Limpieza y Transformación de Datos con R

Josué Romero J.*
* Corporación Universitaria Minuto de Dios
Bogatá D.C, COA Engativa, CLL 80
† josue.romero@uniminuto.edu.co'
Minería de Datos | NRC. 70348 | Narly Sanchez
‡ 08 de octubre de 2024

***Abstract*—Demostramos el uso de una plantilla RMarkdown para el estilo IEEEtran. Por ahora es solo estilo de conferencia, pero eventualmente ampliaremos (quizás).**

I. Introducción

Este archivo de demostración está destinado a servir como un "archivo de inicio" para artículos de conferencia IEEE producidos bajo LaTeX usando IEEEtran.cls versión 1.8b y posteriores. Te deseo el mayor de los éxitos.

mds

08 de octubre de 2024

Usa `rmarkdown::render()` para crear este documento; esencialmente llama a `knit()` para ir de RMD a MD, y luego `pandoc` (con todas las configuraciones en el YAML) para ir de MD a PDF.

Se podría intentar compilar a HTML, pero por supuesto ninguno de los estilos IEEE se aplicará. Y si se ha incluido algún LaTeX crudo en el documento (como es típico en un artículo, ya que podrías necesitar el poder adicional de LaTeX para proporcionar un diseño específico), esto no se compilará en HTML.

II. Ejemplos

A. Knitr

Puedes usar knitr como de costumbre. La opción de bloque `echo=F` debería establecerse (a menos que desees mostrar el código R en el artículo). Además, dado que este es un diseño de dos columnas, probablemente se desbordará, así que necesitarás

- ajustar el código tú mismo (por defecto knitr no ordena el código), o
- habilitar el ajuste de código y especificar el ancho: `opts_knit$set(tidy=T, tidy.opts=list(width.cutoff=40))`.
- NB: la opción de bloque `size` (por ejemplo `opts_chunk$set(size="small")`) solo funciona en Rnw, no en Rmd).

El ancho es bastante pequeño. Para este documento, puedes ajustar aproximadamente 42 caracteres antes de que se desborde (ver el ejemplo en sección II-B).

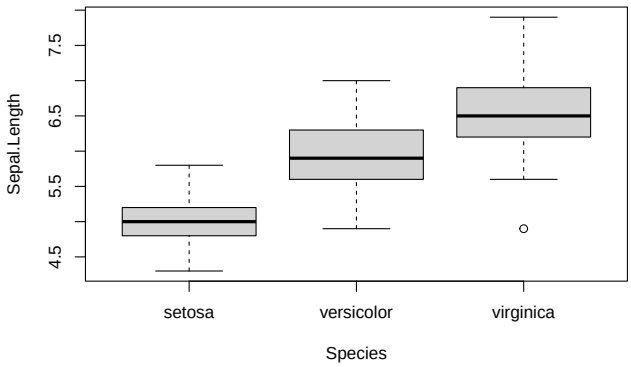


Fig. 1: Longitudes de sépalos para varias especies de iris.

B. Figuras

Por supuesto, puedes generar gráficos usando R y se insertarán con knitr. Sin embargo, dado que knitr va de MD a RMD, se insertarán en formato markdown, no en formato TeX. He configurado knitr para colocar figuras en el directorio `figure/` (`opts_chunk$set(fig.path='figure/')`), así que ahí es donde estará el gráfico.

```
plot(Sepal.Length ~ Species, iris)
```

Ver figura 1. (No estoy seguro por qué esto es "Fig. 1" en la leyenda… ¿es un asunto de knitr/rmarkdown/pandoc, o un asunto de IEEEtran?)

