Business Analytics and Data Science

INTRODUCTION

Know your Instructor

http://linkedin.com/in/ajayohri

- Author "R for Business Analytics"
- Author " R for Cloud Computing"
- Founder "Decisionstats.com"
- University of Tennessee, Knoxville MS (courses in statistics and computer science)
- MBA (IIM Lucknow,India-2003)
- B.Engineering (DCE 2001)

Classroom Rules

- From Instructor

- From Audience
  - mobile phones should be kindly switched off
    - Yes, this includes Whatsapp
  - Ask Questions at end of session
  - Take Notes
  - Please Take Notes

Introduce Yourself

Name
Education Degree from Institute
Work Ex in Years in Domain

Introduce Yourself

Name
Education Degree from Institute
Work Ex in Years in Domain
What expectations from this training

Expectations

How Data Science can help your career ?

Support Team

Madhuresh

Introduction to Data Science

Course Outline

Class 1 : Introduction

- Understanding Business Analytics
- Introduction to RFM Analysis and LTV Analysis
- Refresher in Statistics
- Basics of Data Driven Decision Making
- Installation of R, Rtools, R Studio, R packages and GUIs
- Using RStudio and Using GUIs

Quiz 1

Class 2 : Data Input

- Data Input using standard data sources
- Data Input for large data (RDBMS)
- Using SQL from within R
- Web Scraping
- Specialized Packages for other data sources

Quiz 2

Dataset - Iris

Introduction to Data Science

Information Ladder

The **information ladder** was created by education professor Norman Longworth to describe the stages in human learning.

According to the ladder, a learner moves through the following progression to construct “wisdom”from “data”

Data →

Information →

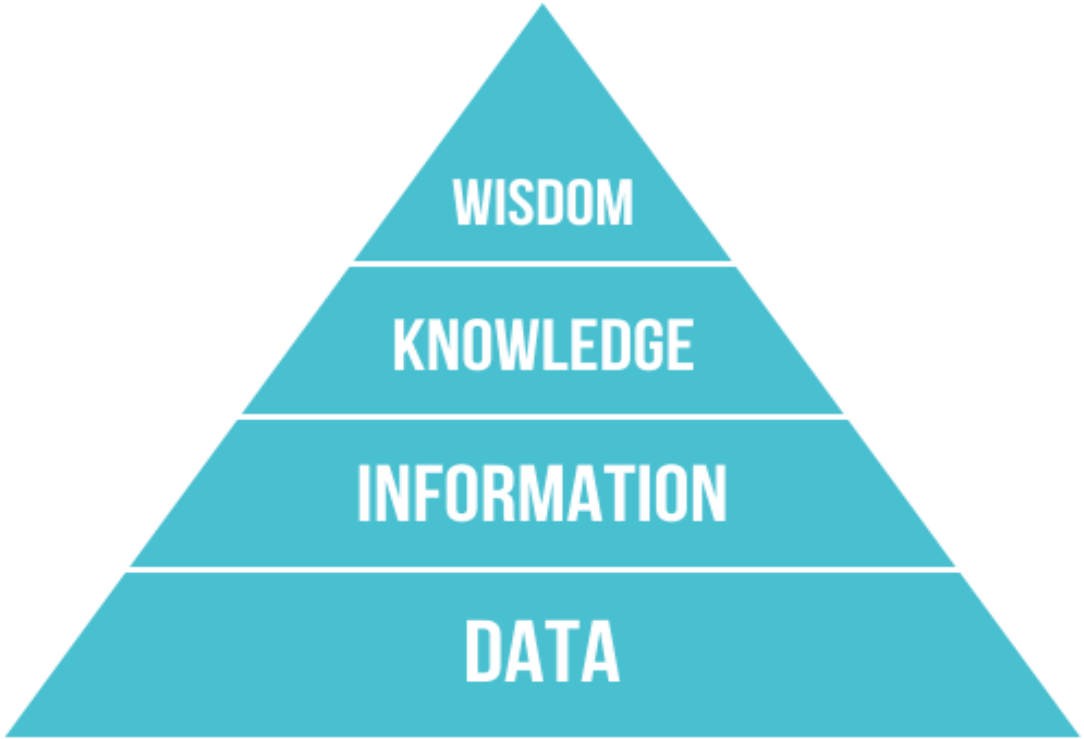
Knowledge →

Understanding →

Insight →

Wisdom

DIKW



Basics of Data Science

Data Science is the extraction of knowledge from data,[1][2] which is a continuation of the field data miningand predictive analytics, also known as knowledge discovery and data mining (KDD). It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information theory and information technology, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing,data compression, computer programming, and high performance computing. Methods that scale to Big Data are of particular interest in data science, although the discipline is not generally considered to be restricted to such data. The development ofmachine learning, a branch of artificial intelligence used to uncover patterns in data from which predictive models can be developed, has enhanced the growth and importance of data science.

CONFUSING?

http://en.wikipedia.org/wiki/Data_science

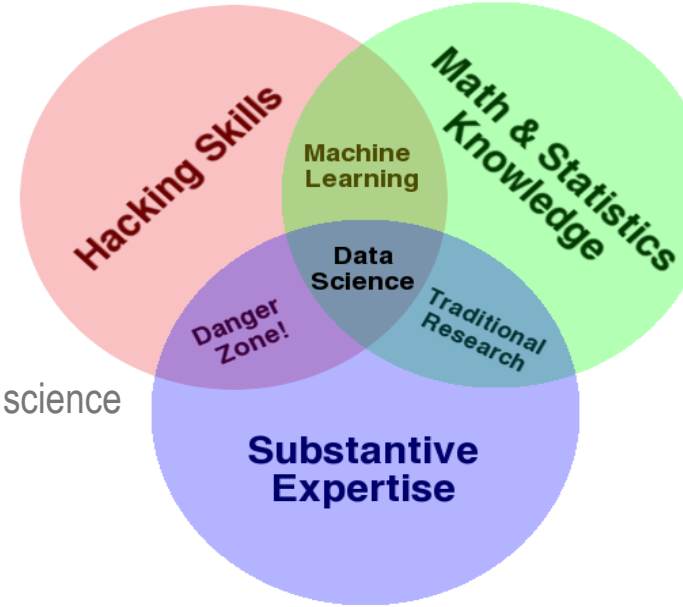
Basics of Data Science

the culture of academia, which does not reward researchers for understanding technology.

DANGER ZONE- this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created

Being able to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically; these are the hacking skills that make for a successful data hacker.

data plus math and statistics only gets you machine learning, which is great if that is what you are interested in, but not if you are doing data science



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Business Intelligence

Business intelligence (BI) is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.

The key general categories of business intelligence tools are:

- Spreadsheets
- Reporting and querying software: tools that extract, sort, summarize, and present selected data
- OLAP: Online analytical processing
- Digital dashboards
- Data mining
- Data warehousing
- Local information systems

Definition – study of business data using statistical techniques and programming for creating decision support and insights for achieving business goals

Predictive- To predict the future.

Descriptive- To describe the past.

So what is a Data Scientist ?

a data scientist is simply a data analyst living in **california**

What is a Data Scientist

a data scientist is simply a person who can

write code
understand statistics
derive insights from data

Oh really, is this a Data Scientist ?

A data scientist is simply a person who can

write code = in R,Python,Java, SQL, Hadoop (Pig,HQL,MR) etc

= for data storage, querying, summarization, visualization

= how efficiently, and in time (fast results?)

= where on databases, on cloud, servers

and understand **enough** statistics

to derive **insights** from data

so **business** can make **decisions**

Guide for Data Scientists

http://www.kdnuggets.com/2014/05/guide-to-data-science-cheat-sheets.html

By Ajay Ohri, May 2014.

Over the past few years, as the buzz and apparently the demand for data scientists has continued to grow, people are eager to learn how to join, learn, advance and thrive in this seemingly lucrative profession. As someone who writes on analytics and occasionally teaches it, I am often asked - How do I become a data scientist?

Adding to the complexity of my answer is data science seems to be a multi-disciplinary field, while the university departments of statistics, computer science and management deal with data quite differently.

But to cut the marketing created jargon aside, a data scientist is simply a person who can write code in a few languages (primarily R, Python and SQL) for data querying, manipulation , aggregation, and visualization using enough statistical knowledge to give back actionable insights to the business for making decisions.

