son equivalentes:

```
1 a+b  
2 a + b  
3 a  +  b
```

Los espacios y las tabulaciones sirven para hacer nuestro código más leíble.

Sintaxis de R

3. Las expresiones deben estar cerradas.

Existen diferentes caracteres especiales que marcan el principio y el final de algo, por ejemplo, el principio y el final de una cadena de caracteres o de una expresión:

```
1 "hello world"  
2 a*(b+c)  
3 x[1]
```

Las sentencias más complejas contienen expresiones anidadas. Las expresiones anidadas se evalúan de dentro a fuera.

```
1 y[c(1, 3)]
```

Para cada paréntesis abierta, comilla, etc. debe haber una contrapartida de cierre en el orden correcto. Esto sería incorrecto:

```
1 y[c(1, 3)]
```

Error: <text>:1:9: unexpected ']'

```
1: y[c(1, 3]  
      ^
```

Sintaxis de R

4. Coma y puntos

Las comas dividen cosas: Principalmente argumentos (parámetros u objetos) de funciones.

```
1 log(x, 5)
```

La coma no se puede utilizar para agrupar dígitos en números grandes:

```
1 population <- 3,350,000
```

```
Error: <text>:1:16: unexpected ','
```

```
1: population <- 3,  
                  ^
```

Se utiliza un punto como punto decimal:

```
1 3.1415
```

Paquetes en R

The screenshot displays the RStudio interface with a presentation slide titled "RStudio" on the right and a list of installed packages in the bottom right pane.

RStudio Presentation Content:

- RStudio is an **Integrated development environment (IDE)** for R
- it's a comfortable **interface** to R
- analogy: if R is the engine, then RStudio is the car around it
- offers:
 - interactive console
 - script editor with error checking
 - package manager
 - data, plot and file viewers

Installed Packages (User Library):

Name	Description	Version
abind	Combine Multidimensional Arrays	1.4-5
arules	Mining Association Rules and Frequent Itemsets	1.6-1
assertthat	Easy Pre and Post Assertions	0.2.0
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.2
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.66.0-1
bindr	Parametrized Active Bindings	0.1.1
bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
bit	A Class for Vectors of 1-Bit Booleans	1.1-14
bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
bitops	Bitwise Operations	1.0-6
blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.1.1
broom	Convert Statistical Analysis Objects into Tidy Tibbles	0.5.0
carData	Companion to Applied Regression Data Sets	3.0-1
caret	Classification and Regression Training	6.0-80
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
cli	Helpers for Developing Command Line Interfaces	1.0.0
colorspace	Color Space Manipulation	1.3-2

Paquetes en R

- Los paquetes (también conocidos como “bibliotecas”) amplían la funcionalidad de R.
- A la derecha, la pestaña “Paquetes” permite ver, instalar y actualizar paquetes de R desde CRAN

```
1 install.packages("tidyverse")  
2 library(tidyverse)
```

- Si olvida cargar un paquete, se encontrará con errores como éstos:

```
1 wb_search()
```

```
Error in wb_search(): could not find function "wb_search"
```

Ayuda en R

Para obtener más información sobre una función (con nombre) u objeto en R, consulte la documentación de “help”. Por ejemplo:

```
1 help(plot)
```

O, simplemente, solo usen `?`:

```
1 # Esta es la manera más común de usar la ayuda.  
2 ?plot
```

Nota 1: Comentarios en R se demarcan con `#`.

- Click **Ctrl+Shift+c** en RStudio para comentar/borrar el comentario de secciones completas de código subrayado.

Nota 2: Vean la sección *Examples* al final del archivo de ayuda.

- Pueden correr los ejemplos usando la función `example()`. Intenten con: `example(plot)`.

