Data Mining with R/ORE

Minming Duan

iTech Solution Profile

Agenda

1. R/ORE Overview
2. XML output generation using SQL
3. Integration with IBP and BIEE
4. Oracle R for Hadoop Connector
5. R vs. SPSS
6. FAQ

Why analysts use R

- R is a statistics language similar to Base SAS or SPSS statistics.
- R environment is…
  - • Powerful
  - • Extensible
  - • Graphical
  - • Extensive statistics
  - • OOTB functionality with many ‘knobs’ but smart defaults
  - • Ease of installation and use
  - • ***Free***

Limitations of R

R is a client and server bundled together as 1 executable - like Excel

- Single user tool
- Not multi-threaded
- Cannot leverage CPU capacity even on a user's laptop/desktop

R requires data it operates on to be first loaded into memory

- Loading data may not be a limitation given RAM available on laptops/desktops
- R's *call by value semantics means as data flows into functions, for each function invocation, many copies of the data are made*
- As a result you quickly run into memory limits

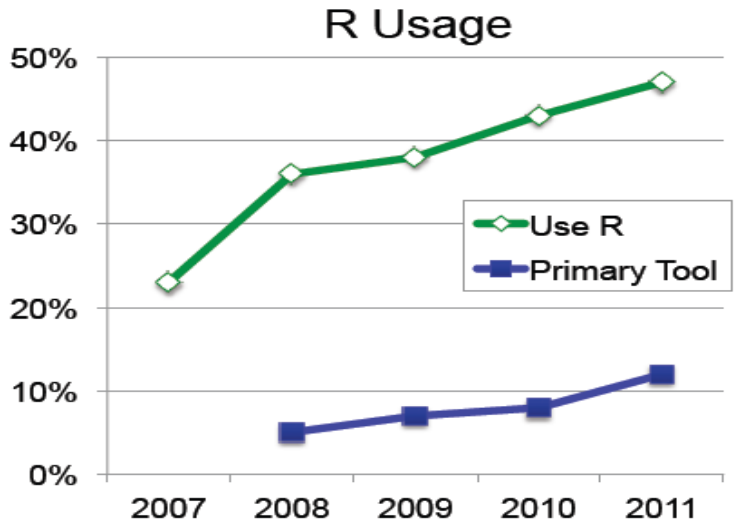
Why should you be interested in R?

- Emerging trends
  - It's the next "big thing" in advanced analytics
  - Colleges and universities use R for statistics classes
    (replacing more traditional software tools)
  - Advanced Analytics as a critical differentiator of the DWH technology stack
- Augment Oracle deployments
  - Enhance results with powerful graphics
  - Integrate R results and graphics with BI Publisher documents and OBIEE dashboards
- A scalable R via Oracle R Enterprise
  - Leverage Oracle-engineered solutions
  - A viable alternative to SAS/SPSS

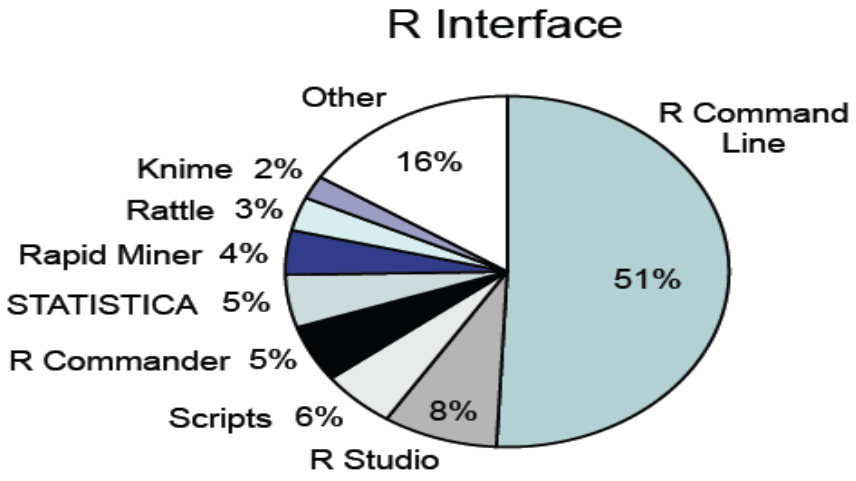
Rexer Analytics Survey 2011

The Popularity of R Software is Growing Fast

- The proportion of data miners using R is rapidly growing!
  - R is also the #1 most used data mining tool (in both 2010 & 2011). Up from #5 in 2007.
- An increasing number of data miners consider R their primary tool.
  - R is now #2 in primary tool rankings. Up from #7 in 2008.
- Half of R data miners use the command line interface. Among the rest, R Studio, scripts, R Commander, and STATISTICA are popular interfaces.

