

Introducción a la ciencia de datos con el lenguaje R

FLISoL 2018 - Tucumán

Fernando Flores

27/04/2018

Ciencia de datos

Ciencia de datos

Campo que reúne disciplinas de estadística y matemática con ciencias de la computación, basado en el método científico y aplicadas a diferentes dominios.

Objetivo principal: Descubrir nuevos conocimientos **accionables** a partir del análisis y modelado de datos.

"The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."

~ John Tukey

Definiendo éxito en ciencia de datos

- Se crea nuevo conocimiento
- Se crea un producto de datos (presentación, app, reporte) con impacto
- Decisiones son tomadas en base a los resultados del experimento
- Se aprende si los datos pueden o no responder a la pregunta de interés

Tipos de preguntas

1. **Descriptiva:** Caracterización de los datos
2. **Exploratoria:** Investigación generadora de hipótesis
3. **Inferencial:** Testeo de hipótesis, muestra vs población
4. **Predictiva:** En base a eventos observados, predecir comportamiento y patrones futuros
5. **Causal:** Ensayos aleatorizados, A/B testing
6. **Mecanicista:** Determinar la mecánica de por qué un resultado se altera según un factor/variable

Complejidad y costo aumentan a medida que avanzamos en la lista.

Ciclo de vida

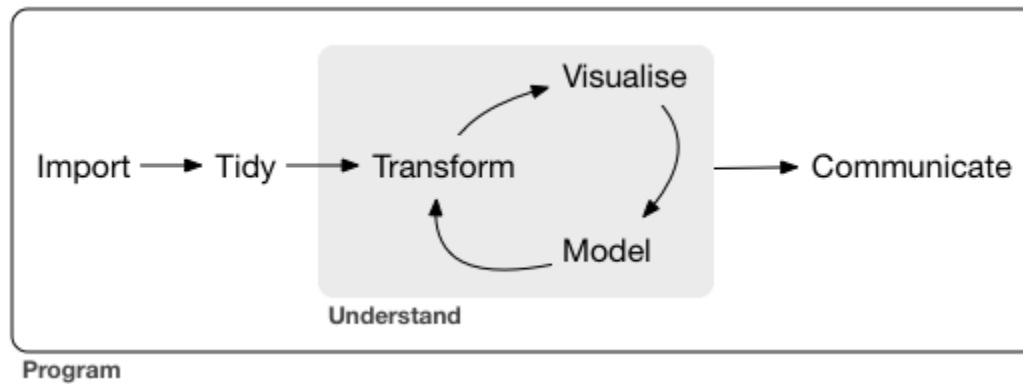


Imagen: Hadley Wickham y Garrett Grolemund, [R for Data Science](#).

Ciencia de datos con R

R



- Lenguaje de programación
- Software libre
- Creado por estadísticos para estadísticos
- Extensible



Importación (archivo)

```
library(tidyverse)
aeropuertos <- read_csv("./data/aeropuertos-AR.csv")
head(aeropuertos, 4)
```

```
## # A tibble: 4 x 20
##       id ident          type          name
##   <int> <chr>          <chr>          <chr>
## 1  5781  SAEZ  large_airport  Ministro Pistarini International Airport
## 2  5771  SABE medium_airport  Jorge Newbery Airpark
## 3  5806  SARI medium_airport  Cataratas Del Iguazú International Airport
## 4  5835  SAWH medium_airport  Malvinas Argentinas Airport
## # ... with 16 more variables: latitude_deg <dbl>, longitude_deg <dbl>,
## #   elevation_ft <int>, continent <chr>, iso_country <chr>,
## #   iso_region <chr>, municipality <chr>, scheduled_service <int>,
## #   gps_code <chr>, iata_code <chr>, local_code <chr>, home_link <chr>,
## #   wikipedia_link <chr>, keywords <chr>, score <int>, last_updated <dtm>
```

Fuente: Airports in Argentina, [Data World](#)