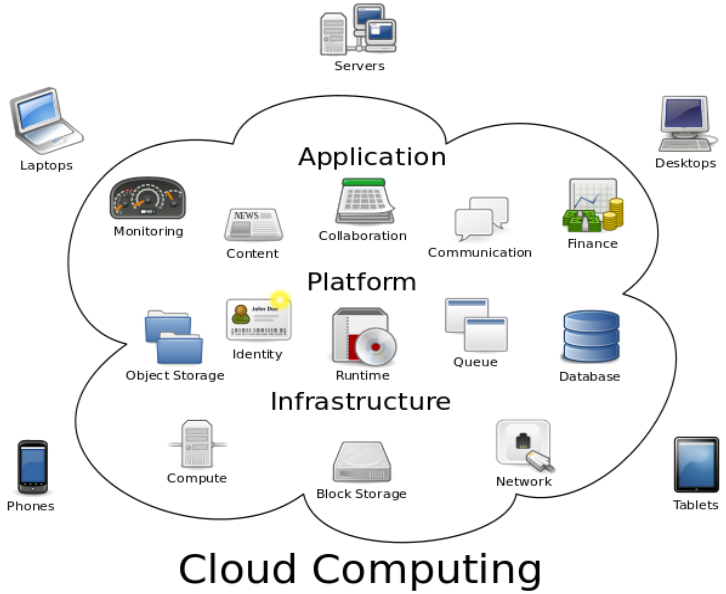
http://www.slideshare.net/ajayohri/cheat-sheets-for-data-scientists

- Business Analytics
    - Understanding what solution business needs

- Data Science
    - Primarily R programming skills
    - Some Applied Statistical Methods
    - Exposure to new domains and techniques

Cloud Computing

1. the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.



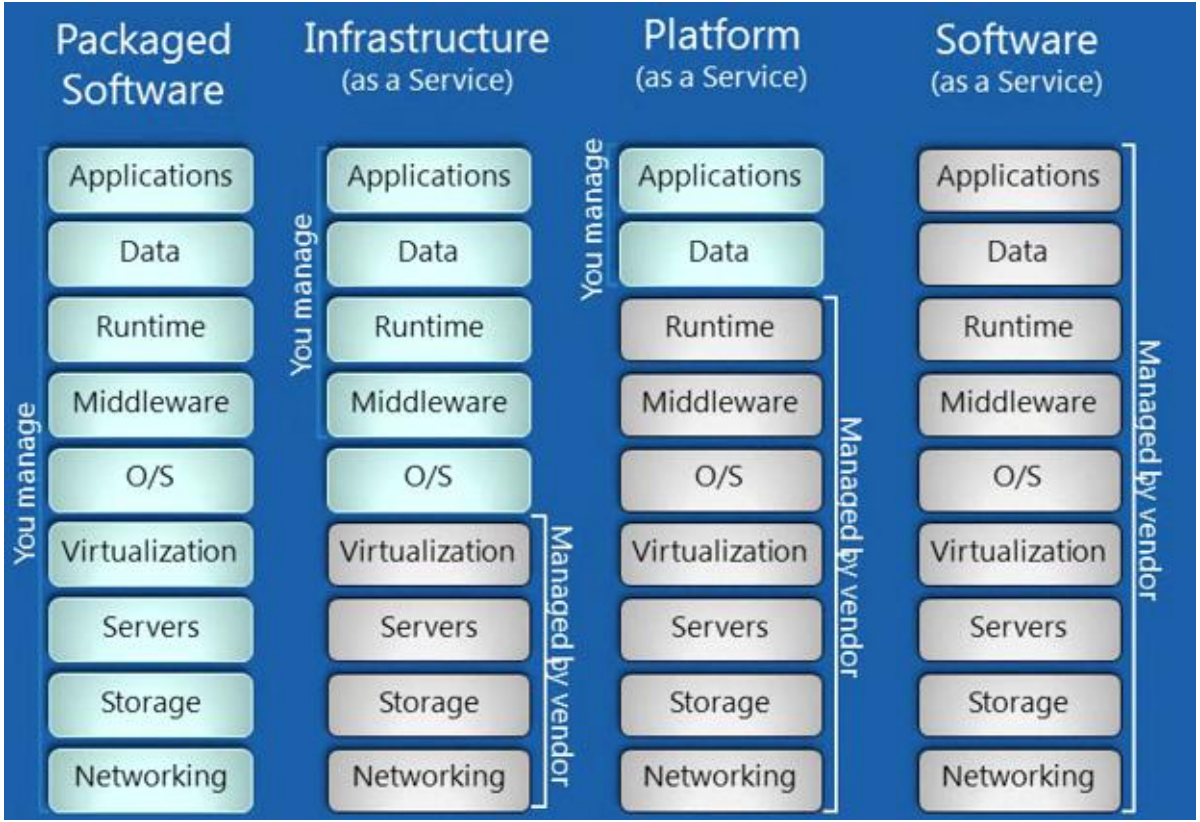
http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

Cloud Computing

1. the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

Cloud Computing



LTV Analytics

Life Time Value (LTV) will help us answer 3 fundamental questions:

1. Did you pay enough to acquire customers from each marketing channel?

2. Did you acquire the best kind of customers?

3. How much could you spend on keeping them sweet with email and social media?

Breaking Down LTV Further

LTV WILL BE DIFFERENT FOR DIFFERENT KINDS OF CUSTOMERS

Step 2 in this graphic is intended to help you determine LTV as a total average (an average of all your customers). To do this, companies will typically average the data from randomly chosen customers (as shown in Step 1 above). Sometimes it's helpful to break down the average further and perform separate LTV calculations for different kinds of customers. Try and segment your customer base by total purchases over a long time period, and it will help you determine the LTV of a "good" customer versus an "average" one. This type of analysis will help you determine how much more you should pay in order to acquire a "good" customer. See chart below.

INVESTING IN "GOOD" CUSTOMERS

Companies should be worried about the lasting impact of "buying cheap customers." How likely are these customers to buy another product, or hang around for a few years? Sometimes it pays to invest in "good" customers. "Good" customers might cost more to acquire, but they'll likely be more profitable as well.

Let's say that the LTV of an "average" customer is $8,000, and the LTV of a "good" customer is $10,000. By subtracting the two LTVs, you can see that you might expect to pay $2,000 more to acquire "good" customers.

Customer Satisfaction Boosts LTV

One of the most effective ways to boost LTV is to increase customer satisfaction. Research has found that a 5% increase in customer retention can increase profits by 25% to 95%. The same study found that it costs six to seven times more to gain a new customer than to keep an existing one.

LTV Analytics

Questions. Fill in the yellow boxes and the spreadsheet will take care of the rest.

	Best Customers	Average Customers
Acquisition Cost .How much did you pay to acquire these customers?	£40.00	£12.00
Average order value. How much do they spend per order?	£92.00	£70.00
Orders per year? Quite simply, How many orders do they place per year?	5	2
Retention? How many years will they be customers for?	3	2
Net profit? What is the net profit percentage of goods sold?	10%	10%

Answers. These cells will be magically calculated based on the values you put in the table on the left

	Best Customers	Average Customers
Lifetime Gross Revenue	£1,380.00	£280.00
Life Time Net Profit	£98.00	£16.00

http://www.kaushik.net/avinash/analytics-tip-calculate-ltv-customer-lifetime-value/

LTV Analytics

Questions. Fill in the yellow boxes and the spreadsheet will [illegible] of the rest.

Screen
Full Screen

	Year 1	Year 2	Year 3	Year 4	Year 5
r Segment. How many of a specific group of customers will you start with?					
Acquisition Cost? How much did you pay for each new customer? We won't use this figure - see note to explain					
Retention Rate. What % of customers will you keep from one year to the next?				75%	80%
Total Orders. How many orders/sales per customer per year? They may place more in future years	3	3	4	4	5
Average order value. How much is each sale or order worth, and will this rise over time?	£60.00	£65.00	£70.00	£75.00	£80.00
Net Profit. What % of each order is left after all costs have been accounted for?	10%	12%	12%	15%	15%
Discount Rate. This recognises our money *could* be better spent on something else - see note to explain				0.729	0.656

This worksheet will give you an individual lifetime value and the total revenue/profit for a group of customers. They could

Life Time Value often looks at existing customers who may have been acquired years ago. If this really is Year 1 of a customer's life with you, you could subtract this figure from the first year's net profit, but this worksheet won't do it for you!

Some companies include this in their LTV calculations, especially where the investment is high over a long time period. Just set all the fields to 1 if you'd prefer to ignore it!

Answers. These cells will be magically calculated based on the values you put in the table above.

