Análisis de datos con R

Rubén Fernández Casal

UNIVERSIDADE DA CORUÑA

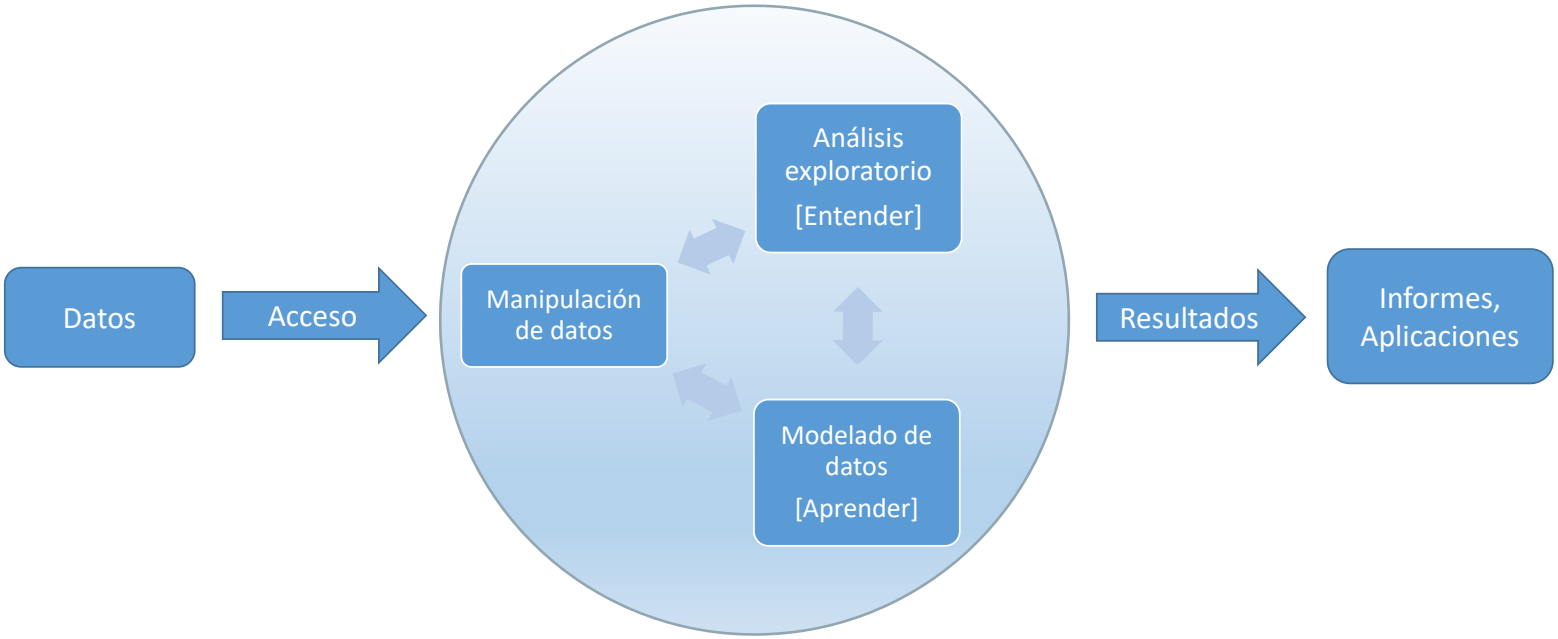
Big Data en PyMEs

Diciembre 2016

Indice

Indice

- **El objetivo es la obtención de información**
  (para la toma de decisiones).
- El proceso:



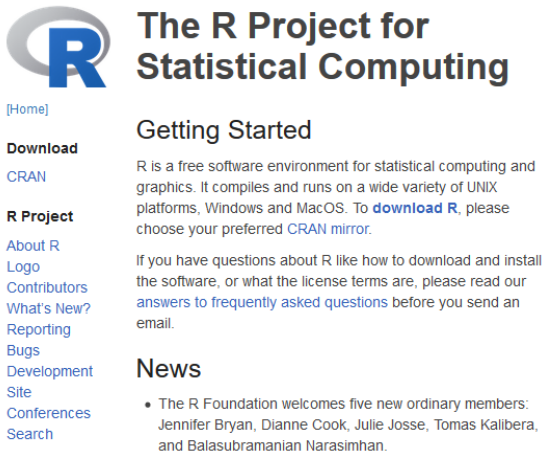
- El entorno estadístico `R` puede ser una herramienta de gran utilidad en todo este proceso.

