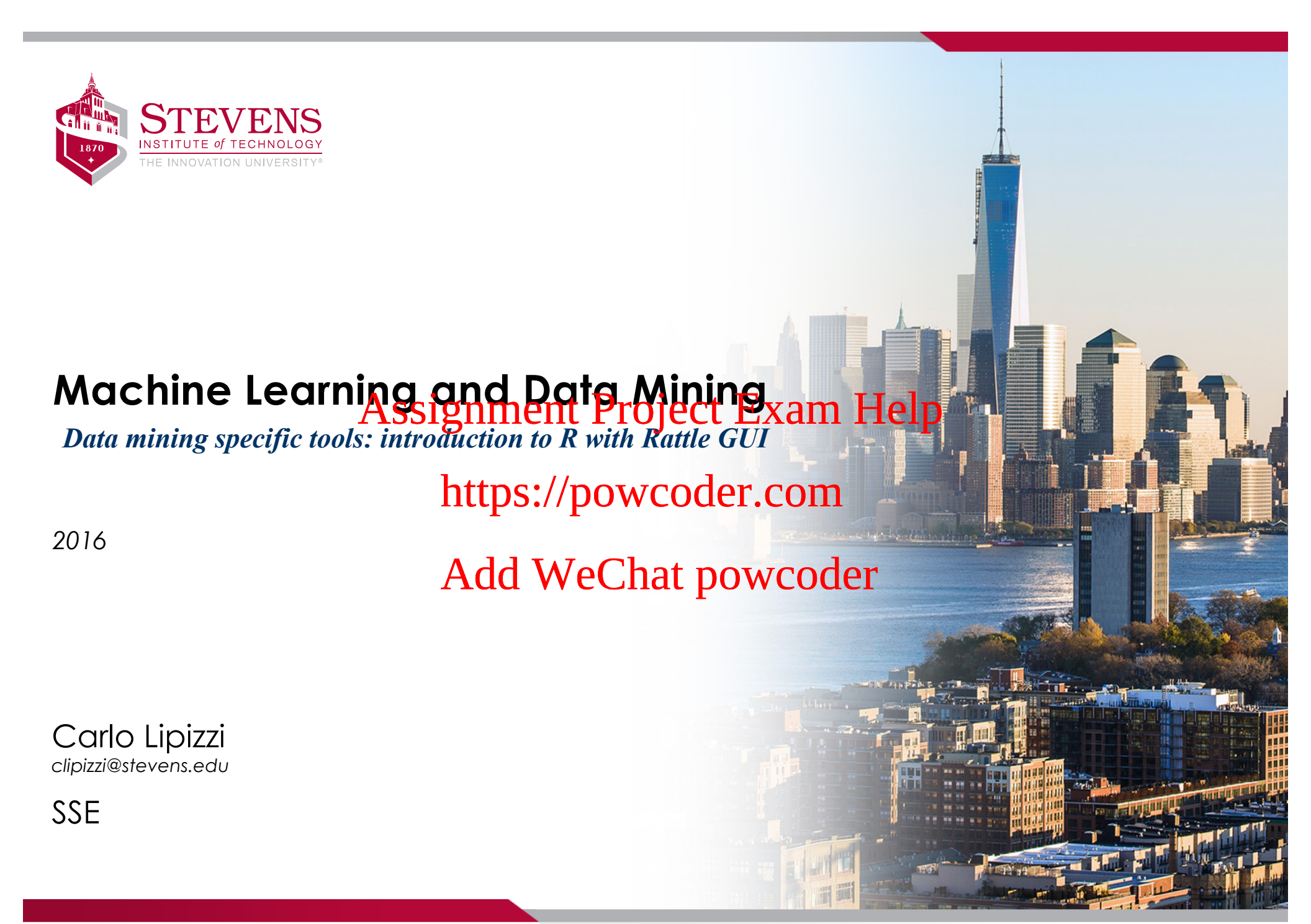


Machine Learning and Data Mining

Data mining specific tools: introduction to R with Rattle GUI

2016

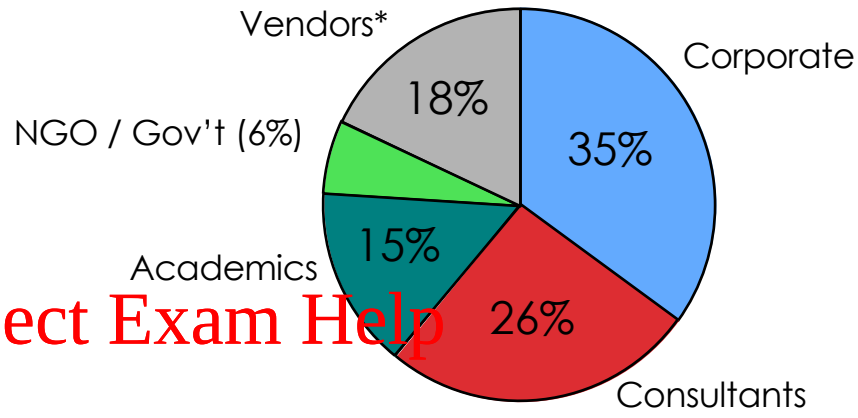
Carlo Lipizzi
clipizzi@stevens.edu

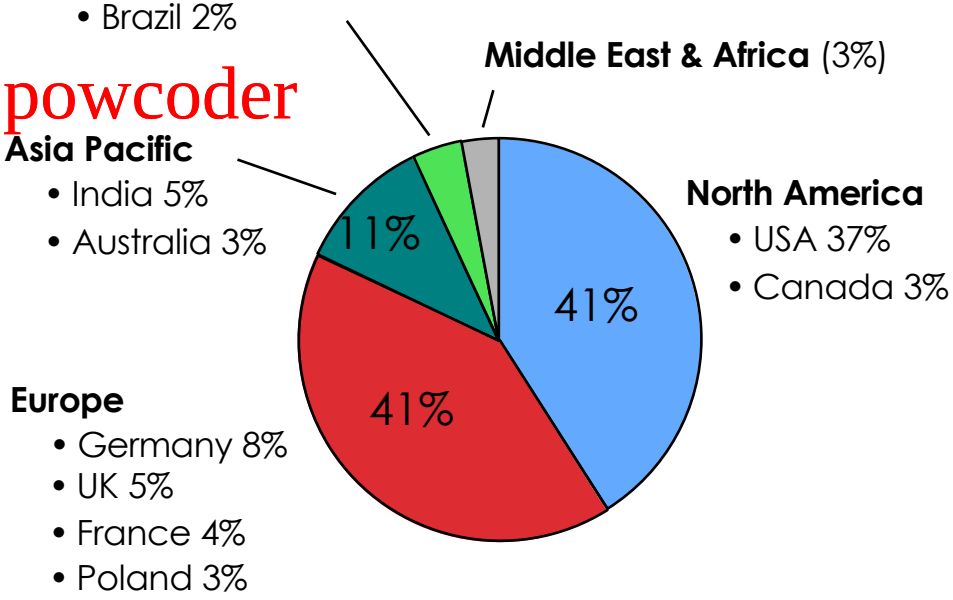
SSE

2013 Data Miner Survey: Overview

REXER ANALYTICS
ANALYTIC AND CRM CONSULTING

- **6th survey since 2007**
- **68 questions**
- **10,000+ invitations emailed, plus promoted by newsgroups, vendors, and bloggers**
- **Respondents:**
  - 1,259 data miners
  - from 75 countries
- **Data collected in first half of 2013**





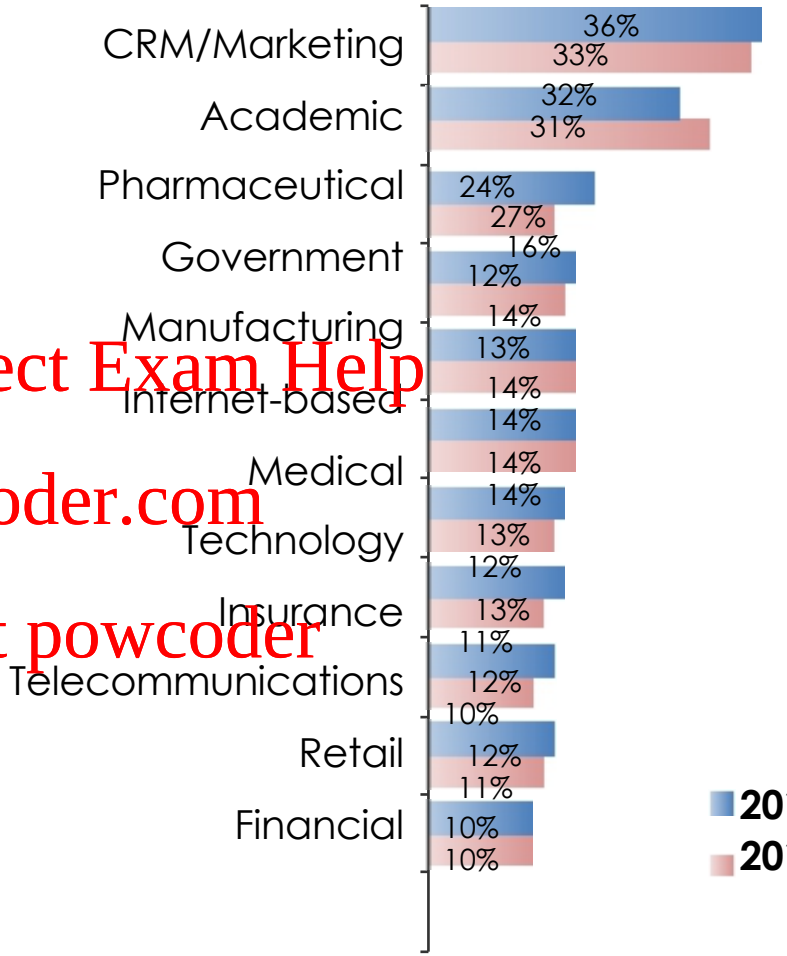
*Data from software vendors is excluded from analyses in this presentation unless otherwise noted

Some Key Findings

- FOCUS ON CRM: In the past few years, there has been an increase among data miners in the already substantial area of customer-focused analytics. Respondents are looking for a better understanding of customers and seeking to improve the customer experience. This can be seen in their goals, analyses, big data endeavors, and in the focus of their text mining.
- BIG DATA: Many in the field are talking about the phenomena of Big Data. There are clearly some areas in which the volume and sources of data have grown. However it is unclear how much Big Data has impacted the typical data miner. While data miners believe that the size of their datasets have increased over the past year, data from previous surveys indicate that the size of datasets have been fairly consistent over time.
- THE ASCENDANCE OF R: The proportion of data miners using R is rapidly growing, and since 2010, R has been the most-used data mining tool. While R is frequently used along with other tools, an increasing number of data miners also select R as their primary tool.