	Year 1	Year 2	Year 3	Year 4	Year 5
Total Customers. The number of customers at the start of each year from the original segment	3,000	1,800	1,170	819	614
Total Revenue per Customer. This is the total revenue per year for individual customers	£180	£195	£280	£300	£400
Total Revenue. Annual revenue generated by all the customers in that year	£5,40,000	£3,51,000	£3,27,600	£2,45,700	£2,45,700
Cumulative Revenue. The revenue generated from the (remaining) original customers every year	£5,40,000	£8,91,000	£12,18,600	£14,64,300	£17,10,000
Annual Net profit per customer. Simply, the profit each customer generates in that year.	£18.00	£23.40	£33.60	£45.00	£60.00
Total Net Profit. Profit generated by all the original customers in that year.	£54,000	£42,120	£39,312	£36,855	£36,855
Profit at Net Present Value. The profit made each year, even if we offset a better way of spending it!	£54,000	£37,908	£31,843	£26,867	£24,177
Cumulative Net Profit at NPV. The profit generated in successive years from the original customers.	£54,000	£96,120	£1,35,432	£1,72,287	£2,09,142
Individual LTV at NPV. The cumulative amount of net profit each original customer is worth each year.	£18.00	£32.04	£45.14	£57.43	£69.71

nstructions | Comparison LTV | Detailed LTV | Sheet1

LTV Analytics

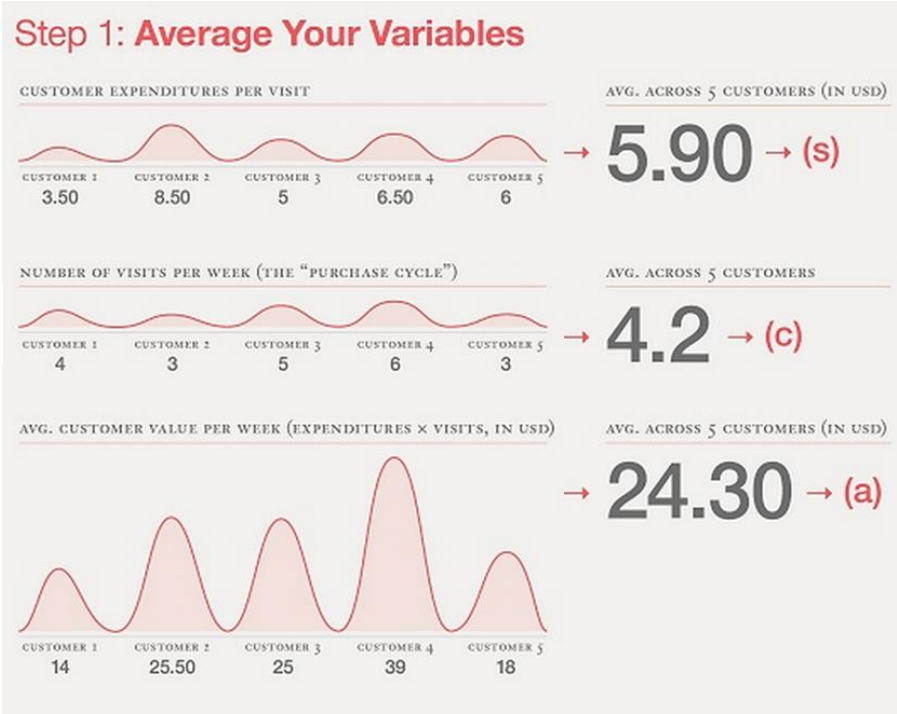
Download the zip file from http://www.kaushik.net/avinash/avinash_ltv.zip

Do the class exercise based on numbers given by instructor

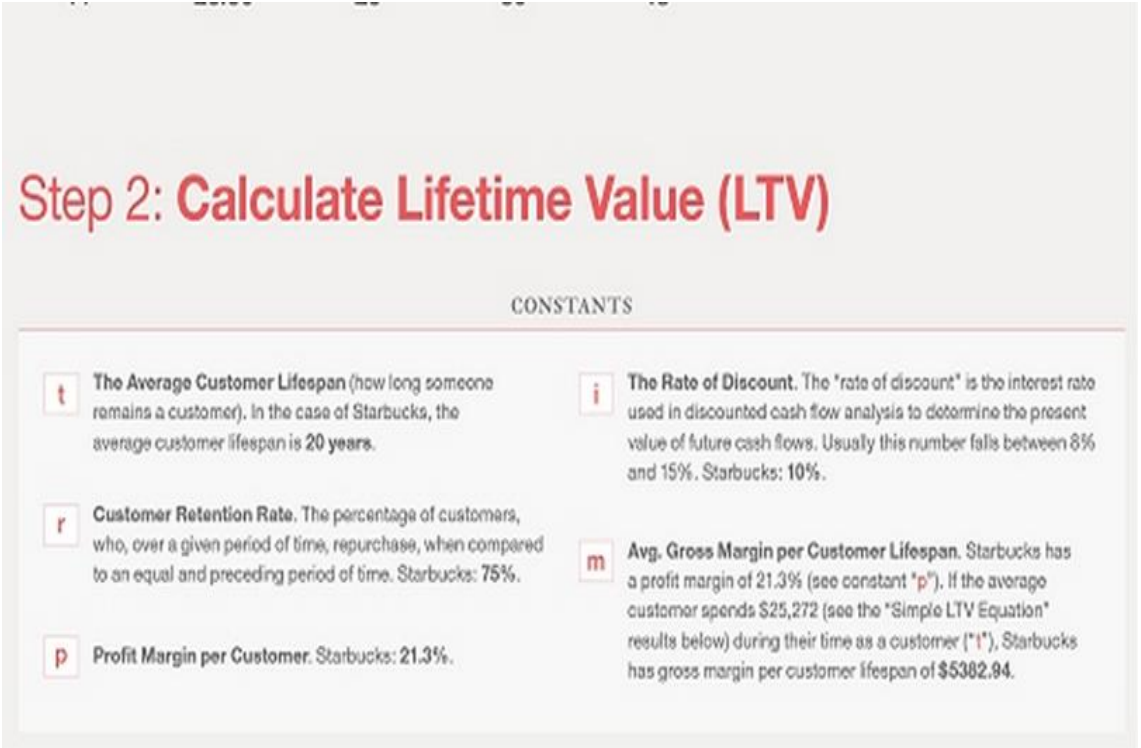
Give a brief supporting statement on analysis

LTV Analytics :Another Approach

https://blog.kissmetrics.com/how-to-calculate-lifetime-value/



LTV Analytics



Step 2: Calculate Lifetime Value (LTV)

CONSTANTS

t **The Average Customer Lifespan** (how long someone remains a customer). In the case of Starbucks, the average customer lifespan is **20 years**.

r **Customer Retention Rate.** The percentage of customers, who, over a given period of time, repurchase, when compared to an equal and preceding period of time. Starbucks: **75%**.

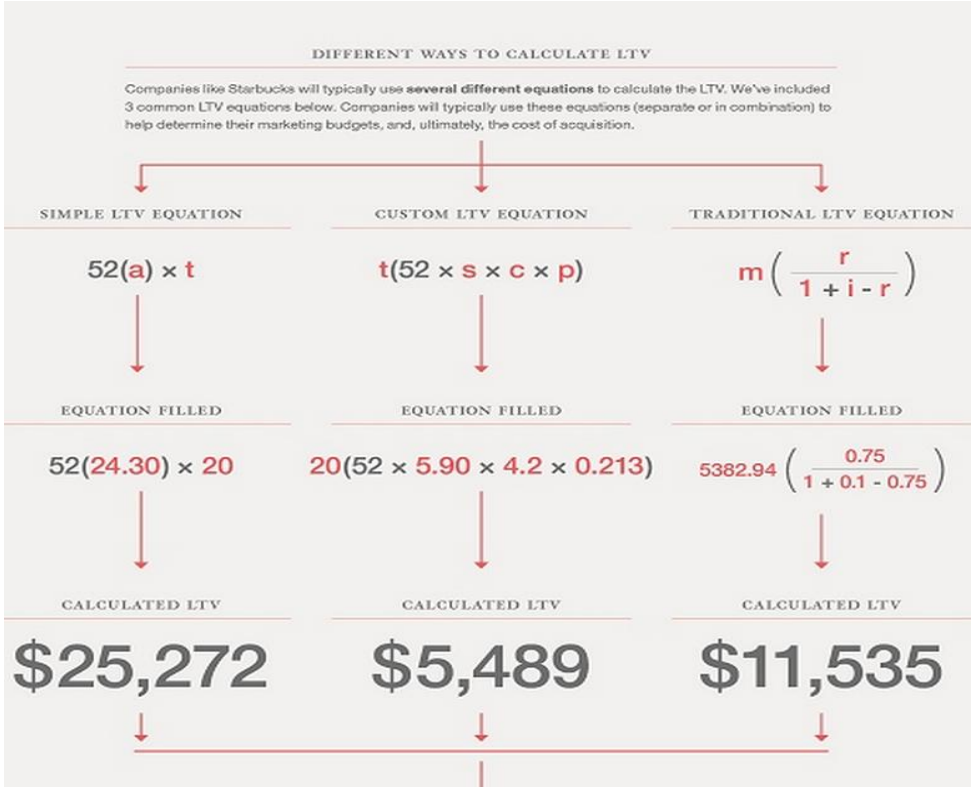
p **Profit Margin per Customer.** Starbucks: **21.3%**.

i **The Rate of Discount.** The "rate of discount" is the interest rate used in discounted cash flow analysis to determine the present value of future cash flows. Usually this number falls between 8% and 15%. Starbucks: **10%**.

m **Avg. Gross Margin per Customer Lifespan.** Starbucks has a profit margin of 21.3% (see constant "p"). If the average customer spends $25,272 (see the "Simple LTV Equation" results below) during their time as a customer ("t"), Starbucks has gross margin per customer lifespan of **$5382.94**.