El entorno estadístico R

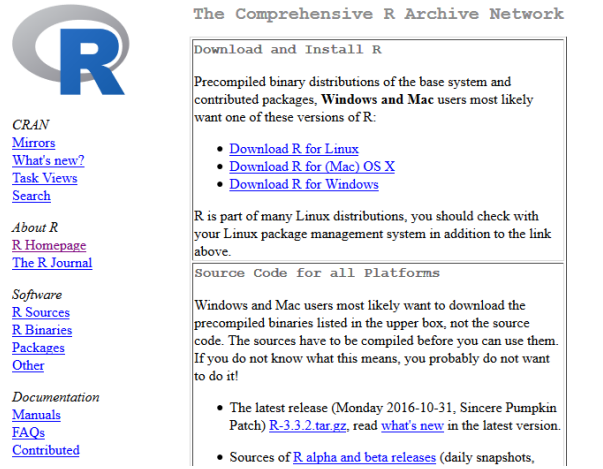
R es un lenguaje de programación desarrollado específicamente para el análisis estadístico y la visualización de datos.

- Lenguaje interpretado (similar a Matlab o Phyton)
  - derivado del S (Laboratorios Bell).
- Libre de código abierto (licencia GPL).
- Multiplataforma (Linux, Windows, MacOS, ...).

http://www.r-project.org

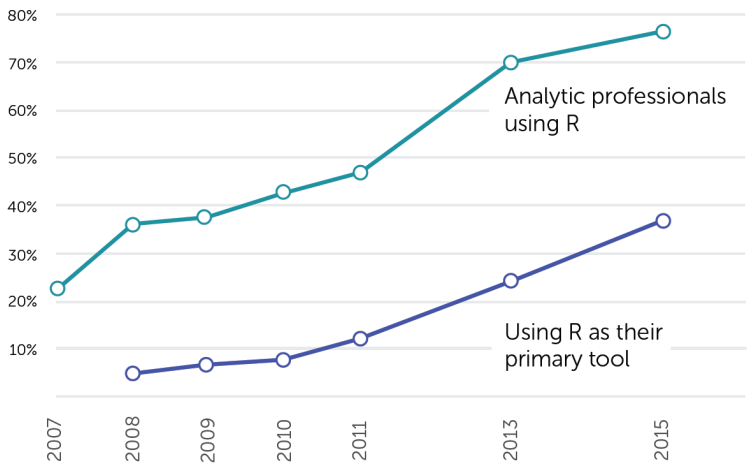


http://cran.es.r-project.org



- Dispone de múltiples librerías (paquetes) que cubren "literalmente" todos los campos del análisis de datos.
  - Repositorios: CRAN (9705), Bioconductor (1289), ...
- Comunidad R muy dinámica (con muchas contribuciones).
  - R es muy popular...

Encuesta sobre el uso de R



Rexer Data Miner Survey 2007-2015

Popularidad lenguajes

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

IEEE Spectrum Top Programming Languages, 2016

- Puntos débiles (a priori): velocidad, memoria, ...

Ventajas de R respecto a otras alternativas

R destaca especialmente en:

- Representaciones gráficas.
- Métodos estadísticos "avanzados"

  (*Data Science*: *Data Mining*, *Machine Learning*, *Statistical Learning*, *Business Intelligence*, ...):

  - Datos funcionales.
  - Estadística espacial.
  - ...
- Análisis de datos "complejos":
  - Big Data.
  - Lenguaje natural (*Text Mining*).
  - Análisis de redes.
  - ...