En la práctica, probablemente querrás escribir tu código de figura en LaTeX crudo para tener un mayor control. En el bloque de configuración de este Rmd hay una función `latex.figure` que es un ejemplo de salida de LaTeX crudo para una figura. Ajusta como desees. (Seguramente hay una biblioteca como `xtable` para esto).

```
latex.figure(
  'figure/iris.plot-1.pdf',
  caption='Otro gráfico de longitudes de sépalos
           para las diversas especies de iris.',
```

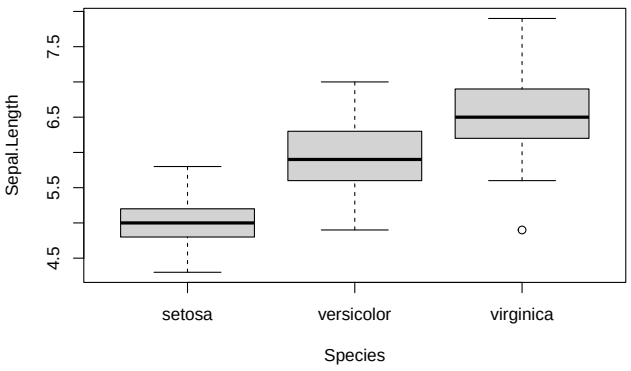


Fig. 2: Otro gráfico de longitudes de sépalos para las diversas especies de iris.

```
  label='fig:iris2')
```

El comando `latex.figure` también tiene soporte básico para subfiguras: simplemente proporciona múltiples rutas de las imágenes. Si hay tantas leyendas como figuras, se utiliza una para cada una.
Si hay una leyenda más que la cantidad de figuras, la primera se usa como la leyenda "principal" y el resto como leyendas de subfiguras. Si solo hay una leyenda, se utiliza para la figura y no se añaden subleyendas.
Consulta figura **??** para ver el resultado.

```
# generar y guardar algunas imágenes
n = 1:5
figs = sprintf('figure/x%i.png', n)
for (nn in n) {
  png(filename=figs[nn], width=480, height=3
  plot(1:10, (1:10)^nn)
  dev.off()
}

# mostrar como figura flotante con 3 subfiguras
latex.figure(
  figs,
  caption=c("Polinomios",
            sprintf("$x^%i$", n)),
  label='fig:polynomials',
  linebreaks.after=3,
  width='.6\\columnwidth',
  floating=T)
```

Nota que frecuentemente los artículos de IEEE con subfiguras no usan leyendas para las subfiguras, sino que en su lugar las referencian/describen como (a), (b), etc., dentro de la leyenda principal.

TABLE I: Ejemplo del conjunto de datos iris

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
7.70	2.60	6.90	2.30	virginica
5.90	3.00	5.10	1.80	virginica
5.50	3.50	1.30	0.20	setosa
5.00	3.50	1.60	0.60	setosa
6.40	2.80	5.60	2.20	virginica
5.10	3.70	1.50	0.40	setosa

También nota que típicamente IEEE coloca los elementos flotantes solo en la parte superior, incluso cuando esto resulta en que un gran porcentaje de una columna esté ocupado por figuras flotantes.

C. *Tables*

No debes usar la sintaxis de pandoc, ya que utiliza el paquete `longtable` (esto está codificado) y `longtable` no funciona bien con entradas de dos columnas. Usa algo como Hmisc o xtable para generar salida en LaTeX y proporcionar mayor control (por ejemplo, tabla I).

```
print(xtable(
  iris[sample(nrow(iris), 6), ],
  caption='Ejemplo del conjunto de datos iris',
  label='tbl:iris.xtable',
  align=c(rep('r', 5), 'l')))
```

Podrías desear que la tabla abarque varias columnas. Usa `table*` en lugar de `table` (tabla II).
Nota que el argumento `floating.environment` pertenece a `print.xtable`, no a `xtable`.