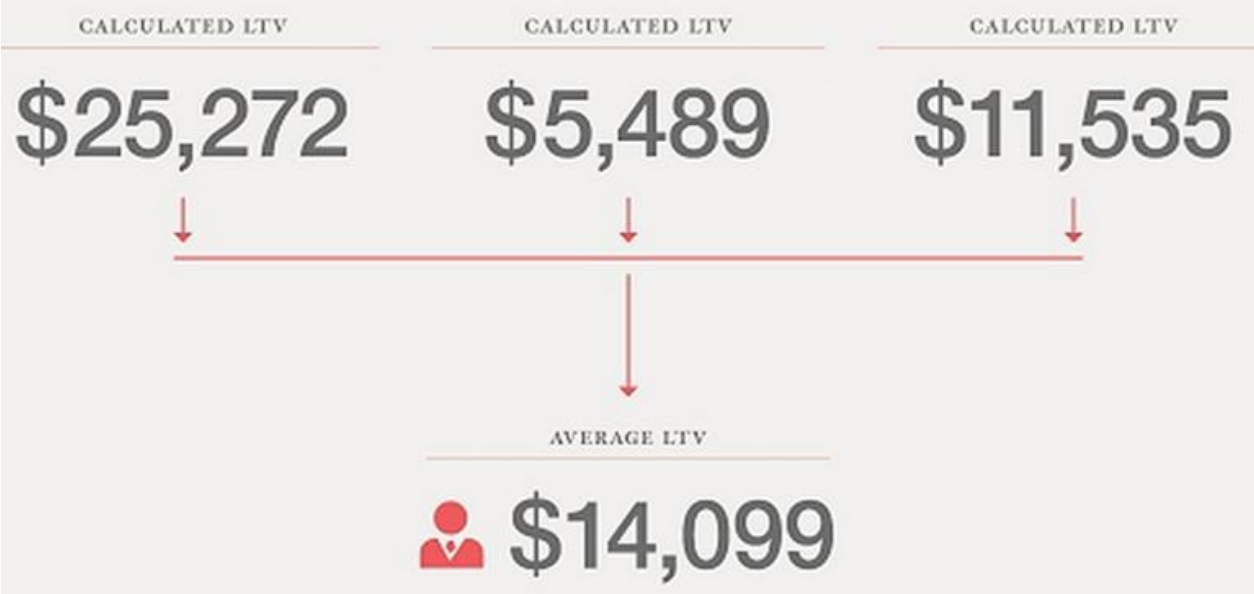
https://blog.kissmetrics.com/how-to-calculate-lifetime-value/

LTV Analytics



https://blog.kissmetrics.com/how-to-calculate-lifetime-value/

LTV Analytics



https://blog.kissmetrics.com/how-to-calculate-lifetime-value/

The **Pareto principle** (also known as the **80–20 rule**, the **law of the vital few,** and the **principle of factor sparsity**) states that, for many events, roughly 80% of the effects come from 20% of the causes

- 80% of a company's profits come from 20% of its customers
- 80% of a company's complaints come from 20% of its customers
- 80% of a company's profits come from 20% of the time its staff spend
- 80% of a company's sales come from 20% of its products
- 80% of a company's sales are made by 20% of its sales staff

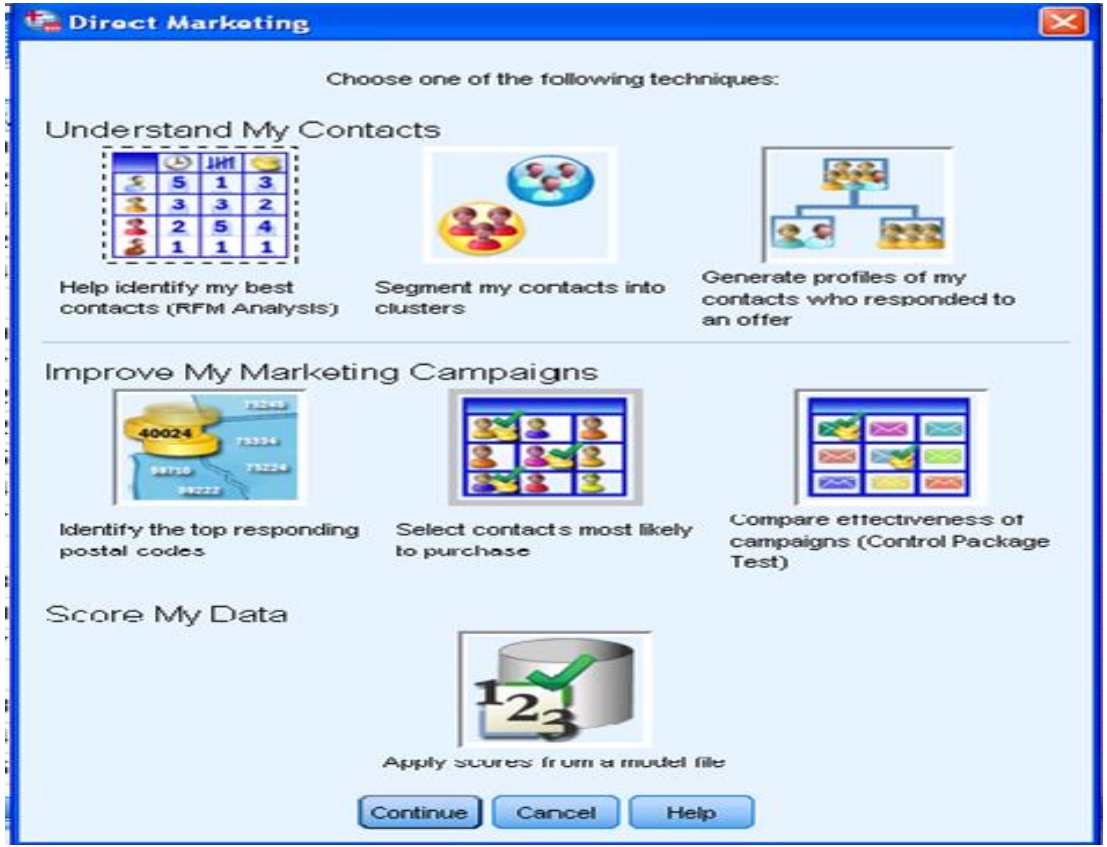
Several criminology studies have found 80% of crimes are committed by 20% of criminals.

RFM Analysis

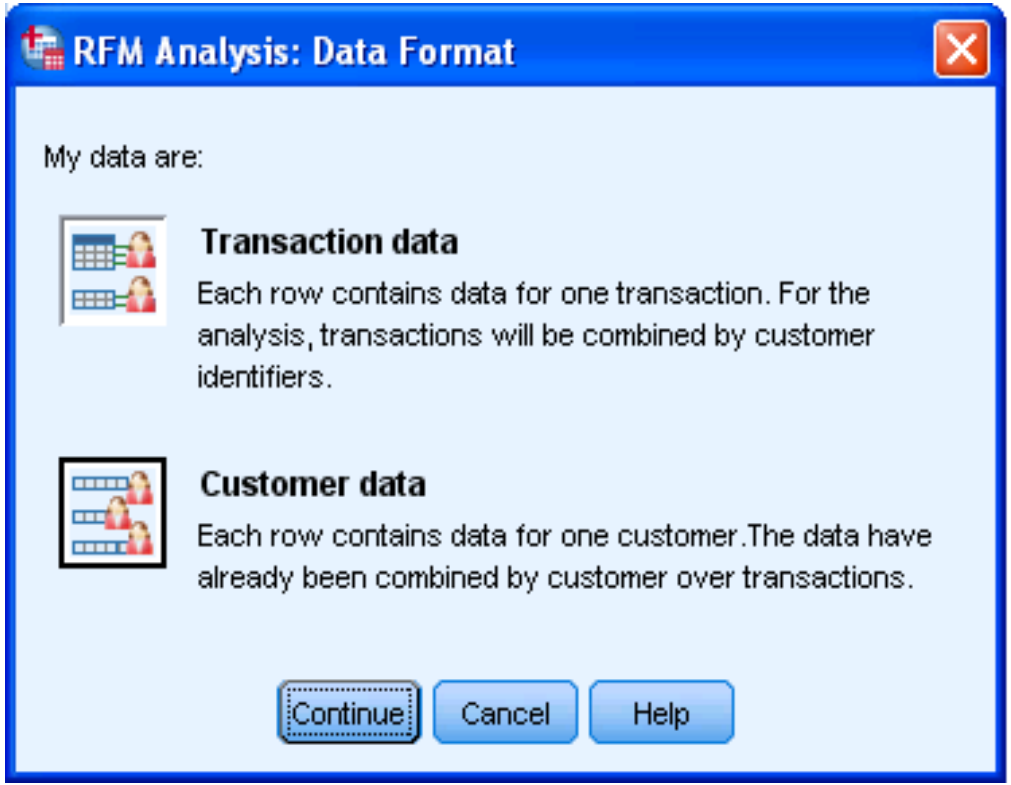
RFM is a method used for analyzing customer value.