Importación (base de datos)

```
db_conn <-  
  DBI::dbConnect(  
    RMariaDB::MariaDB(),  
    host = "localhost",  
    dbname = "mydatabase",  
    user = "myusername",  
    password = rstudioapi::askForPassword("Database password"))  
  
aeropuertos <- tbl(db_conn, "aeropuertos-AR")
```



Importación (Apache Spark)

```
library(sparklyr)

spark_conn <- spark_connect(master = "local")

aeropuertos <- tbl(spark_conn, "aeropuertos-AR")
```

Tidy data

- Cada variable debe tener su propia columna
- Cada observación debe tener su propia fila
- Cada valor debe tener su propia celda

Tidy data

```
publicidad <- read_csv("./data/Advertising.csv")  
publicidad
```

```
## # A tibble: 200 x 5  
##       X1      TV radio newspaper sales  
##   <int> <dbl> <dbl>      <dbl> <dbl>  
## 1     1  230.1  37.8      69.2  22.1  
## 2     2   44.5  39.3      45.1  10.4  
## 3     3   17.2  45.9      69.3   9.3  
## 4     4  151.5  41.3      58.5  18.5  
## 5     5  180.8  10.8      58.4  12.9  
## 6     6    8.7  48.9      75.0   7.2  
## 7     7   57.5  32.8      23.5  11.8  
## 8     8  120.2  19.6      11.6  13.2  
## 9     9    8.6   2.1       1.0   4.8  
## 10    10  199.8   2.6      21.2  10.6  
## # ... with 190 more rows
```

Dataset Advertising.csv: G. James, D. Witten, T. Hastie and R. Tibshirani, [An Introduction to Statistical Learning, with applications in R](#) (Springer, 2013)



Tidy data

```
tidy_publicidad <-  
  publicidad %>%  
    select(market = X1, TV, radio, newspaper, sales) %>%  
    gather(key = "type", value = "budget", -market, -sales)  
head(tidy_publicidad, 3)
```

```
## # A tibble: 3 x 4  
##   market sales  type budget  
##   <int> <dbl> <chr>  <dbl>  
## 1      1  22.1   TV   230.1  
## 2      2  10.4   TV    44.5  
## 3      3   9.3   TV    17.2
```

```
tail(tidy_publicidad, 3)
```

```
## # A tibble: 3 x 4  
##   market sales      type budget  
##   <int> <dbl>    <chr>  <dbl>  
## 1    198  12.8 newspaper    6.4  
## 2    199  25.5 newspaper   66.2  
## 3    200  13.4 newspaper    8.7
```



Transformación

```
aero_data <-  
  aeropuertos %>%  
  filter(type != "closed") %>%  
  mutate(elevation_mts = elevation_ft * 0.3048) %>%  
  select(iata_code, type, elevation_mts, iso_region) %>%  
  add_count(iso_region)  
  
head(aero_data, 5)
```

