---

title: "Analítica de Datos"

subtitle: "Introducción a la analítica de datos"

author: "Carlos Cardona Andrade"

format:

revealjs:

theme: dark

slide-number: true

chalkboard:

buttons: false

code-block-height: 250px

preview-links: auto

css: styles.css

---

# Tabla de contenido

1. [Introducción](#introduccion)

2. [Estructura del Curso](#struct)

3. [Analítica de Datos](#andatos)

4. [¿Por qué R?](#por-que-r)

5. [Fundamentos básicos de R](#intro-r)

# Introducción {#introduccion}

## Introducción personal

```{r}

#| echo: false

library(fontawesome)

```

- Carlos Cardona Andrade

- `r fa('envelope')` [carlos.cardonaa@javeriana.edu.co](mailto: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 {.smaller}

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.

#

![](images/datavsgut.png){fig-align="center" width=60%}

<div style="text-align: center; font-size: 0.8em;">[Why Insights from Data Analytics are Underestimated in Companies?](https://medium.com/@ytxia/why-insights-from-data-analytics-are-underestimated-in-companies-7f89d3aecf4a)

<div style="text-align: center; font-size: 0.9em;">Las personas encargadas de tomar decisiones deben evolucionar junto con un entorno que se caracteriza por un flujo continuo de información.

# Estructura del curso {#struct .smaller}

## 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.

## 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 {#andatos}

#

![](images/big\_data\_boss.png){fig-align="center" width=60%}

<div style="text-align: center; font-size: 0.9em;">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 {.smaller}

::: incremental

- ¿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 {.smaller}

::: incremental

- 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 {.smaller}

`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).

::: incremental

- ¿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 {.smaller}

`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.

::: incremental

- 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 {.smaller}

`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.

::: incremental

- ¿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

![](images/madmen.png){.absolute top="240" left="50" width=30%}

![](images/twitter.png){.absolute top="240" right="120" width=40%}

- [How to use data to make a hit TV show?](https://www.youtube.com/watch?v=vQILP19qABk)

- [The Social Dilemma](https://www.nytimes.com/2020/09/09/movies/the-social-dilemma-review.html), documental en Netflix

# ¿Por qué R? {#por-que-r}

## Número de Empleos

![](images/R\_jobs.png){fig-align="center" width=60%}

## Visualización de Datos en R

```{r}

#| code-fold: true

#| code-summary: "Código de la gráfica"

#| echo: true

#| fig-width: 10

#| fig-height: 4.5

library(dplyr)

library(tidyverse)

library(gapminder)

library(countrycode)

library(ggimage)

library(rnaturalearth)

library(rnaturalearthdata)

# Damos un nombre a la base de datos "gapminder"

GAPMINDER<-gapminder

GAPMINDER<-filter(GAPMINDER, year == max(year))

GAPMINDER$iso2c<-countryname(GAPMINDER$country, destination = 'iso2c')

selected\_countries <- GAPMINDER %>%

filter(country %in% c('United States', 'China', 'Colombia', 'Brazil','Nigeria', 'Finland', 'Iran', 'India'))

selected\_countries$iso2c<-countryname(selected\_countries$country, destination = 'iso2c')

plot <- selected\_countries %>% ggplot(aes(x= reorder(country, lifeExp),y=lifeExp))+

geom\_flag(y = -5, aes(image = iso2c)) +geom\_bar(stat = "identity", color="black", fill = "maroon", width=0.6) +

labs(title = "Life Expectancy for Several Countries ",

subtitle = "Source: GAPMINDER, 2007 ",

x = "Country",

y = "Life Expectancy")+ coord\_flip() +

expand\_limits(y = -5)

plot+ theme\_minimal()

```

## Visualización de Datos en R

```{r}

#| code-fold: true

#| code-summary: "Código de la gráfica"

#| echo: true

#| fig-width: 10

#| fig-height: 4.5

box\_plot <- ggplot(GAPMINDER, aes(x = continent, y = lifeExp)) +

geom\_boxplot(fill = "lightblue", color = "black") +

geom\_jitter(width = 0.2, size = 2, color = "maroon")+

labs(

title = "Life Expectancy by Continent",

subtitle = "Source: GAPMINDER, 2007",

x = "Continent",

y = "Life Expectancy"

)

box\_plot+ theme\_minimal()