Applications

- CRM / Marketing remains the #1 area to which data mining is applied.
- The roots of data mining in customer focused analytics are strong. In each of the 6 Data Miner Surveys, more people report applying their analytics in the field of CRM / Marketing than in any other field.
- In 2013, 36% of data miners indicated that they are commonly involved in CRM / Marketing data mining, up slightly from 2011. The number of data miners working in the overlapping area of Retail analytics is also increasing

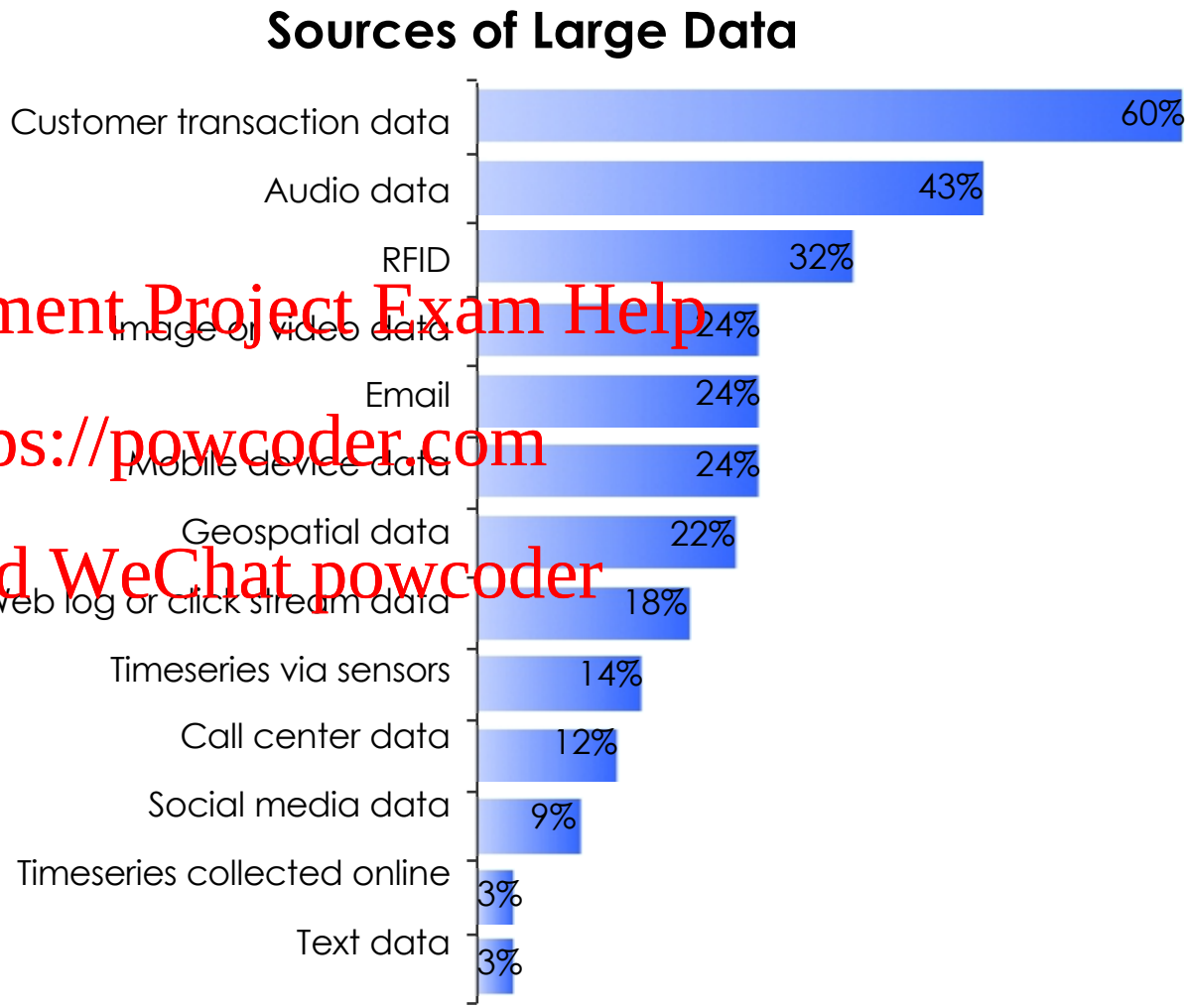


Data miners also report working in Non-profit (5%), Hospitality / Entertainment / Sports (4%), Military / Security (2%), and Other (10%).