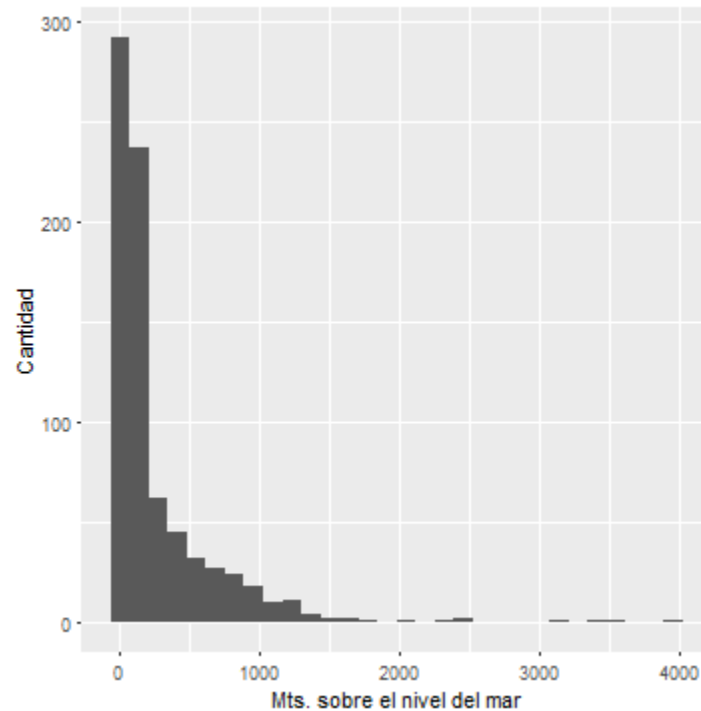
```
## # A tibble: 5 x 5
```

##	iata_code	type	elevation_mts	iso_region	n
##	<chr>	<chr>	<dbl>	<chr>	<int>
## 1	EZE	large_airport	20.4216	AR-B	239
## 2	AEP	medium_airport	5.4864	AR-C	12
## 3	IGR	medium_airport	279.1968	AR-N	13
## 4	USH	medium_airport	31.0896	AR-V	19
## 5	FTE	medium_airport	203.9112	AR-Z	34



Visualización (Histograma)

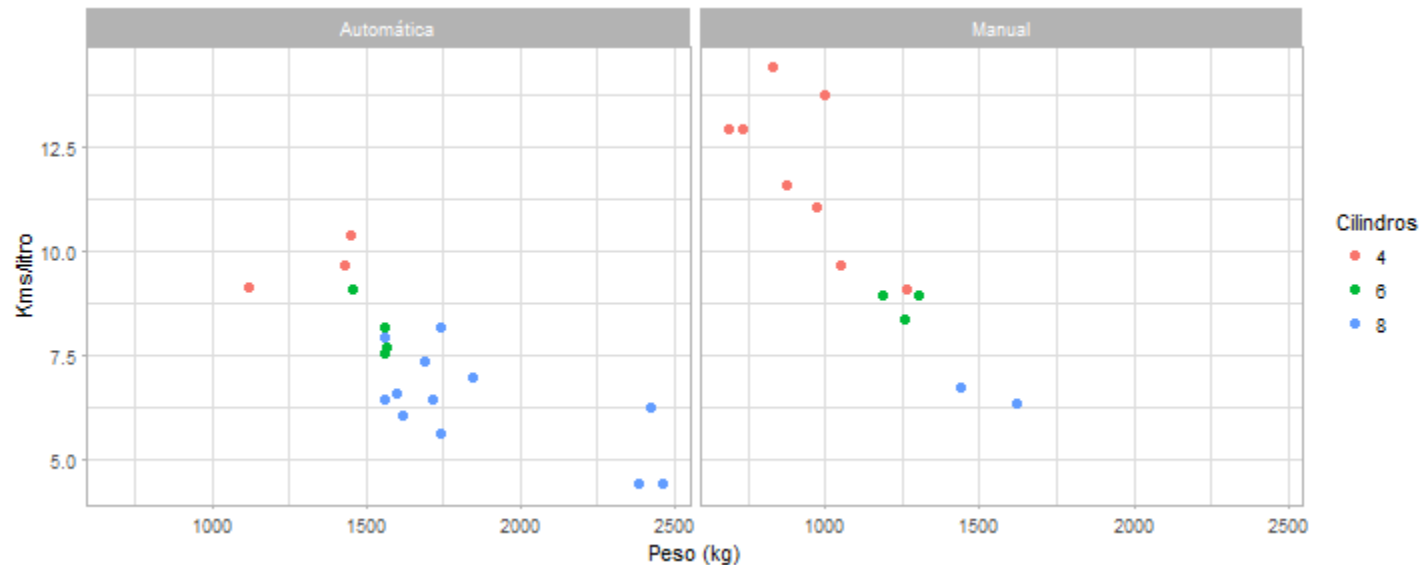
```
ggplot(data = aero_data, aes(x = elevation_mts)) +  
  geom_histogram() +  
  labs(x = "Mts. sobre el nivel del mar",  
       y = "Cantidad")
```