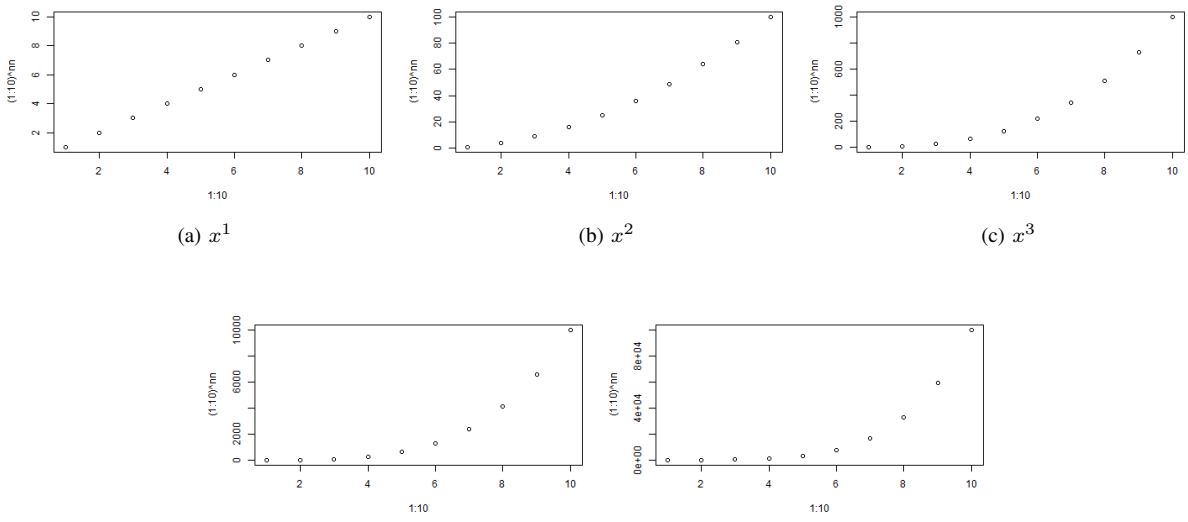
```
print(xtable(
    head(mtcars),
    caption='Ejemplo del conjunto de datos
             de pruebas de automóviles de motor t
    label='tbl:xtable.floating'),
  floating.environment='table*')
```

Note que, para las tablas en el estilo IEEE, dado que los títulos de las tablas funcionan como encabezados, las leyendas suelen escribirse con mayúscula inicial en todas las palabras, excepto aquellas como: a, an, and, as, at, but, by, for, in, nor, of, on, or, the, to y up, que generalmente no se capitalizan, a menos que sean la primera o última palabra de la leyenda. El texto de las tablas usará por defecto `\footnotesize`, ya que el IEEE normalmente emplea esta fuente más pequeña para las tablas.

Note que IEEE típicamente coloca los flotantes solo en la parte superior, incluso cuando esto resulta en que un gran porcentaje de una columna esté ocupada por flotantes.

D. *Citando*

Ejemplos de citar un autor [1] y dos autores [1, 2].



(a) x^1 (b) x^2 (c) x^3

(d) x^4 (e) x^5

Fig. 3: Polinomios

TABLE II: Ejemplo del conjunto de datos de pruebas de automóviles de motor trend

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00
18.10	6.00	225.00	105.00	2.76	3.46	20.22	1.00	0.00	3.00	1.00

E. Ecuaciones

Son como cabría esperar. Puede utilizar la sintaxis de pandoc-crossref para generar etiquetas. Es decir,

```
$$
e = m c^2
$$ {#eq:einstein}
```

Que es igual a

$$e = mc^2. \tag{1}$$

Se puede usar `@eq:einstein` para referirse a la ecuación, por ejemplo, 1. El único inconveniente es que la ecuación debe estar en su propio párrafo si desea numerarla, lo que significa que en el archivo tex y pdf resultante, la ecuación estará en su propia línea. (Si no desea numerar la ecuación, no tiene que estar en su propio párrafo y se renderizará en el párrafo como cabría esperar).

Aún no he encontrado una buena solución para esto. Es un requisito de `pandoc-crossref`. Debe editar el archivo TeX y eliminar las líneas en blanco adicionales (donde sea apropiado) antes de compilar. Agrego un comentario `% FIXME ALIGNMENT` a estas ecuaciones para hacerlas más fáciles de encontrar.

III. Conclusión

Espero que se le haya dado un breve recorrido por las capacidades de esta configuración y que ahora proceda a escribir artículos al estilo IEEEtran utilizando RMarkdown con (relativa) facilidad.

Agradecimientos

Esta plantilla no sería posible sin los archivos IEEEtran de Michael Shell, pandoc, pandoc-crossref, knitr, rmarkdown, y una exhaustiva búsqueda en StackOverflow. Merece reconocimiento también Rstudio. No es necesario para esto, pero ciertamente facilita todo el proceso. Y a cualquiera que haya olvidado mencionar.

References

[1] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society B*, pp. 192–236, 1974.

[2] ——, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society B*, pp. 259–302, 1986.