Question: In what fields do you TYPICALLY apply data mining? (Select all that apply)

Customer Transactions: #1 Source of Large Data

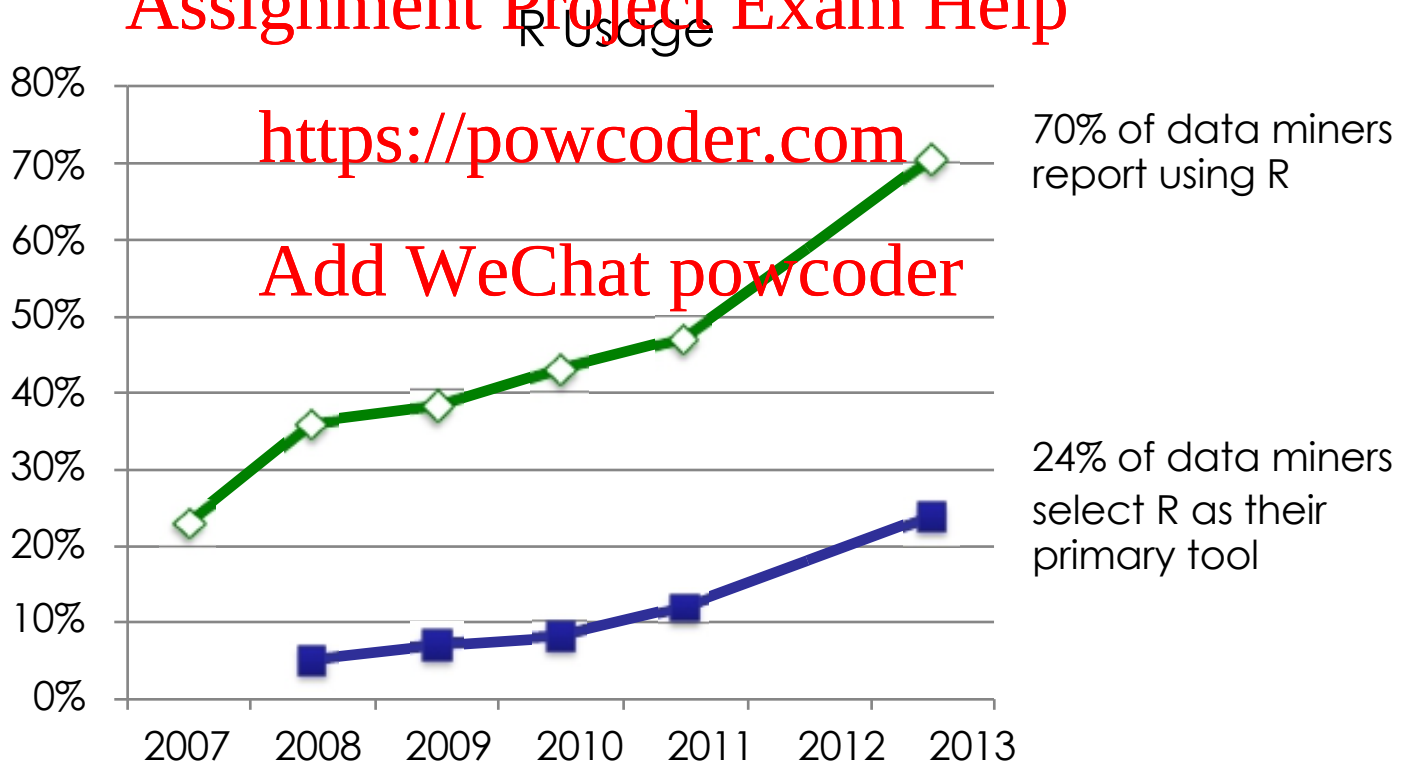
- Customer transactional data often affords the opportunity for a wide range of analytics due to the depth and scope of available data
- Among respondents who reported increases in data volume, 60% identified customer transaction data as a source of their large data sets



Question: What are the sources of data for your large datasets? (select all that apply)

The Popularity of R

The proportion of data miners using R is rapidly growing, and since 2010, R has been the most-used data mining tool. While R is frequently used along with other tools, an increasing number of data miners also select R as their primary tool. Among data miners who say they are likely to switch their primary package in the coming year, R is frequently identified as the tool they are plan to switch to – more than 2.5 times more often that any other tool



Priorities and Characteristics of R Users

- While data miners overall consider quality and accuracy of model performance, dependability of software, and data manipulation capabilities the most important factors when choosing a data mining tool, those using R as their primary tool identify the ability to write one's own code as their most important priority.

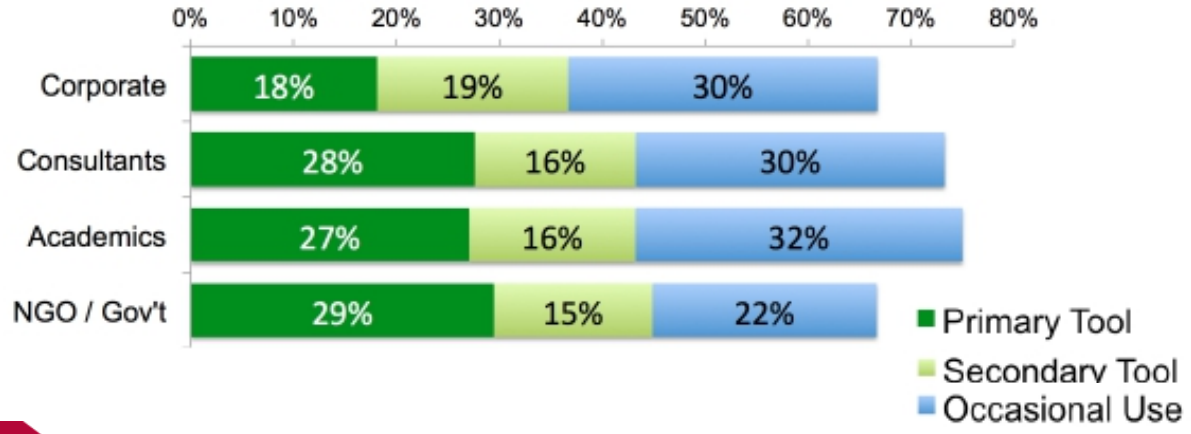
Important Factors in Selecting Software

R is primary tool	All data miners
#1: Ability to write own code	#1: Quality & accuracy of model performance
#2: Quality & accuracy of model performance	#2: Dependability of software
#3: Data manipulation capabilities	#3: Data manipulation capabilities

- The quality of the user interface was rated as significantly less important by primary R users than by other data miners.