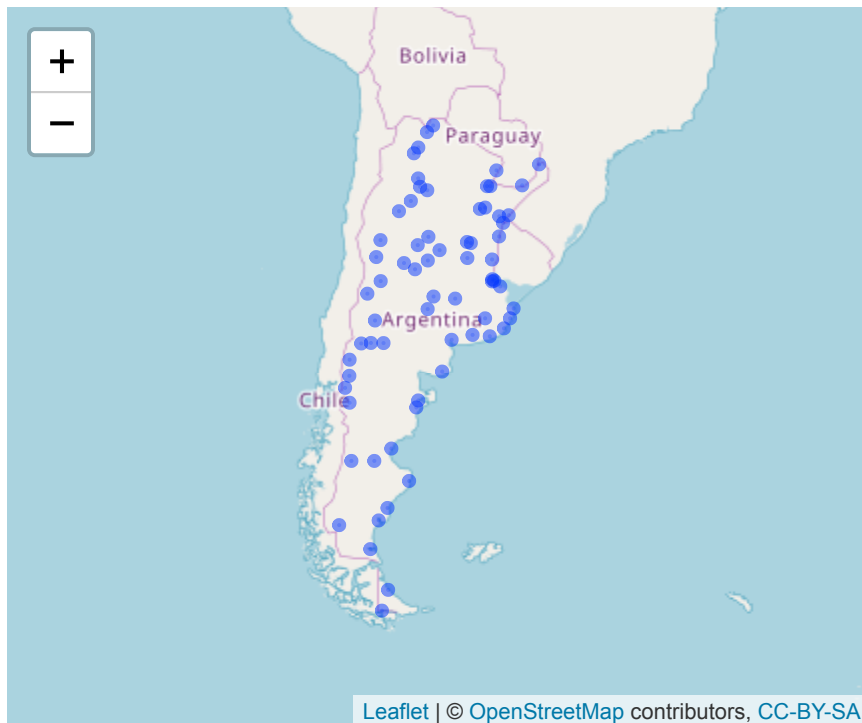
Visualización (Puntos)

```
mtcars %>%  
  mutate(km_lt = mpg * 0.425144,  
         kgs = wt * 1000 * 0.453592,  
         transm = if_else(am == 0, "Automática", "Manual")) %>%  
  ggplot(aes(x = kgs, y = km_lt, color = factor(cyl))) +  
  geom_point(size = 2) +  
  labs(x = "Peso (kg)", y = "Kms/litro", color = "Cilindros") +  
  facet_grid(~transm) +  
  theme_light()
```



Visualización (Mapas interactivos)

```
library(leaflet)
aeropuertos %>%
  filter(type == "medium_airport") %>%
  leaflet() %>%
  addTiles() %>%
  addCircles(lng = ~longitude_deg, lat = ~latitude_deg)
```



Modelos

Modelamos datos.

| "All models are wrong, some are useful."

~ George Box

Evaluamos modelos.

| "If you torture the data enough, nature will always confess."