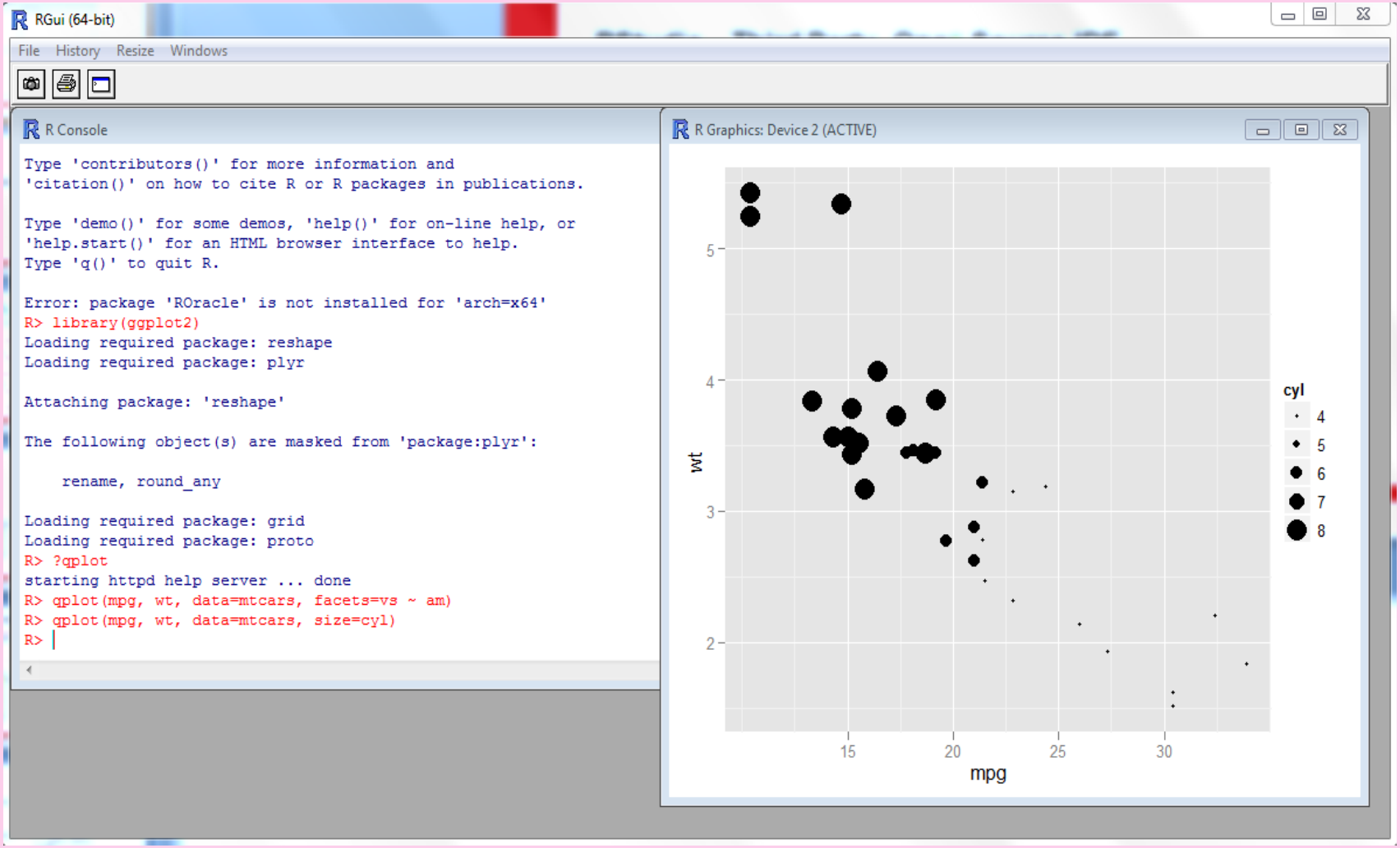


Vendors were excluded from these analyses.