James *et al.* (2008). *An Introduction to Statistical Learning: with Aplications in R*. Springer (disponible en: *http://www-bcf.usc.edu/~gareth/ISL*).

Williams (2011). *Data Mining with Rattle and R*. Springer.

Compañías que usan R

Cada vez son más las empresas que utilizan R.

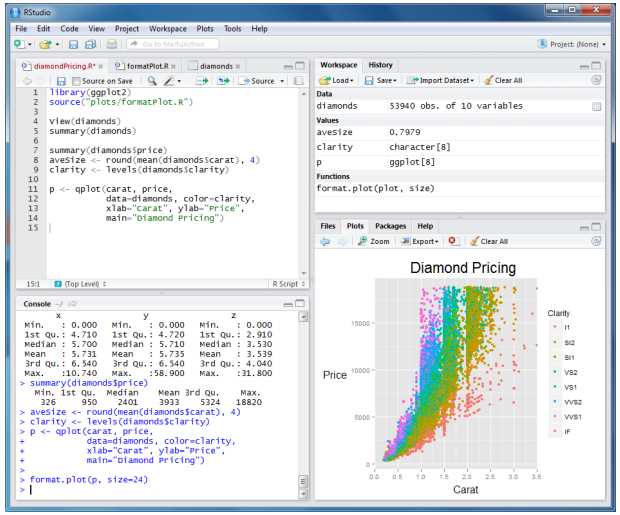
- R Consortium

  Grupo de empresas que apoyan a la Fundación R y a la comunidad R.

- Otras compañías:
  - Facebook, Twitter, Bank of America, Monsanto, ...
  - http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html
  - http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html

- Microsoft R Server
  Diseñado para entornos Big Data y computación de altas prestaciones.
- Microsoft R Open
  Versión de R con rendimiento mejorado.
  - Microsoft R Application Network:
    http://mran.revolutionanalytics.com
- Integracion de R con:
  SQL Server, PowerBI, Azure y Cortana Analytics.

Microsoft no reveló el precio de compra;
la capitalización total de Revolution era de unos 40 millones de dólares en esa fecha.
http://thomaswdinsmore.com/2015/01/26/microsoft-buys-revolution-analytics



- RStudio Desktop

  Entorno de desarrollo (IDE) con múltiples herramientas.

- RStudio Server

  Interfaz web que permite ejecutar RStudio en el servidor.

  - Evita el movimiento de datos a los clientes.
  - Ediciones Open Source y Professional.

- Compañía muy activa en el desarrollo de R:

  - Múltiples paquetes: Shiny, rmarkdown, knitr, ggplot2, dplyr, tidyr, ...
  - Hadley Wickham (Jefe científico de RStudio).

Van der Loo y de Jonge (2012). *Learning RStudio for R Statistical Computing*. Packt Publishing.

Indice

Acceso a datos

Hay una gran cantidad de paquetes de R para el acceso a distintos formatos/infraestructuras de datos:

- Archivos de datos
  - csv, xlsx, xml, json, sas, spss, stata, ...
- Bases de datos relacionales
  - `DBI`: Controladores de bases de datos nativos (máximo rendimiento).
    - `RMySQL`, `ROracle`, `RSQLite`, `RPostgreSQL`, ...
  - `RODBC`, `RJDBC`, ...
- Big Data
  - Hadoop (`RHadoop`).
  - Hive (`RHive`), Spark (`SparkR`), ...
- Web scraping
  - `rvest`, `RSelenium`, `twitteR`, `Rfacebook` , ...

Munzert *et al.* (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Wiley.

Ejemplo: Estadística Espacial

Por ejemplo, con R (`osmar`) se pueden importar mapas de OpenStreetMap y manipularlos fácilmente (`sp, igraph, ...`).