~ Ronald Coase

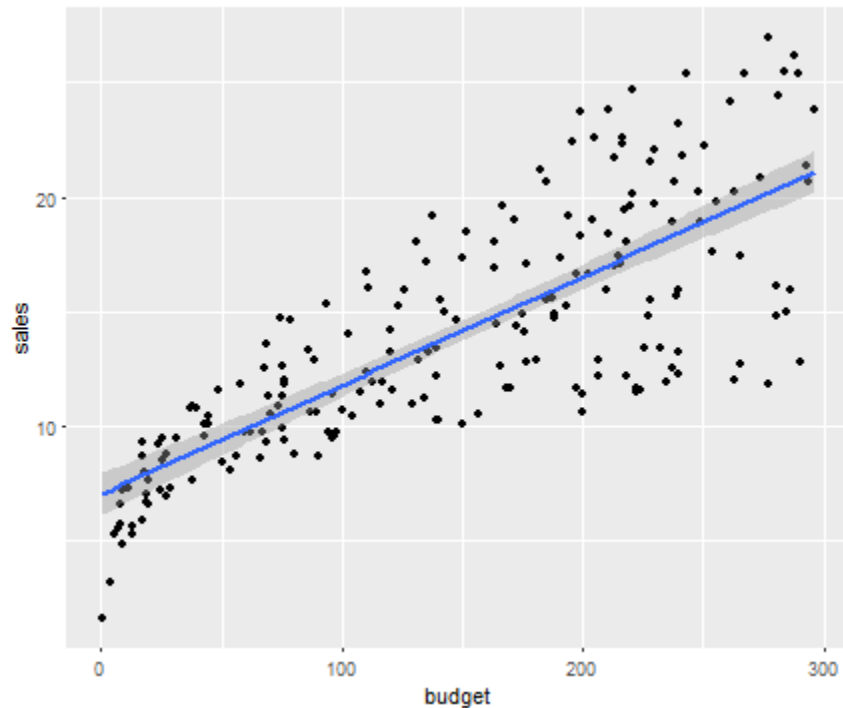
Modelos

```
tidy_publicidad_tv <- tidy_publicidad %>% filter(type == "TV")
modelo <- lm(sales ~ budget, data = tidy_publicidad_tv)
summary(modelo)
```

```
##
## Call:
## lm(formula = sales ~ budget, data = tidy_publicidad_tv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.032594    0.457843   15.36  <2e-16 ***
## budget       0.047537    0.002691   17.67  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Modelos

```
ggplot(tidy_publicidad_tv,  
       aes(x = budget, y = sales)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```



Comunicación

- Investigación reproducible
- Programación literaria (literate programming)

Comunicación

Reportes con R Markdown (Tufte handout):

Figures

Margin Figures

Images and graphics play an integral role in Tufte's work. To place figures or tables in the margin you can use the `fig.margin` knitr chunk option. For example:

```
library(ggplot2)
qplot(Sepal.Length, Petal.Length, data = iris,
      color = Species)
```

Note the use of the `fig.cap` chunk option to provide a figure caption. You can adjust the proportions of figures using the `fig.width` and `fig.height` chunk options. These are specified in inches, and will be automatically scaled down to fit within the handout margin.

Equations

You can also include \LaTeX equations in the margin by explicitly invoking the `marginfigure` environment.

Note the use of the `\caption` command to add additional text below the equation.