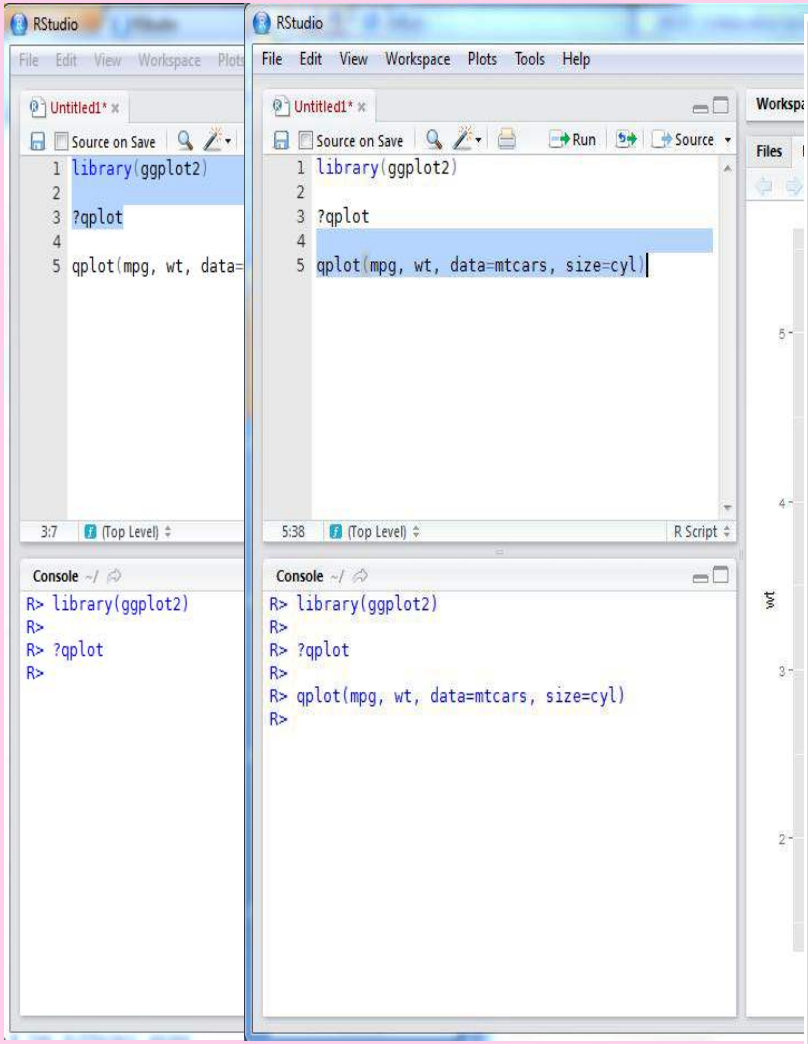


Question: If you use the R software package, what is your primary interface to R?