```

## Visualización de Datos en R

```{r}

#| code-fold: true

#| code-summary: "Código del mapa"

#| echo: true

#| fig-width: 10

#| fig-height: 4.5

world <- ne\_countries(scale = "medium", returnclass = "sf")

# Rename the variable using rename()

GAPMINDER <- GAPMINDER %>%

rename(iso\_a2 = iso2c)

JOIN<-left\_join(world,GAPMINDER, by="iso\_a2")

ggplot(data = JOIN) +

geom\_sf(aes(fill = lifeExp)) +

scale\_fill\_viridis\_c(option = "plasma", trans = "sqrt")+

labs(fill = "Life Expectancy",title = "Life Expectancy by Country",

subtitle = "Source: GAPMINDER, 2007") + theme\_bw() +

theme(panel.grid.major = element\_blank(),

panel.border = element\_blank(),

title =element\_text(size=14),

axis.ticks.y = element\_blank(),

axis.text.y = element\_blank(),

axis.ticks.x = element\_blank(),

axis.text.x = element\_blank(),

legend.background = element\_rect(fill = "white"),

legend.title=element\_text(size=10),

legend.key = element\_rect( fill = "white"),

legend.position = c(0.15, 0.45),

legend.key.size = unit(0.6, 'cm'),

legend.text=element\_text(size=10),

legend.spacing.y = unit(0.1, "cm"))

```

## Visualización de Datos en R

```{r}

#| code-fold: true

#| code-summary: "Código del mapa"

#| echo: true

#| fig-width: 10

#| fig-height: 4.5

library(leaflet)

leaflet() %>%

addTiles() %>% # Add default OpenStreetMap map tiles

addMarkers(lng=-74.06454, lat=4.63006, popup="Universidad Javeriana")

```

# Algunos recursos útiles

- [\*R for Data Science\*](https://r4ds.had.co.nz/index.html) de Hadley Wickham and Garett Grolemund.

- [\*Data Visualization: A Practical Guide\*](https://socviz.co/makeplot.html) de Kieran Healy.

- [Swirl](https://swirlstats.com/) Website

- [Data Science Course](https://github.com/uo-ec607/lectures) de Grant McDermott

- [Introducción a la ciencia de datos](http://rafalab.dfci.harvard.edu/dslibro/r-basics.html) de Rafael Irizarry

# Fundamentos básicos de R {#intro-r}

## Instalación y pasos a seguir {.smaller}

1. Descargar [R](https://www.r-project.org/).

2. Descargar [RStudio](https://www.rstudio.com/products/rstudio/download/preview/).

3. ¿Tienen la versión más reciente de R?

```{r}

#| echo: true

version$version.string

```

4. ¿Tienen la versión más reciente de RStudio? (La [versión previa](https://www.rstudio.com/products/rstudio/download/preview/) también sirve.)

```R

RStudio.Version()$version

# Requiere la versión interactiva pero debería mostrar algo como "[1] ‘2023.3.0.386’"

```

## Algunos fundamentos básicos de R {.smaller}

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.

. . .

<br/>Puntos 1. y 2. pueden ser resumidos como un enfoque de [programación orientada a objetos](https://es.wikipedia.org/wiki/Programaci%C3%B3n\_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

![](images/consola.png){fig-align="center" width=60%}

## Aritmética básica {.smaller}

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

```{r}

#| echo: true

1+2 ## Adición

6-7 ## Resta

5/2 ## División

2^3 ## Exponencial

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

```

## Sintáxis de R {.smaller}

Algunas reglas generales:

1. Cada línea es una sentencia ("comando"), varias sentencias se evalúan de arriba a abajo.

```{r, eval=FALSE, echo=TRUE}

c<-a+b

d<-sqrt(c)

```

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

```{r, eval=FALSE, echo=TRUE}

a\*(b

+c

+d)

```

## Sintáxis de R {.smaller}

2. Los espacios suelen ignorarse.

Todos son equivalentes:

```{r, eval=FALSE, echo=TRUE}

a+b

a + b

a + b

```

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

## Sintáxis de R {.smaller}

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:

```{r, eval=FALSE, echo=TRUE}

"hello world"

a\*(b+c)

x[1]

```

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

```{r, eval=FALSE, echo=TRUE}

y[c(1, 3)]

```

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

```{r, echo=TRUE, error=TRUE}

y[c(1, 3])

```

## Sintáxis de R {.smaller}

4. Coma y puntos

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

```{r, eval=FALSE, echo=TRUE}

log(x,5)

```

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

```{r, echo=TRUE, error=TRUE}

population <- 3,350,000

```

Se utiliza un punto como punto decimal:

```{r, eval=FALSE, echo=TRUE}

3.1415

```

## Paquetes en R

![](images/packages.png){fig-align="center" width=60%}

## Paquetes en R {.smaller}

- 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

```R

install.packages("tidyverse")

library(tidyverse)

```

::: {.fragment .fade-in}

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

```{r, echo=TRUE, error=TRUE}

wb\_search()

```

:::

## Ayuda en R {.smaller}

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

```R

help(plot)

```

O, simplemente, solo usen `?`:

```R

# Esta es la manera más común de usar la ayuda.

?plot

```

. . .

<br/>\*\*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)`.