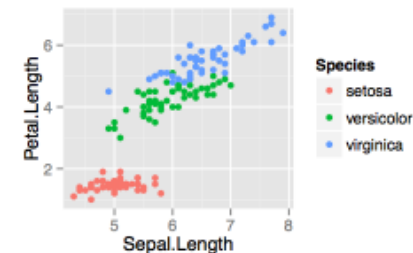
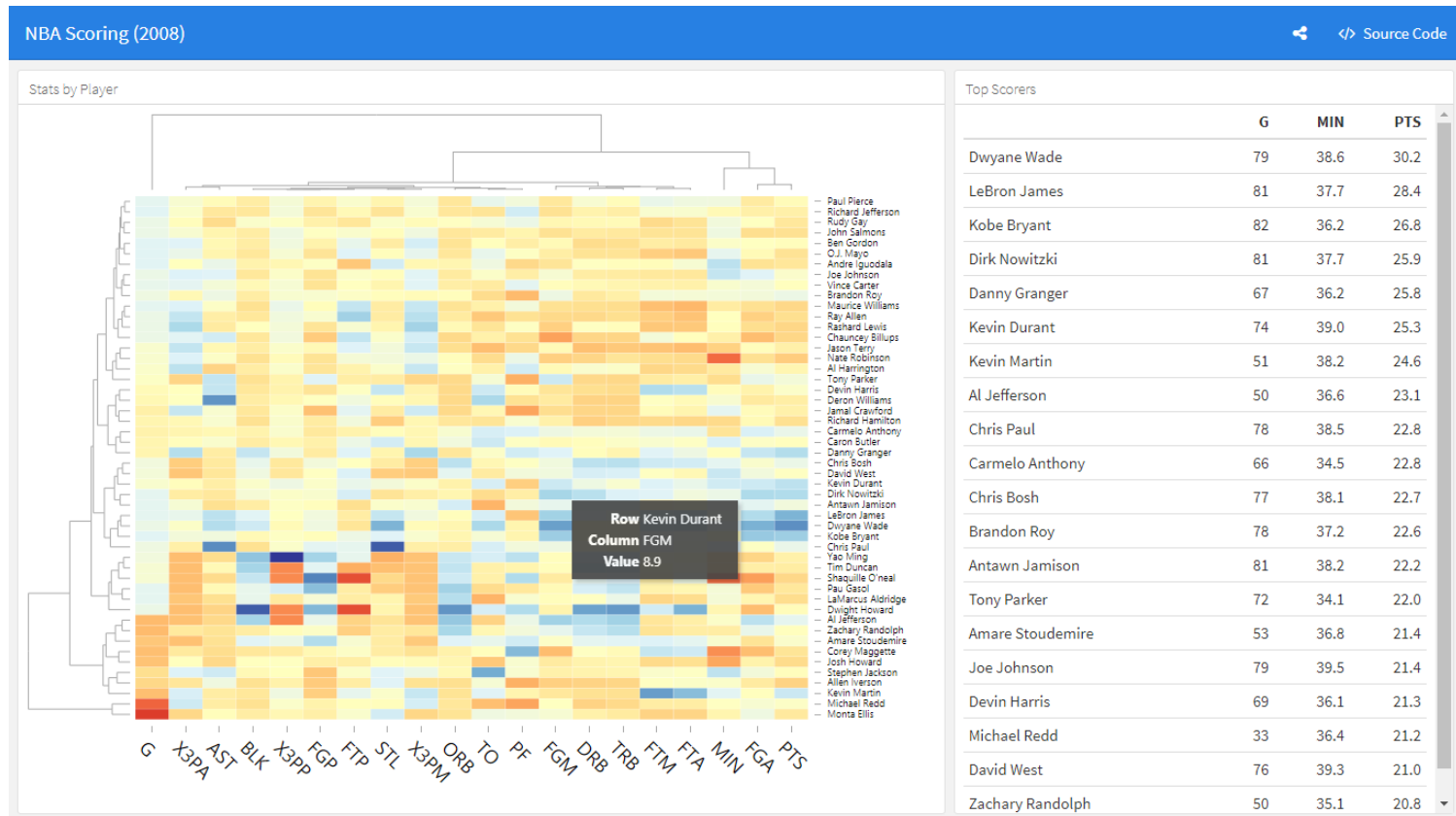