- Gran cantidad de paquetes para la manipulación y el análisis estadístico de datos espaciales (CRAN Task View: Analysis of Spatial Data):
  - El paquete `sp`
    Bivand *et al.* (2008). *Applied Spatial Data Analysis with R*. Springer.
  - Integración con sistemas GIS, Maps y API de Google, ...

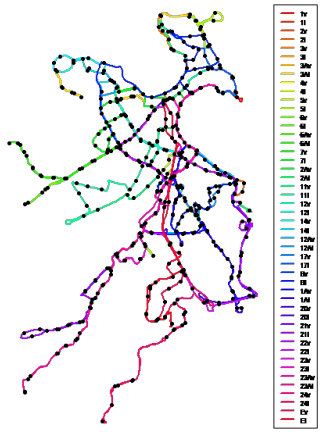
Manipulación de datos

- El paquete básico de R proporciona múltiples herramientas:
  - `data.frame`: objeto de datos base (tabla de datos).
- Para manipular conjuntos de datos (grandes) de forma más eficiente están disponibles otros paquetes:
  - `data.table`, `tibble`, `raster`, ...
- El paquete `dplyr` proporciona un entorno homogéneo para el acceso y la manipulación de datos:
  - Utiliza una sintaxis de la forma:
    `datos %>% filter %>% group_by %>% summarise...`
  - Puede trabajar con datos en distintos formatos:
    - `data.frame`, `data.table`, `tibble`, ...
    - Bases de datos relacionales
    - Hadoop (`plyrmr`), Spark (`sparklyr`), ...

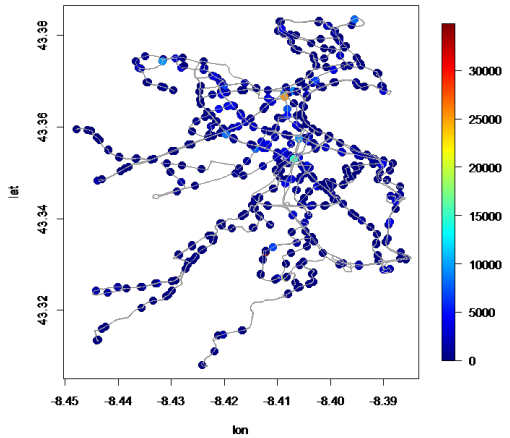
  Evita emplear comandos de otros lenguajes (SQL, HQL, Scale, ...).

Ejemplo: datos en red

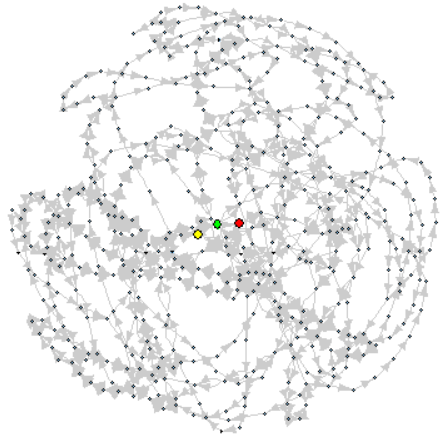
Red de autobuses de A Coruña



Los datos son atributos de nodos (paradas) y arcos (tramos)



Número de billetes



Nodos centrales

Kolaczyk y Csárdi (2014). *Statistical analysis of network data with R*. Springer.

Indice

Análisis exploratorio

Métodos exploratorios (o de aprendizaje no supervisado):

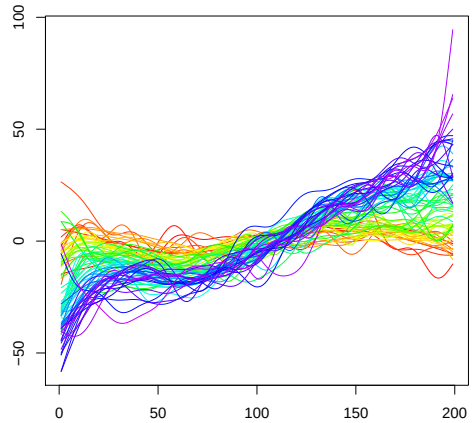
El objetivo es obtener información directamente de los datos, buscar relaciones y patrones, ...