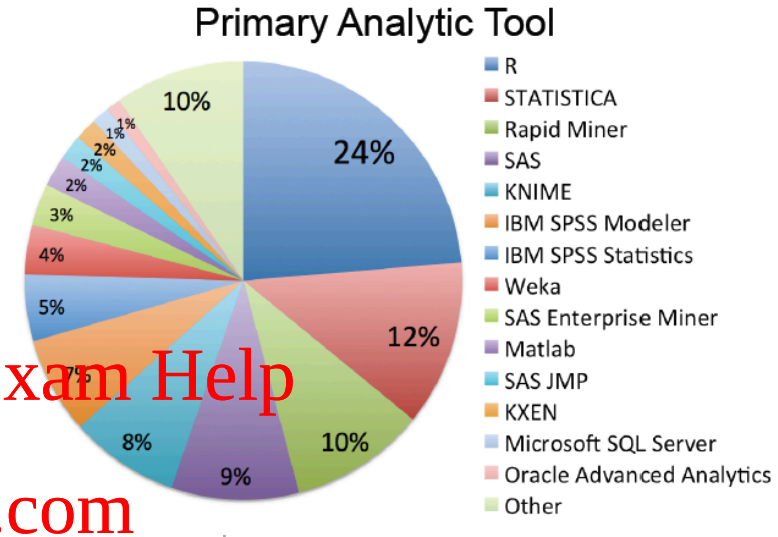
- Interestingly, there was no difference in the stated importance of cost of tool between those using R as their primary package and others. However, primary R users are more satisfied than other tool users with the cost of their software (see page 33). They are also more satisfied with the variety of available algorithms and the ability to modify algorithms to fine-tune analyses.

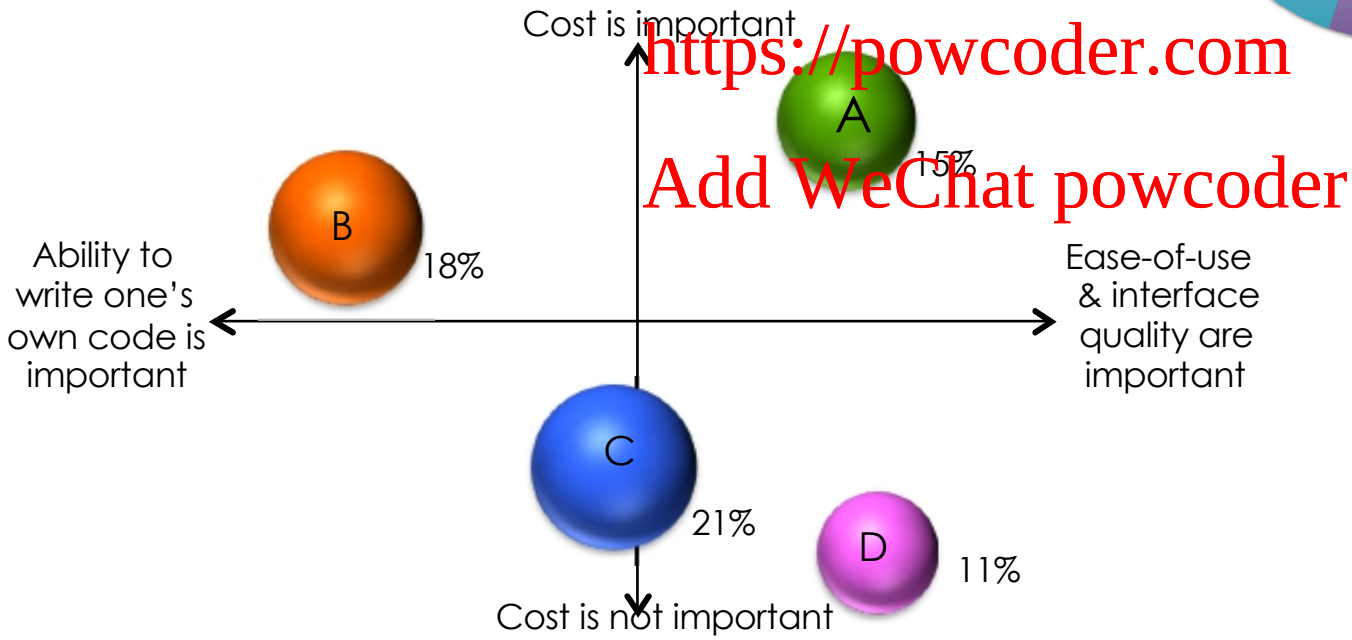
- While R is heavily used among data miners working in all settings, in corporate settings, a smaller proportion of data miners report that R is their primary tool.