- **R**ecency - *How recently did the customer purchase?*
- **F**requency - *How often do they purchase?*
- **M**onetary Value - *How much do they spend?*

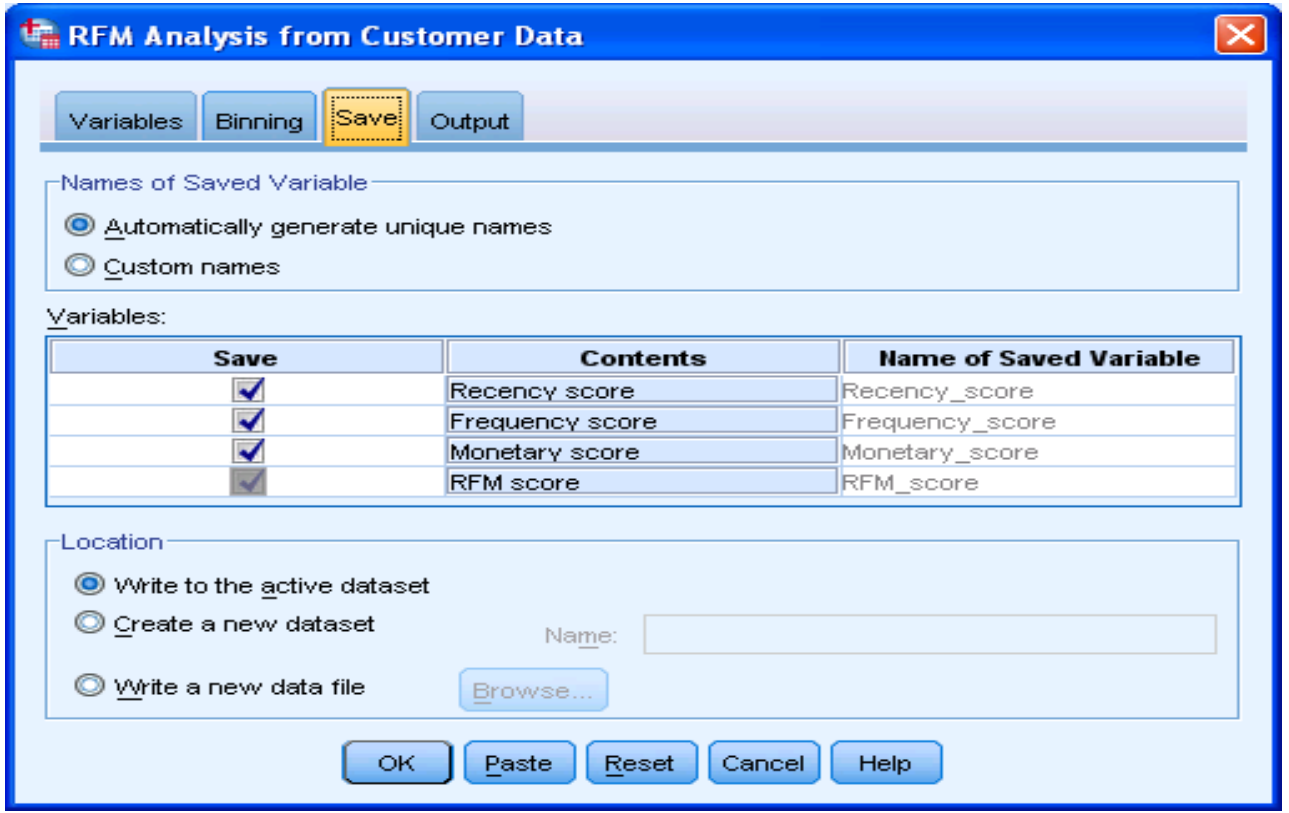
RFM Analysis



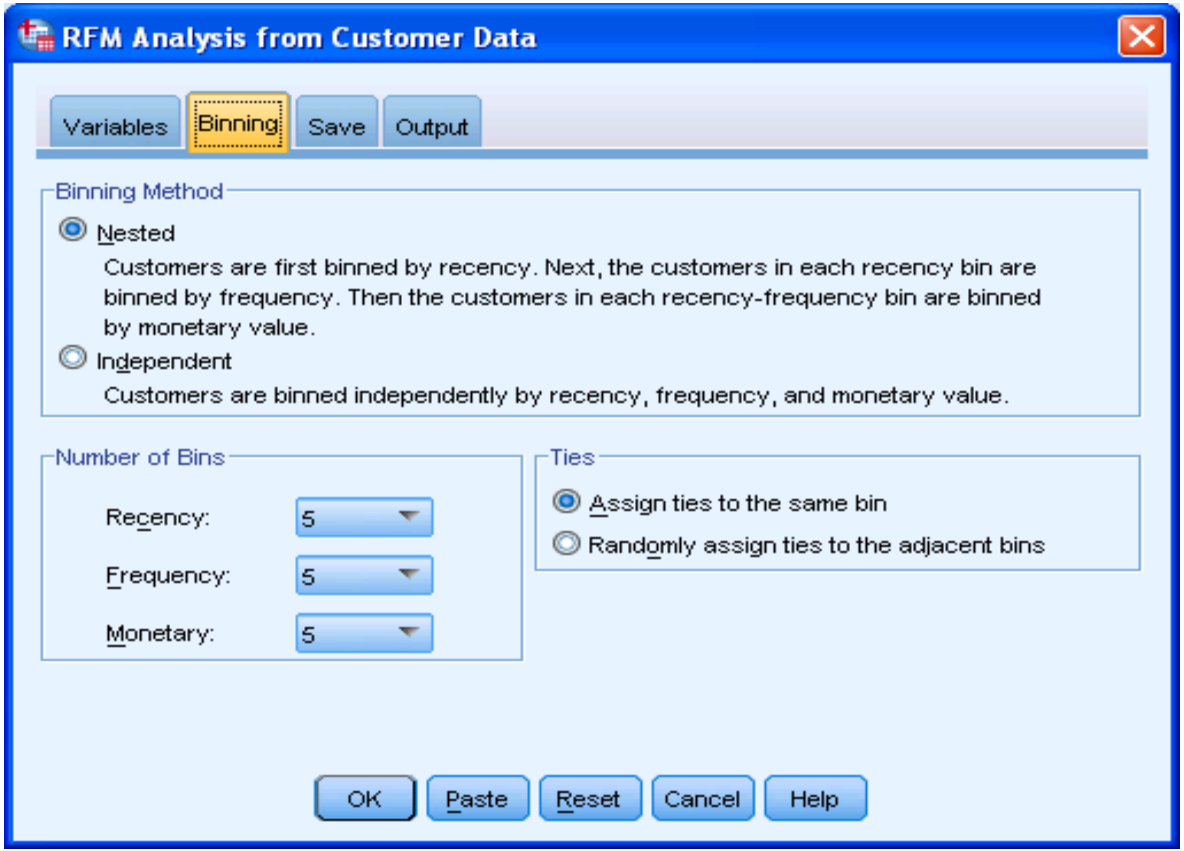
RFM Analysis



RFM Analysis

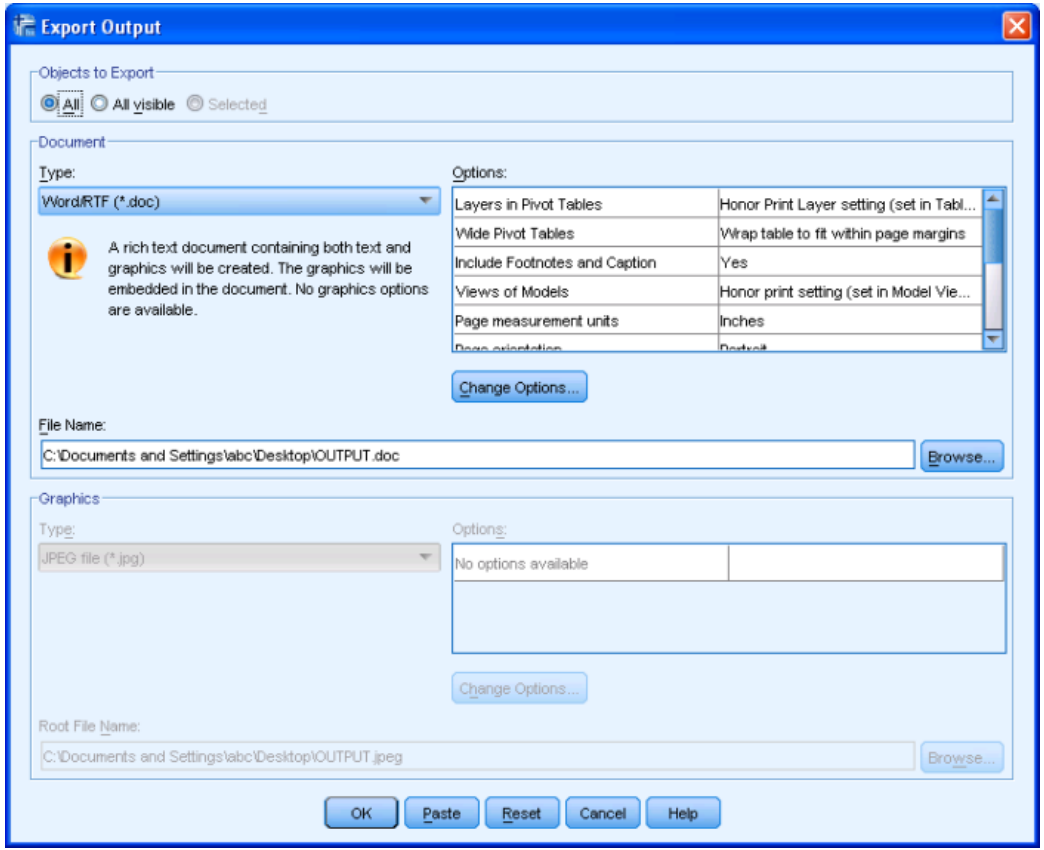


RFM Analysis

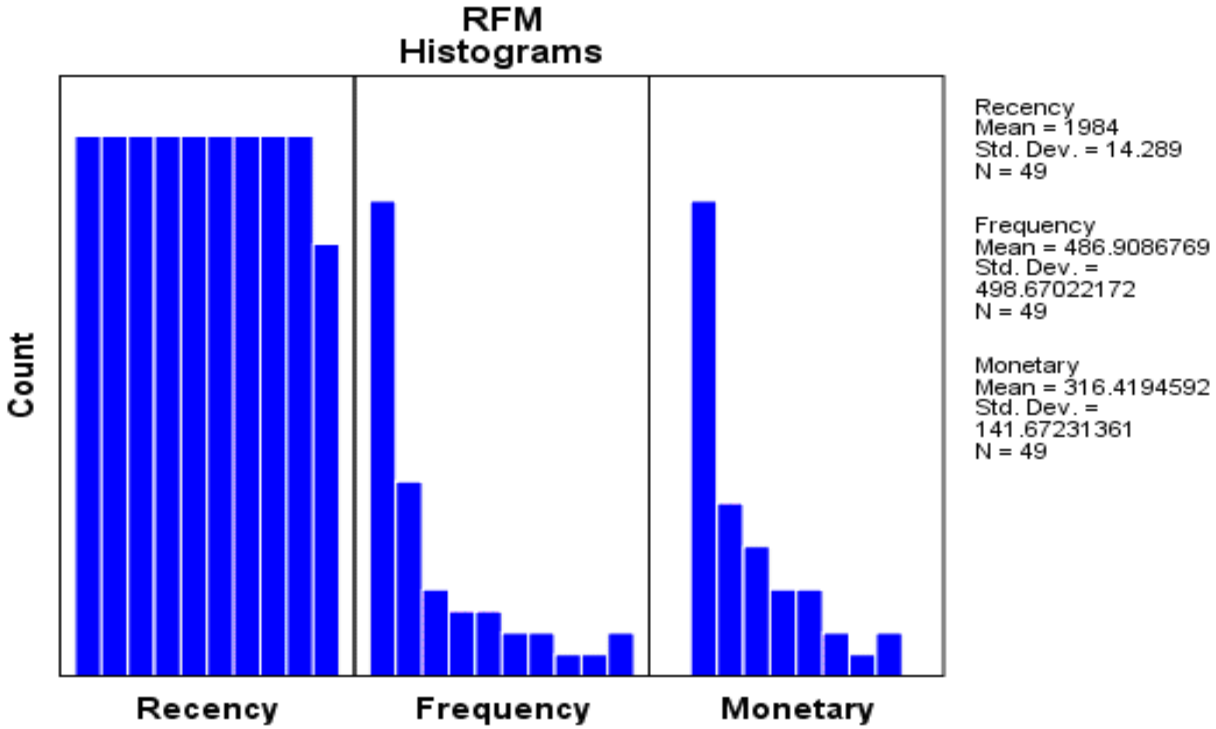


RFM Analysis

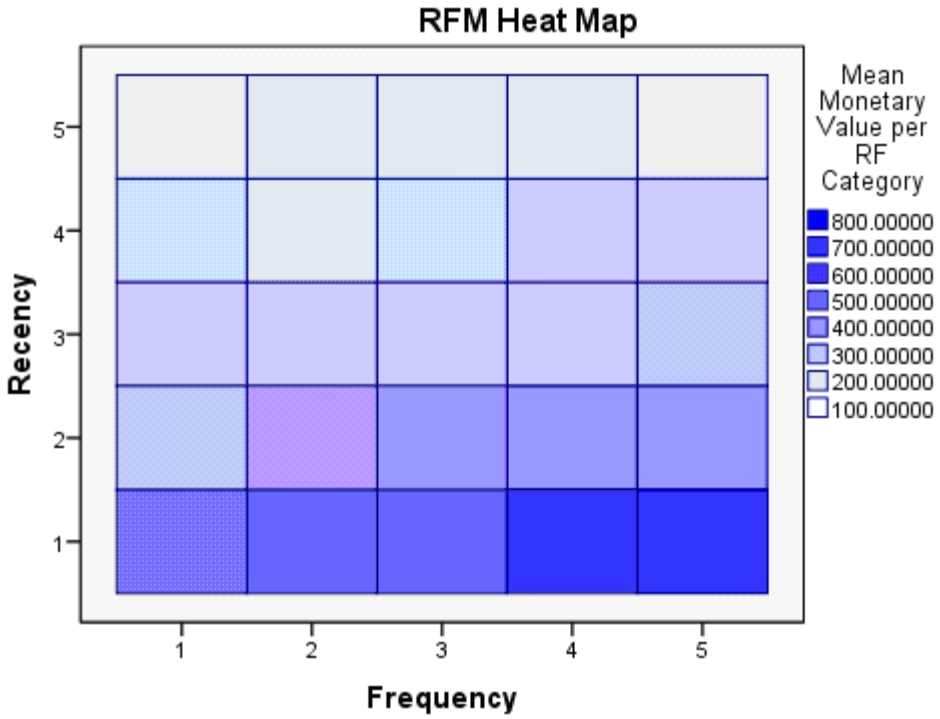
Using SPSS 19 - example



RFM Analysis

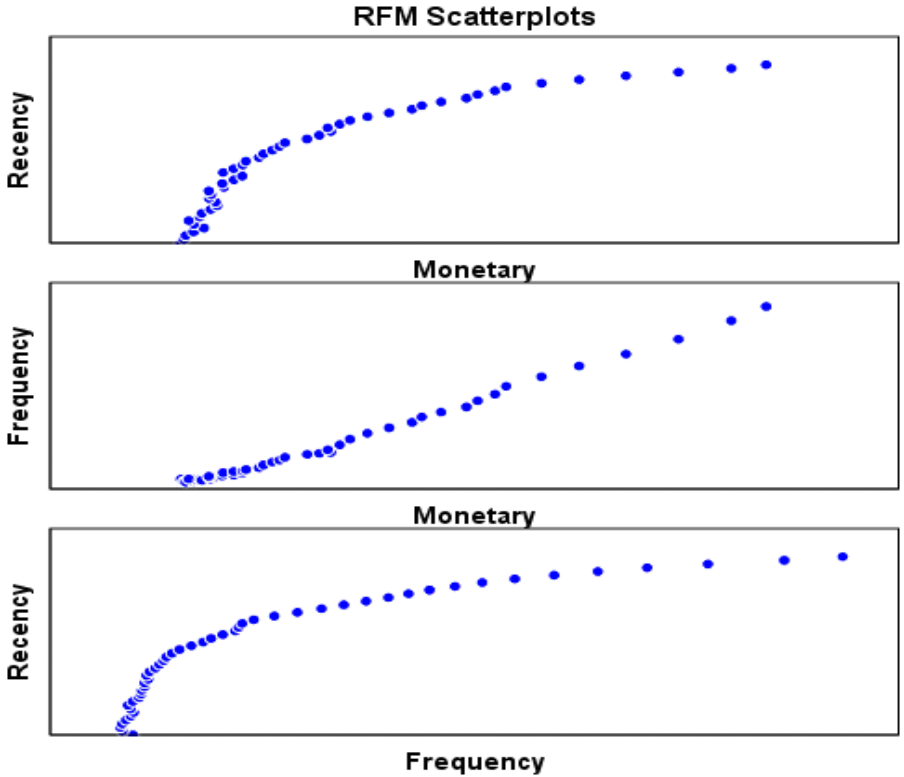


RFM Analysis



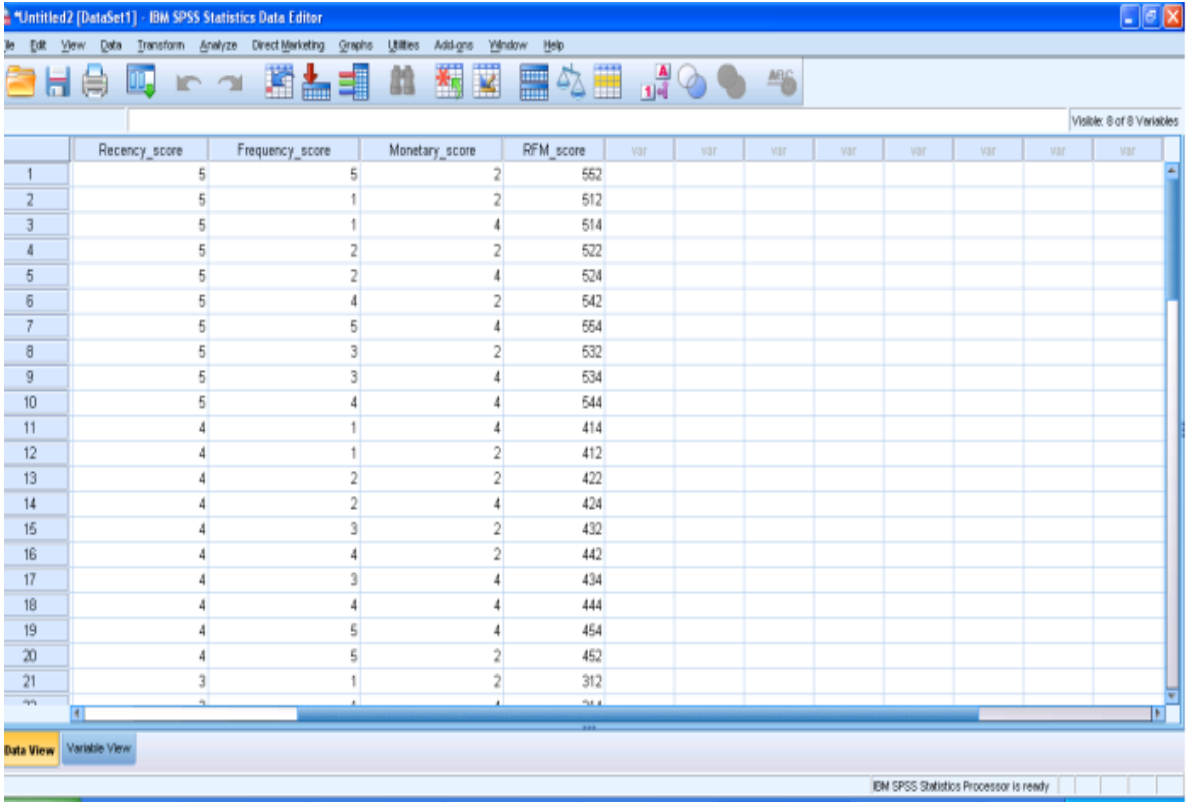
RFM Analysis

Using SPSS 19 - example



RFM Analysis

Using SPSS 19 - example



	Recency_score	Frequency_score	Monetary_score	RFM_score
1	5	5	2	552
2	5	1	2	512
3	5	1	4	514
4	5	2	2	522
5	5	2	4	524
6	5	4	2	542
7	5	5	4	554
8	5	3	2	532
9	5	3	4	534
10	5	4	4	544
11	4	1	4	414
12	4	1	2	412
13	4	2	2	422
14	4	2	4	424
15	4	3	2	432
16	4	4	2	442
17	4	3	4	434
18	4	4	4	444
19	4	5	4	454
20	4	5	2	452
21	3	1	2	312

RFM Analysis

RFM is a method used for analyzing customer value.