(Tukey, 1975)

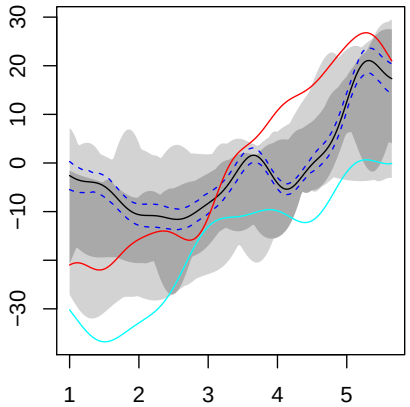
- Análisis descriptivo.
  - Gráficos:
    - Estáticos: estándar, `lattice`, `ggplot2`, ...
    - Dinámicos: `rggobi`, `ggvis`, ...
- Métodos de reducción de la dimensión:
  - Análisis de componentes principales, análisis factorial, ...
  - Análisis Clúster.
- Detección de datos atípicos.
- ...

Ejemplo: datos funcionales

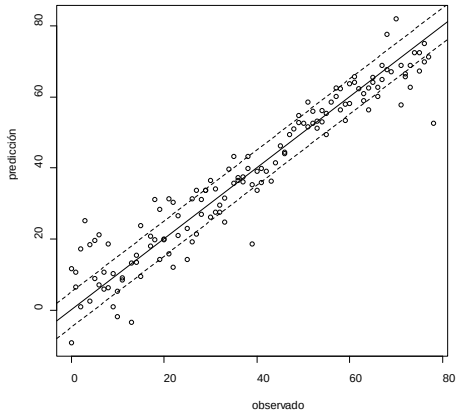
Mediciones durante la producción de silicio (parámetros eléctricos del horno).



Cada observación es una curva (correspondiente a una colada)



Boxplot funcional



Predicciones duración crisol (modelo funcional)

Ramsay *et al.* (2009). *Functional Data Analysis with R and MATLAB*. Springer.

Indice

Modelado de datos

- La realidad puede ser muy compleja por lo que es habitual emplear un modelo para tratar de explicarla.
  - Modelos estocásticos (con componente aleatoria).
    - Tienen en cuenta la incertidumbre debida a no disponer de información suficiente.
    - La inferencia estadística proporciona herramientas para ajustar y contrastar la validez del modelo.
  - El objetivo es disponer de una aproximación simple de la realidad que sea útil (George Box: "En esencia, todos los modelos son falsos, pero algunos son útiles").
    - En ocasiones el objetivo es únicamente predecir.
- Métodos (de aprendizaje supervisado):
  - Clasificación
    - Análisis discriminante, Regresión logística, ...
    - Árboles de decisión, *bagging*, *random forest*, *boosting*
    - *Support vector machines* (SVM)
  - **Regresión**