Figure 1: Sepal length vs. petal length, colored by species

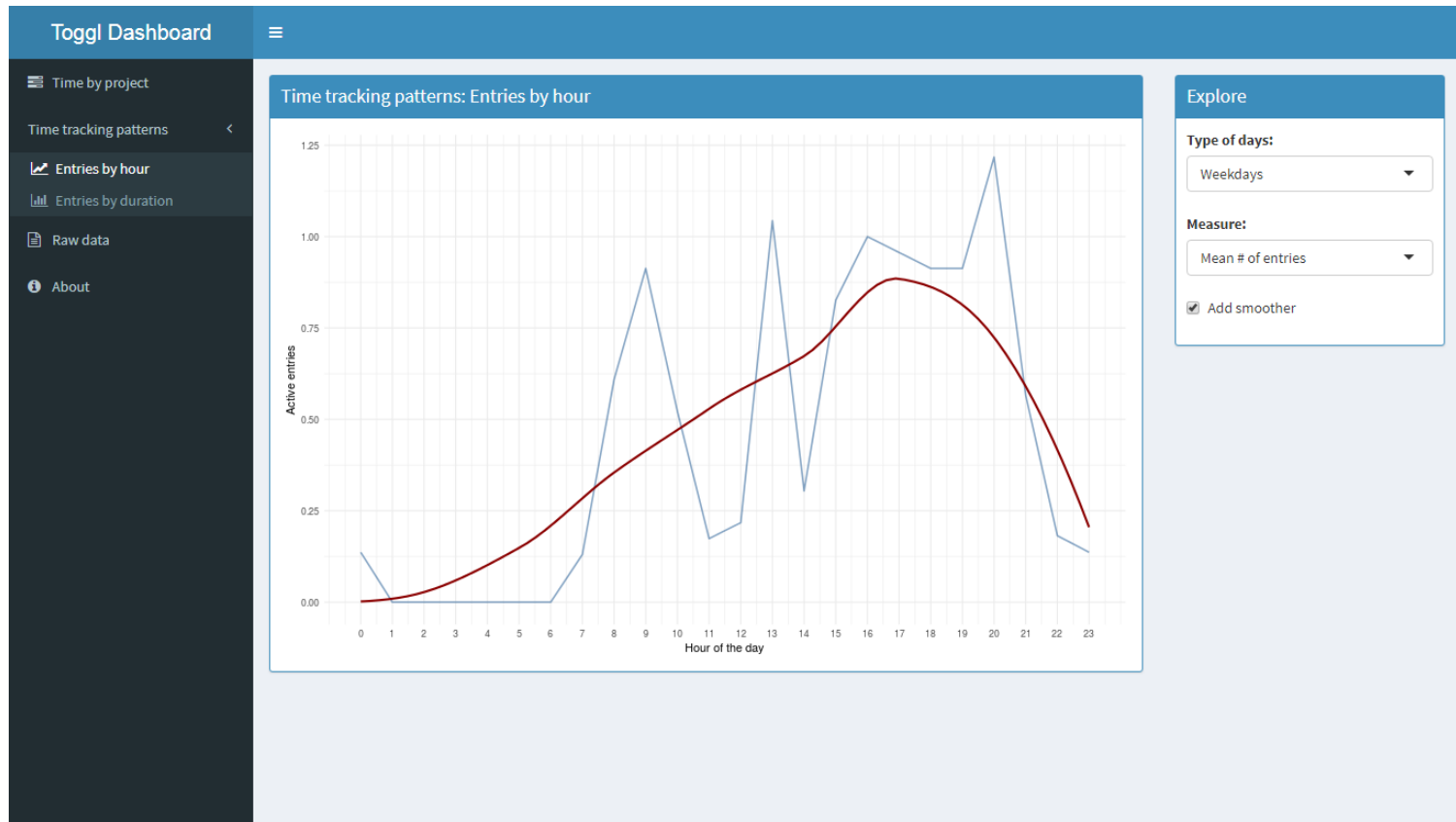
$$\frac{d}{dx} \left(\int_0^x f(u) du \right) = f(x).$$

Figure 2: An equation

Dashboard con Flexdashboard (NBA Scoring):




Aplicaciones web con Shiny (**Time tracking Dashboard**):









Sitios y blogs con blogdown (**Website personal**):

FERNANDO FLORES

[Home](#) [Projects](#) [Talks](#) [Blog](#) [Reading](#) [Contact](#)



Fernando Flores
Data science, software development
and engineering



I'm a Data scientist at the [Coordination for Digital Education \(AR\)](#) and the Technical director of [Tucma Software](#). I also work as a consultant in data science and software development.

During my career I've worked in the public and private sectors, within on-site and globally distributed teams in industries like software, IT, retail, education and fintech, providing solutions for clients and users in America, Europe and Africa.

I helped companies across all stages of the data science life cycle, from kickstarting their data initiatives to statistical analytics, automation and application/dashboard development.

Do you need collaboration to achieve a measurable, data-driven improvement in your business? Feel free to [book a consultation](#) or connect with me on [LinkedIn](#).