Tool Selection

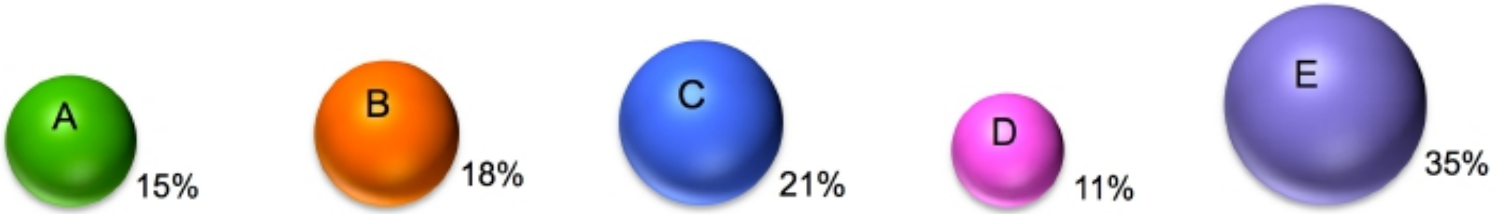
- Data miners are a diverse group who are looking for different things from their data mining tools. They report using multiple tools to meet their analytic needs, and even the most popular tool is identified as their primary tool by just 24% of data miners. Over the years, R and Rapid Miner have shown substantial increases
- Cluster analysis* reveals that, in their tool-selection preferences, data miners fall into 5 groups. The primary dimensions that distinguish them are price sensitivity and code-writing / interface / ease-of-use preferences





*Cluster analysis was conducted on data miners' ratings of the importance of 22 tool selection factors.