- **R**ecency - *How recently did the customer purchase?*
- **F**requency - *How often do they purchase?*
- **M**onetary Value - *How much do they spend?*

A method

- Recency = 10 - the number of months that have passed since the customer last purchased
- Frequency = number of purchases in the last 12 months (maximum of 10)
- Monetary = value of the highest order from a given customer (benchmarked against $10k)

Alternatively, one can create categories for each attribute. For instance, the Recency attribute might be broken into three categories: customers with purchases within the last 90 days; between 91 and 365 days; and longer than 365 days. Such categories may be arrived at by applying business rules, or using a data mining technique, to find meaningful **breaks**.

A commonly used shortcut is to use deciles. One is advised to look at distribution of data before choosing breaks.

Refresher in Statistics

Mean

Arithmetic Mean- the sum of the values divided by the number of values.
The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean) e.g. rates of growth.

Median

the **median** is the number separating the higher half of a data sample, a population, or a probability distribution, from the lower hal

Mode-

The "mode" is the value that occurs most often.

Refresher in Statistics

Range

the **range** of a set of data is the difference between the largest and smallest values.

Variance

mean of squares of differences of values from mean

Standard Deviation

square root of its variance

Frequency

a **frequency distribution** is a table that displays the **frequency** of various outcomes in a sample.

Distributions

Bernoulli

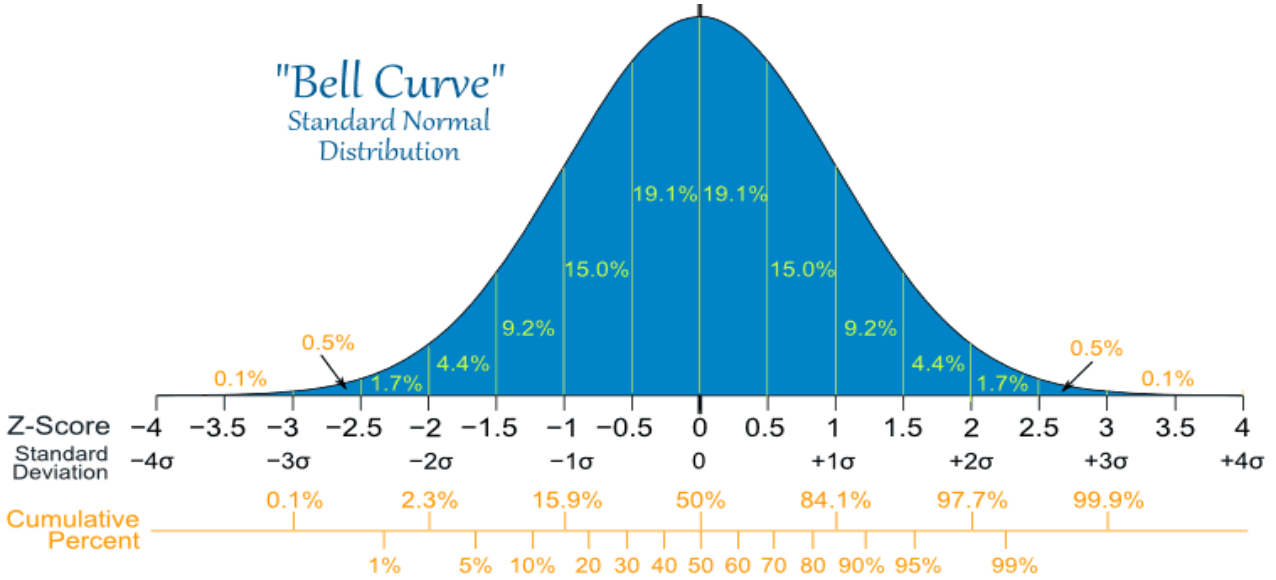
p $q = 1 - p$

Distribution of a random variable which takes value 1 with success probability and value 0 with failure probability. It can be used, for example, to represent the toss of a coin

Distributions

Normal

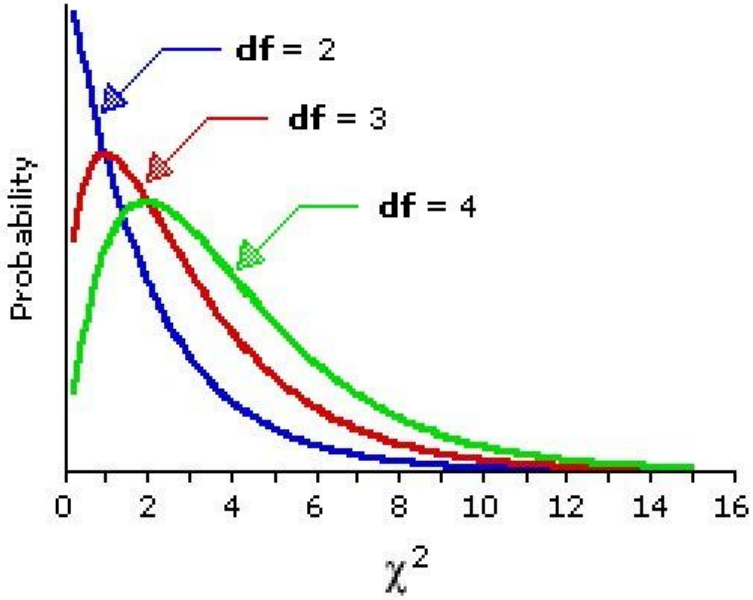
The simplest case of a normal distribution is known as the *standard normal distribution*. This is a special case where μ=0 and σ=1,