Métodos de regresión

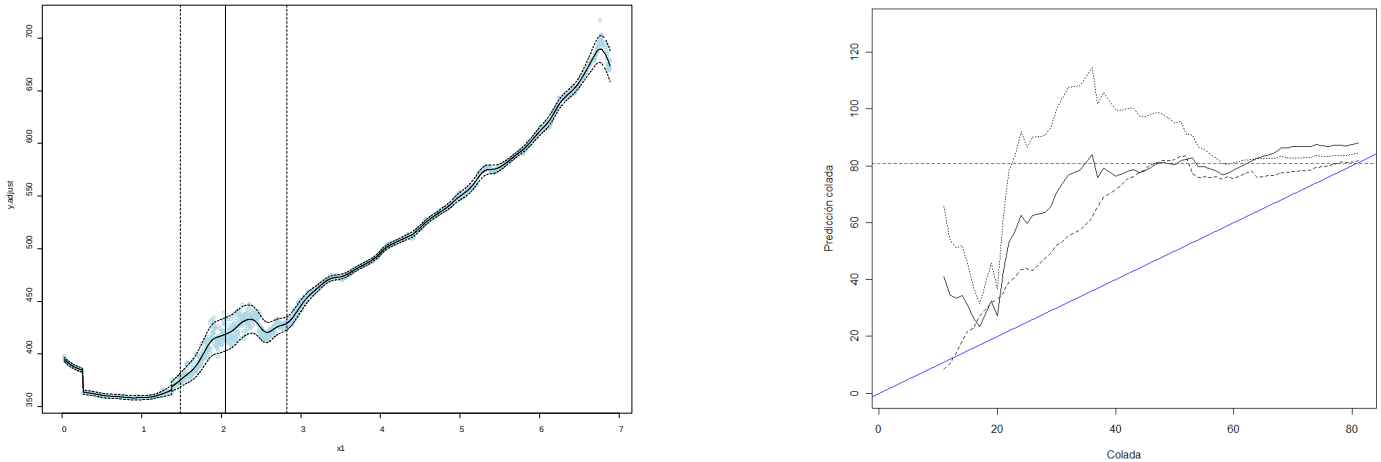
Algunas de las funciones y paquetes disponibles:

- Modelos paramétricos
  - Modelos lineales:
    - Regresión lineal: `lm()` (`aov()`, `lme()`, `biglm`, ...).
    - Regresión lineal robusta: `rlm()` (`MASS`, ...).
    - Métodos de regularización (Ridge regression, Lasso): `glmnet`, ...
  - Modelos lineales generalizados: `glm()` (`bigglm`, ...).
  - Modelos paramétricos no lineales: `nls()` (`nlme`, ...).
- Modelos no paramétricos
  - Regresión local (métodos de suavizado): `loess()`, `KernSmooth`, `sm`, ...
  - **Modelos aditivos generalizados (GAM)**: `gam`, **mgcv**, ...
  - Arboles de decisión (Random Forest, Boosting): `rpart`, `randomForest`, `xgboost`, ...
  - Redes neuronales (`nnet`, ...), ...

Con todos los modelos se trabaja de una forma muy similar en `R`.

Ejemplo: Modelos aditivos

Mediciones durante la producción de silicio (parámetros eléctricos del horno).



Estimación de la media (y de la variabilidad) en una colada
y predicciones del número total de coladas (derivadas del modelado del proceso).

- Modelo:

$$Y = \beta_0 + f_1(\mathbf{X}_1) + f_2(\mathbf{X}_2) + \cdots + f_p(\mathbf{X}_p) + \varepsilon,$$

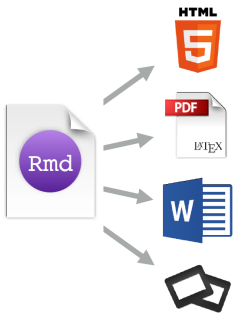
con f_i, $i = 1, ..., p$, funciones cualesquiera (suaves).

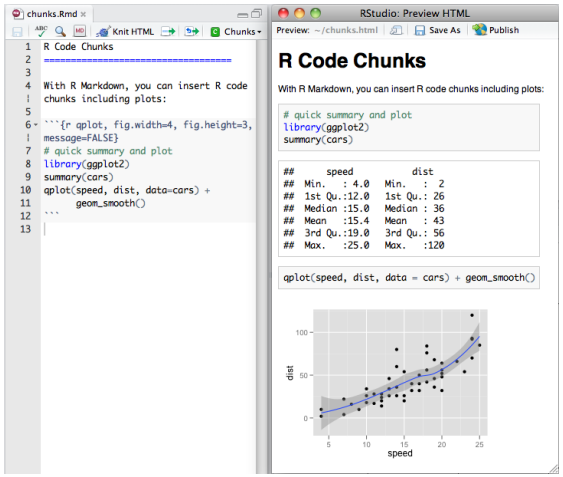
  - Adicionalmente se puede considerar una función link.