Interests

- Data Science
- Machine Learning
- Artificial Intelligence
- Math
- Big Data
- Data Ethics
- AI Safety

Libros en línea con bookdown (Text Mining with R):

Text Mining with R

Welcome to Text Mining with R

Preface

1 The tidy text format

2 Sentiment analysis with tidy data

2.1 The sentiments dataset

2.2 Sentiment analysis with inner ...

2.3 Comparing the three sentime...

2.4 Most common positive and ne...

2.5 Wordclouds

2.6 Looking at units beyond just w...

2.7 Summary

3 Analyzing word and document freq...

4 Relationships between words: n-gr...

5 Converting to and from non-tidy for...

6 Topic modeling

7 Case study: comparing Twitter arc...

8 Case study: mining NASA metadata

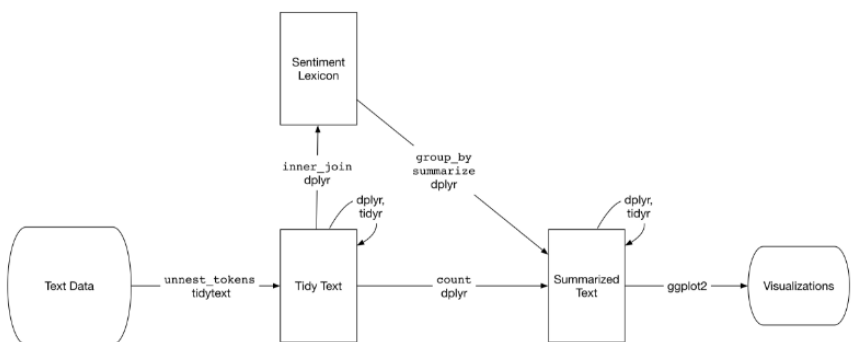
9 Case study: analyzing usenet text

10 References

Published with bookdown

2 Sentiment analysis with tidy data

In the previous chapter, we explored in depth what we mean by the tidy text format and showed how this format can be used to approach questions about word frequency. This allowed us to analyze which words are used most frequently in documents and to compare documents, but now let's investigate a different topic. Let's address the topic of opinion mining or sentiment analysis. When human readers approach a text, we use our understanding of the emotional intent of words to infer whether a section of text is positive or negative, or perhaps characterized by some other more nuanced emotion like surprise or disgust. We can use the tools of text mining to approach the emotional content of text programmatically, as shown in Figure 2.1.



```
graph LR;
    A[Text Data] -- "unnest_tokens<br/>tidytext" --> B[Tidy Text];
    B -- "inner_join<br/>dplyr" --> C[Sentiment Lexicon];
    B -- "dplyr,<br/>tidyr" --> D[Summarized Text];
    C -- "group_by<br/>summarize<br/>dplyr" --> D;
    B -- "count<br/>dplyr" --> D;
    D -- "dplyr,<br/>tidyr" --> E[Visualizations];
    D -- "ggplot2" --> E;
```

Figure 2.1: A flowchart of a typical text analysis that uses tidytext for sentiment analysis. This chapter shows how to implement sentiment analysis using tidy data principles.



Presentaciones con xaringan:

Introducción a la ciencia de datos con el lenguaje R

FLISoL 2018 - Tucumán

Fernando Flores

27/04/2018

1 / 34

Vine por el lenguaje,
me quedé por la comunidad

Ecosistema y comunidad

- Contribuir a proyectos y paquetes. Por ejemplo: **Contribute to the tidyverse**.
- **rOpenSci**: Comunidad de usuarios y desarrolladores que incentivan la investigación científica reproducible con herramientas libres.
- **R-Ladies**: Organización para promover diversidad de género en la comunidad de R.
- **R Forwards**: Grupo de trabajo para incrementar la participación de grupos con poca representación en este campo.

Ecosistema y comunidad

- **R-Bloggers** y **R Weekly**: Posts, tutoriales y noticias contribuidos por la comunidad.
- **RStudio Community**: Comunidad inclusiva y cordial de preguntas y respuestas.
- Hashtags **#rstats** y **#rstatsES**.
- Conferencias, por ejemplo **Latin-R** (Bs.As., 4 y 5 de septiembre).

P&R

¡Gracias!

@ds_floresf

floresfdev.github.io