Default R GUI



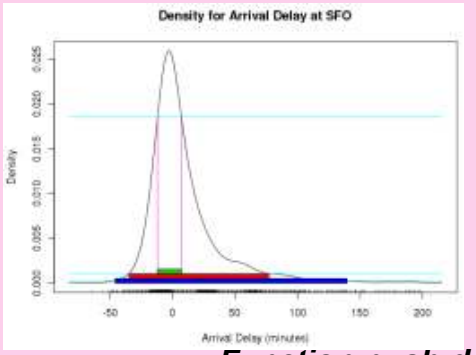
RStudio – Third Party, Open Source IDE



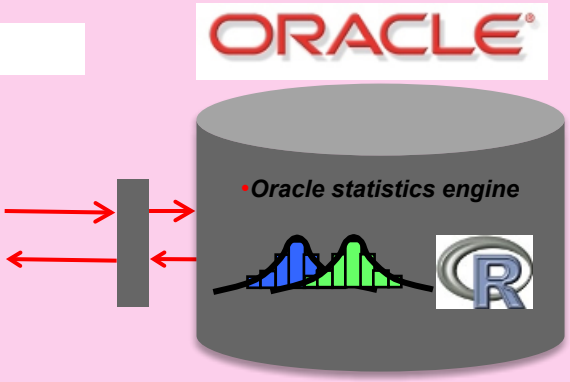
Which R interfaces do you use frequently?	
built-in R console (225)	40%
RStudio (135)	24%
Eclipse with StatET (90)	16%
RapidMiner R extension (80)	14.2%
Tinn-R (62)	11%
ESS (Emacs Speaks Statistics) (59)	10.5%
Rattle GUI (53)	9.4%
R Commander (43)	7.7%
Revolution Analytics (31)	5.5%
RKWard (22)	3.9%
JGR (Java Gui for R) (21)	3.7%
RExcel (18)	3.2%
R via a data mining tool plugin (12)	2.1%
Red-R (8)	1.4%
SciViews-R (6)	1.1%
Other (44)	7.8%

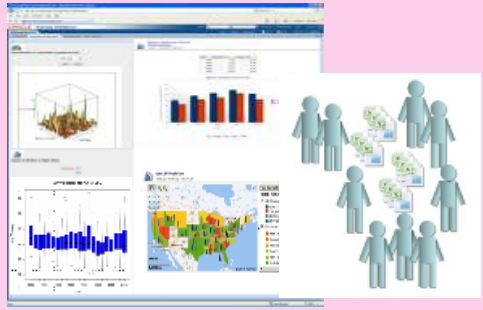
Oracle R Enterprise

•*R workspace console*



•*Function push-down*
– *data transformation & statistics*





•*OBIEE, Web Services*

•*No changes to the user experience*

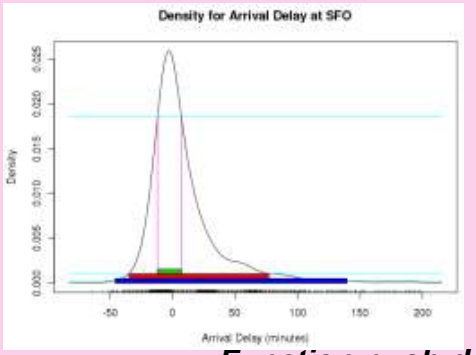
•*Scale to large data sets*

•*Embed in operational systems*

•*Development* → •*Production* → •*Consumption*

Oracle R Enterprise

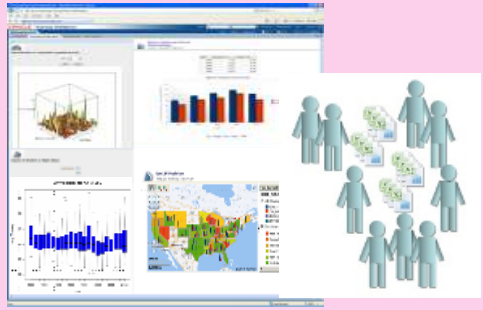
- *R workspace console*



- *Function push-down*
  - *data transformation & statistics*

ORACLE

- *Oracle statistics engine*



- *OBIEE, Web Services*

hadoop

- *Transparently leverage Hadoop for High Performance Analytics to Oracle Big Data Appliance (part of Big Data Connectors software suite)*

Oracle R Enterprise – Key messages

- *Most integrated and complete suite of Enterprise Advanced Analytics software available in the market today*

- *Substantial leap forward from incumbent platforms*
  - *Data volume – using SQL and existing DB functionality*
  - *Data Heterogeneity – Oracle DB + BDA*
  - *Breadth of Analytics – Oracle DB + R packages*
  - *Breadth of User Types – R+SQL+BI report developers, DBAs*

- *Enables enterprise-wide consumption of advanced analytics models via integration with Oracle Exalytics*

iTech Solution Profile

Agenda

1. R/ORE Overview
2. XML output generation using SQL
3. Integration with IBP and BIEE
4. Oracle R for Hadoop Connector
5. R vs. SPSS
6. FAQ

iTech Solution Profile

Agenda

1. R/ORE Overview
2. XML output generation using SQL
3. Integration with IBP and BIEE
4. Oracle R for Hadoop Connector
5. R vs. SPSS
6. FAQ