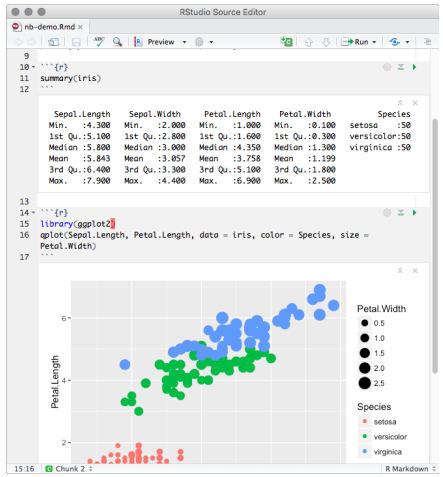
    Wood (2006). *Generalized Additive Models: An Introduction with R*. Chapman.

Indice

Generación de informes



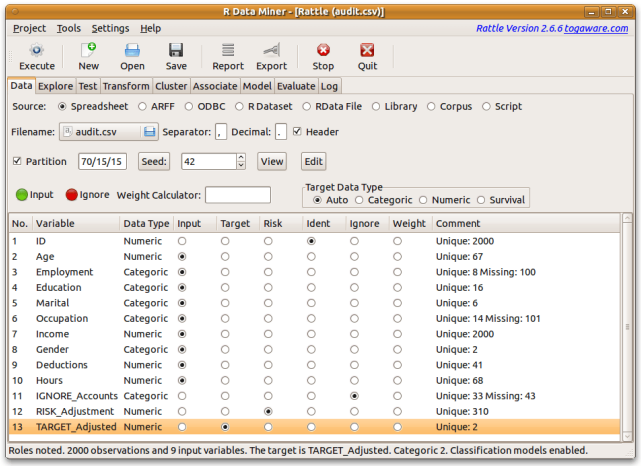




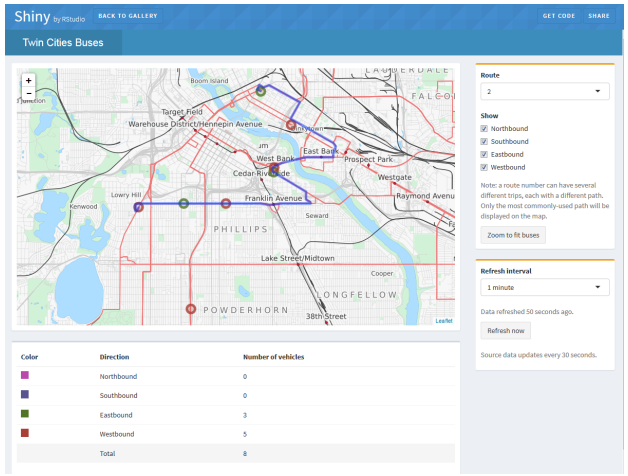
- Informes
  - R-Markdown permite la creación de informes en distintos formatos (HTML, PDF, DOCX, ...) de forma muy sencilla (`rmarkdown`).
  - Otros paquetes: `knitr`, `htmltools`, ...
- Notebooks
  - Permiten la ejecución interactiva (generación automática de resultados).
  - RStudio notebooks, Apache Zeppelin, Jupyter, ...

Yihui Xie (2015). *Dynamic Documents with R and knitr*. Chapman.

Desarrollo de aplicaciones



Rattle (Gtk)



Shiny

- Hay paquetes que permiten emplear librerías gráficas para la creación de aplicaciones:
  - `tcltk2`, `RGtk2`, `rJava`, ...

Lawrence y Verzani (2012). *Programming Graphical User Interfaces in R*. Chapman and Hall/CRC.

- `Shiny` permite crear aplicaciones web interactivas fácilmente (incluso se pueden crear con `rmarkdown`).
  - Otras alternativas: `opencpu`, `httpuv`, ...

Beeley (2015). *Web Application Development with R Using Shiny*. Packt Publishing.

Eso es todo...