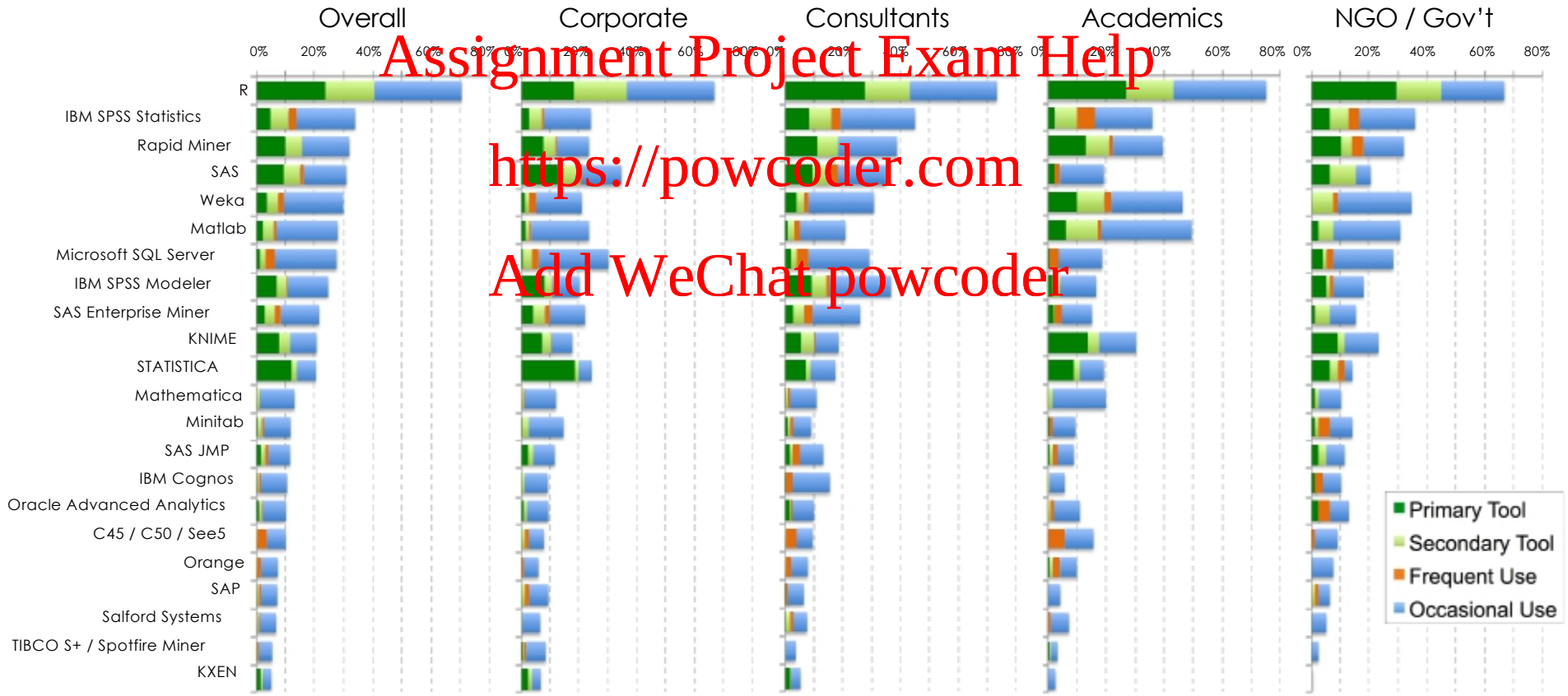
Tool Selection Groups



	A	B	C	D	E
Importance of cost	Very high	High	Moderate	Low / Moderate	Very high
Importance of ease-of-use	High	Low / Moderate	Moderate	High	Very high
Importance of user interface quality	High	Low	High	Very high	Very high
Importance of ability to write one's own code	Low	Very high	High	Low	High
Primary tools	Rapid Miner (26%) IBM Modeler (12%) KNIME (11%)	R (56%) SAS (10%)	R (26%) SAS (19%)	STATISTICA (31%) IBM Modeler (20%) Rapid Miner (12%)	R (19%) STATISTICA (16%) KNIME (10%) Rapid Miner (10%)
Tool use	R (62%) Rapid Miner (50%) IBM Statistics (40%) IBM Modeler (36%) Weka (33%)	R (90%) Weka (37%) SAS (33%) Matlab (31%)	R (73%) SAS (43%) IBM Statistics (35%) Matlab (32%) SQL Server (32%) SAS-EM (32%)	R (51%) IBM Statistics (38%) STATISTICA (37%) IBM Modeler (32%)	R (73%) IBM Statistics (35%) Rapid Miner (34%) Weka (32%) SQL Server (30%) SAS (30%)
Working with Big Data	---	---	---	Less Likely	More Likely
Experience (years)	Many new data miners	Few new data miners	---	Many new data miners	Many experienced data miners

Tool Use Varies by Employment Setting

- R, IBM SPSS Statistics, Rapid Miner, and SAS are the software tools used by the most data miners. The average data miner reports using 5 tools, but conducts 76% of their work in their primary tool. R, STATISTICA, Rapid Miner, and SAS are the primary data mining tools chosen most often. 64% of data miners also report writing their own code – the most common language is SQL (43%), followed by Java (26%) and Python (24%)
- The graphs below summarize the patterns of primary tool selection and overall tool usage, which vary by the setting in which data miners work – e.g., academics are heavier users of Weka and Matlab



R: Introduction

- R is "GNU S" — A language and environment for data manipulation, calculation and graphical display.
  - R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al.
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for interactive data analysis,
  - graphical facilities for data analysis and display either directly at the computer or on hardcopy
  - a well developed programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- The core of R is an interpreted computer language.
  - It allows branching and looping as well as modular programming using functions.
  - Most of the user-visible functions in R are written in R, calling upon a smaller set of internal primitives.
  - It is possible for the user to interface to procedures written in C, C++ or FORTRAN languages for efficiency, and also to write additional primitives.

What R does and does not

- data handling and storage: numeric, textual
- matrix algebra
- hash tables and regular expressions
- high-level data analytic and statistical functions
- classes ("OO")
- graphics
- programming language: loops, branching, subroutines

- is not a database, but connects to DBMSs
- has no graphical user interfaces, but connects to Java, TclTk
- language interpreter can be very slow, but allows to call own C/C++ code
- no spreadsheet view of data, but connects to Excel/MsOffice
- no professional / commercial support

R and statistics

- Packaging: a crucial infrastructure to efficiently produce, load and keep consistent software libraries from (many) different sources / authors
- Statistics: most packages deal with statistics and data analysis
- State of the art: many statistical researchers provide their methods as R packages

Usability

What is Usability?

- Ease of learning
- Ease of use
- Ease of remembering
- Subjective satisfaction
- Efficiency of use
- Effectiveness of use

Usability Engineering

- Usability Engineering (UE):
  - Processes to build "Usability" into products
  - Various methods can be used throughout the design lifecycle
  - Methods can be incorporated into design process easily
  - Methods maintain focus on user throughout design

Benefits of UE to an Organization

- Reduce training costs
- Reduce development costs
  - Identify and fix problems early
- Reduce support costs; minimize need for
  - support personnel/help desks
  - fixes, maintenance, upgrades
- Enhance organization's reputation – positive "word-of-mouth" trade
- Larger numbers of "hit" and "return visit" rates

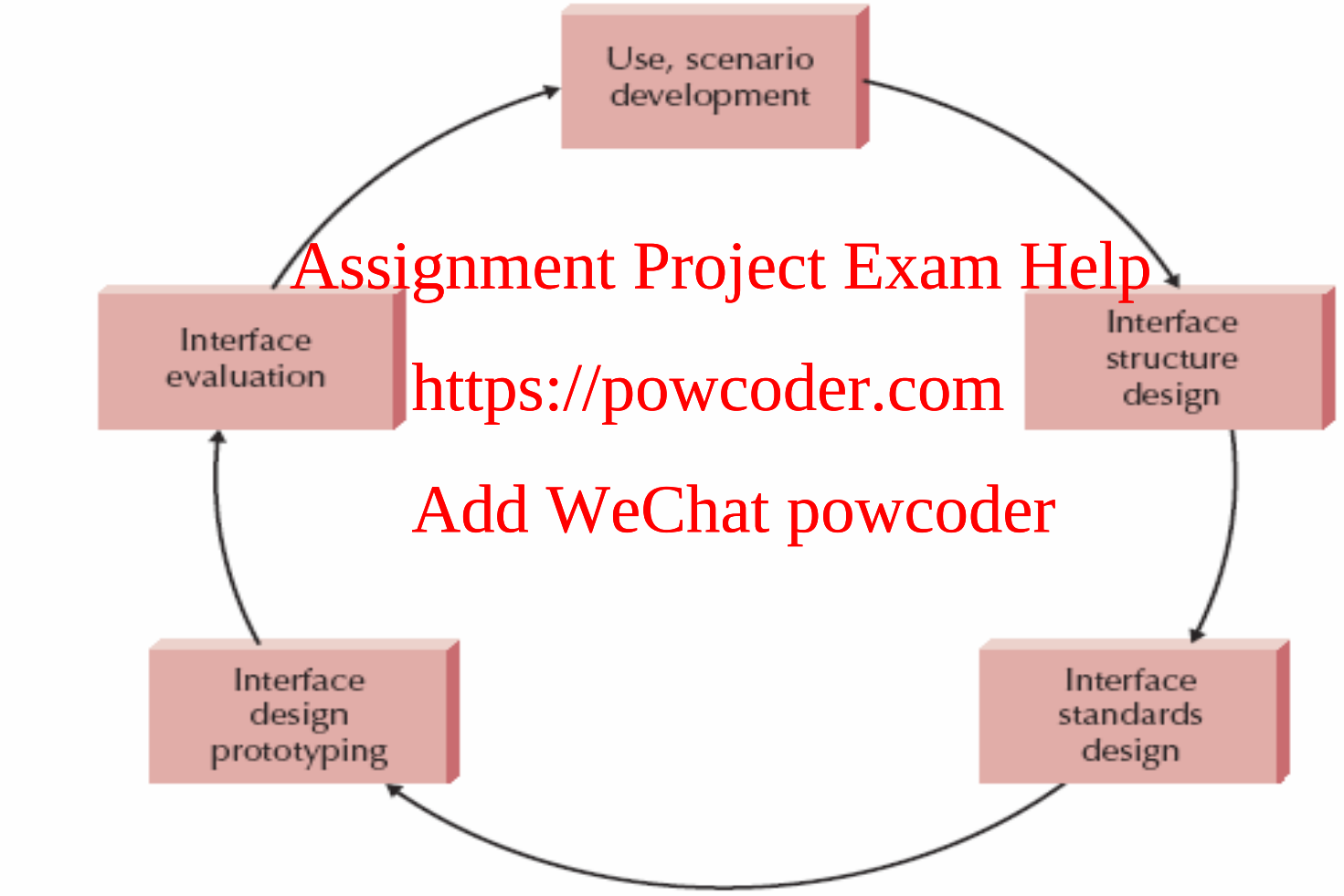
Benefits of UE to the User

- Less time to complete work
- Greater success with tasks
- Increased user satisfaction

GUI Design is Multi Disciplinary

- A team includes
  - Analyst
  - Designer
  - Technology expert
  - Graphic artist
  - Social and behavioral scientist
  - Programmer

Usability Design Process



GUI for R

- Statistics can be complex and traps await
- So many tools in R to deliver insights
- Effective analyses should be scripted
- Scripting also required for repeatability
- R is a language for programming with data

- How to remember how to do all of this in R?
- How to skill up 150 data analysts with Data Mining?

Poll from KDnuggets.com

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%

Deducer

- “An intuitive, cross-platform graphical data analysis system”
- Deducer is based upon rJava and provides access to the Java Swing Network

Related Packages	Description
DeducerExtras	Additional dialogs and functions for Deducer
DeducerPlugIn Example	Deducer Plug-In Example
DeducerPlugInScaling	Reliability and factor analysis plugin
DeducerSpatial	Deducer for spatial data analysis
DeducerSurvival	Add Survival Dialogue to Deducer
DeducerText	Deducer GUI for Text Data

Rattle

- Today, Rattle is used world wide in many industries
  - Health analytics
  - Customer segmentation and marketing
  - Fraud detection
  - Government
- It is used by
  - Consultants and Analytics Teams across business
  - Universities to teach Data Mining
- It is and will remain freely available.

- CRAN and http://rattle.togaware.com

Walk-through

- Loading Data
- Basic Data Exploration
- Explore Distribution
- Explore Correlations

GUI limitations

- Data miners are programmers of data
- A GUI can only do so much
- R is a powerful statistical language

- Professional data mining
- Scripting
- Transparency
- Repeatability

- The log tab: a bridge from GUI to CLI (Command Line Interface)

Exercise 1 – Companies Values – 15’

- **Data set:** *Companies1.xlsx*
  - 25 companies

- **For the 3 numeric variables calculate:**

  - Mean, Mode, Median, Max, Min, Range

- **Can you get any non explicit info from the values you calculated?**
- **Can you create any new variables to get more from your data? What is your goal?**

Exercise 2 on Handling Missing Data – 15’

- Examine *cars.txt* dataset containing records for 261 automobiles manufactured in 1970s and 1980s
- Examine the file and handle missing data

Exercise 3 - 15’

Handling Outliers

- Examine *cars_full.txt* dataset containing full records for 261 automobiles manufactured in 1970s and 1980s
- Examine the file for outliers

Normalization

- Using the *cars_full.txt* dataset normalize the “time-to-60” values
- Use the [0-1] scale

Exercise 4 on Data Analysis

- Using the *cars_full.txt* dataset
- Class will be divided into 3 groups, focusing on
  - Mileage (mpg)
  - Power (hp)
  - Performance (time-to-60)
- Questions:
  - Which are the more correlated variables? Explain
  - What is the impact of "brand" (Country of origin)?
  - Are there differences in normalized vs non-normalized analyses (apply normalization to all the variables)?
- Each group will present their results