Distributions

Chi Square

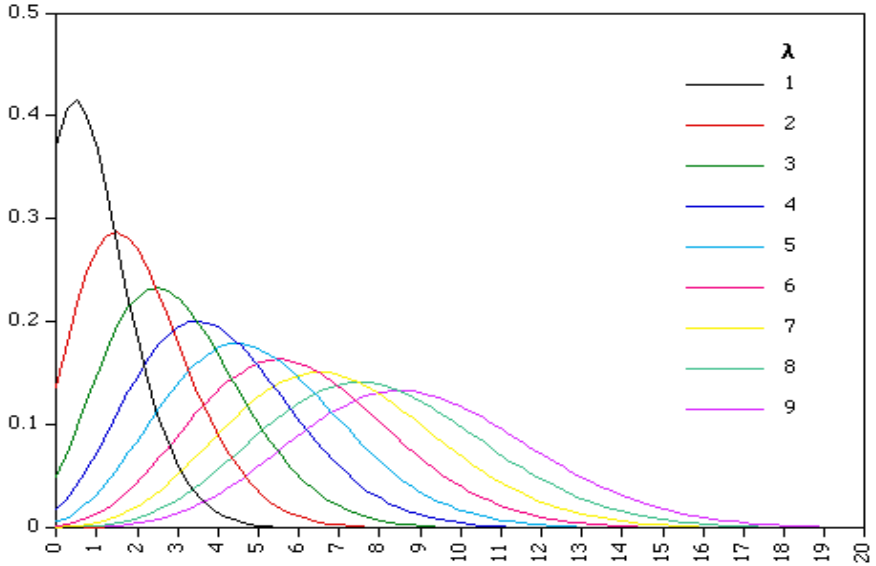
the distribution of a sum of the squares of *k* independent standard normal random variables.



Distributions

Poisson

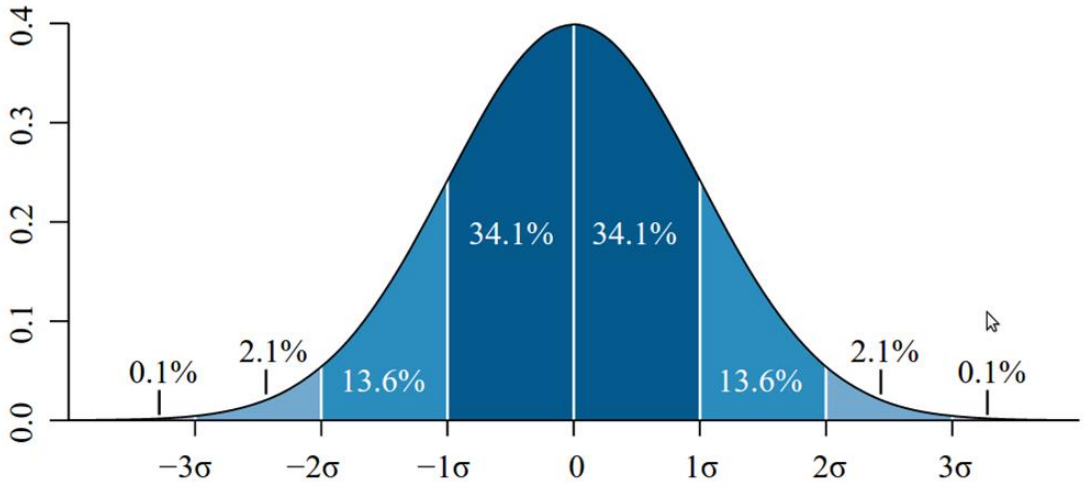
a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event



Refresher in Statistics

Probability Distribution

The probability density function (pdf) of the normal distribution, also called Gaussian or "bell curve", the most important continuous random distribution. As notated on the figure, the probabilities of intervals of values correspond to the area under the curve.



Refresher in Statistics

Central Limit Theorem –

In probability theory, the **central limit theorem** (**CLT**) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independentrandom variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

Introduction to Modeling

Hypothesis testing

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps.

1. Formulate the null hypothesis (commonly, that the observations are the result of pure chance) and the alternative hypothesis (commonly, that the observations show a real effect combined with a component of chance variation).

2. Identify a test statistic that can be used to assess the truth of the null hypothesis.

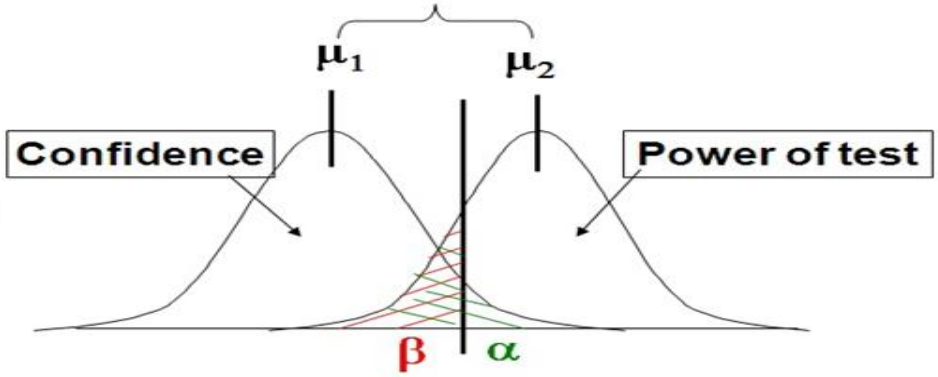
3. Compute the P-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the -value, the stronger the evidence against the null hypothesis.

4. Compare the -value to an acceptable significance value (sometimes called an alpha value). If , that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

http://mathworld.wolfram.com/HypothesisTesting.html

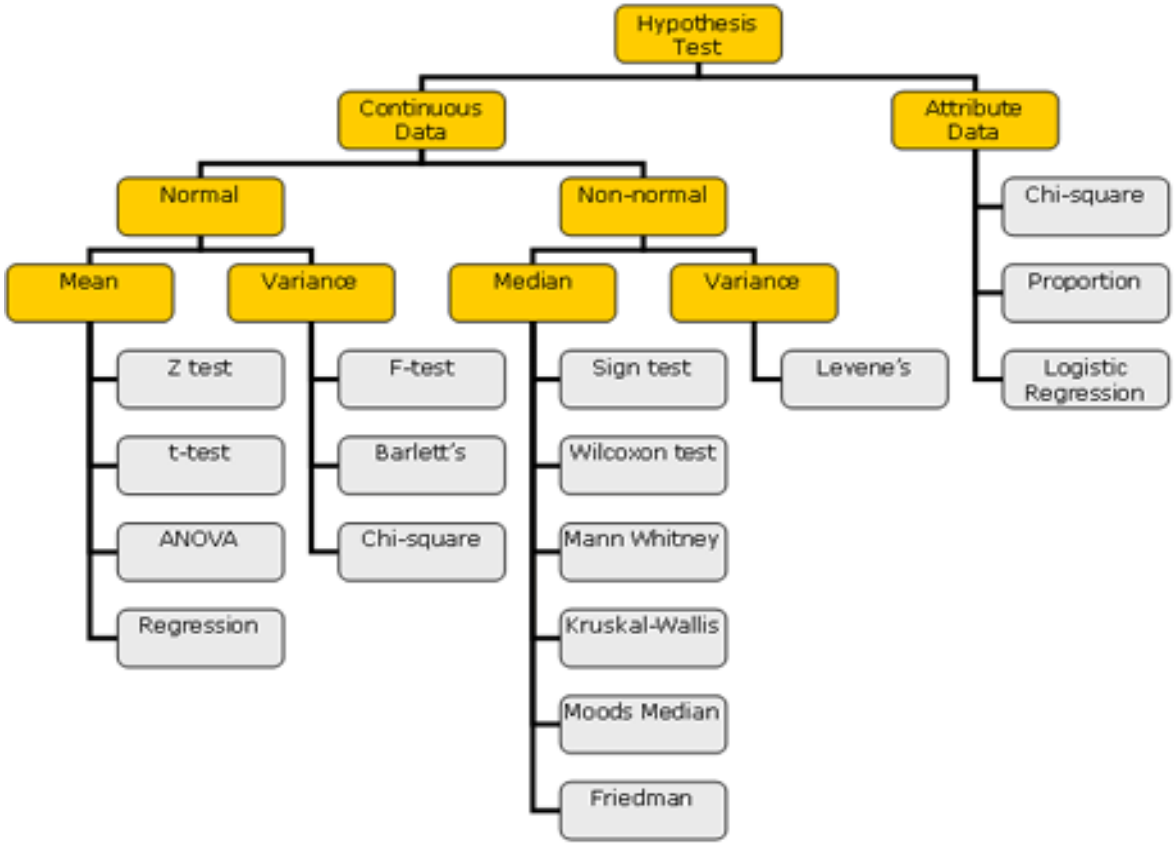
Hypothesis testing

Decision \ Reality	Ho	Ha
Ho	Correct decision	Type II error **β- risk** Producer's risk
Ha	Type I error **α- risk** Consumer' risk	Correct decision



	The Truth (unknown to the researcher)	
The Researcher's Decision	The Null Hypothesis is True	The Null Hypothesis is False
Reject the Null Hypothesis	Type I Error	Correct Decision
Fail to Reject the Null Hypothesis	Correct Decision	Type II Error

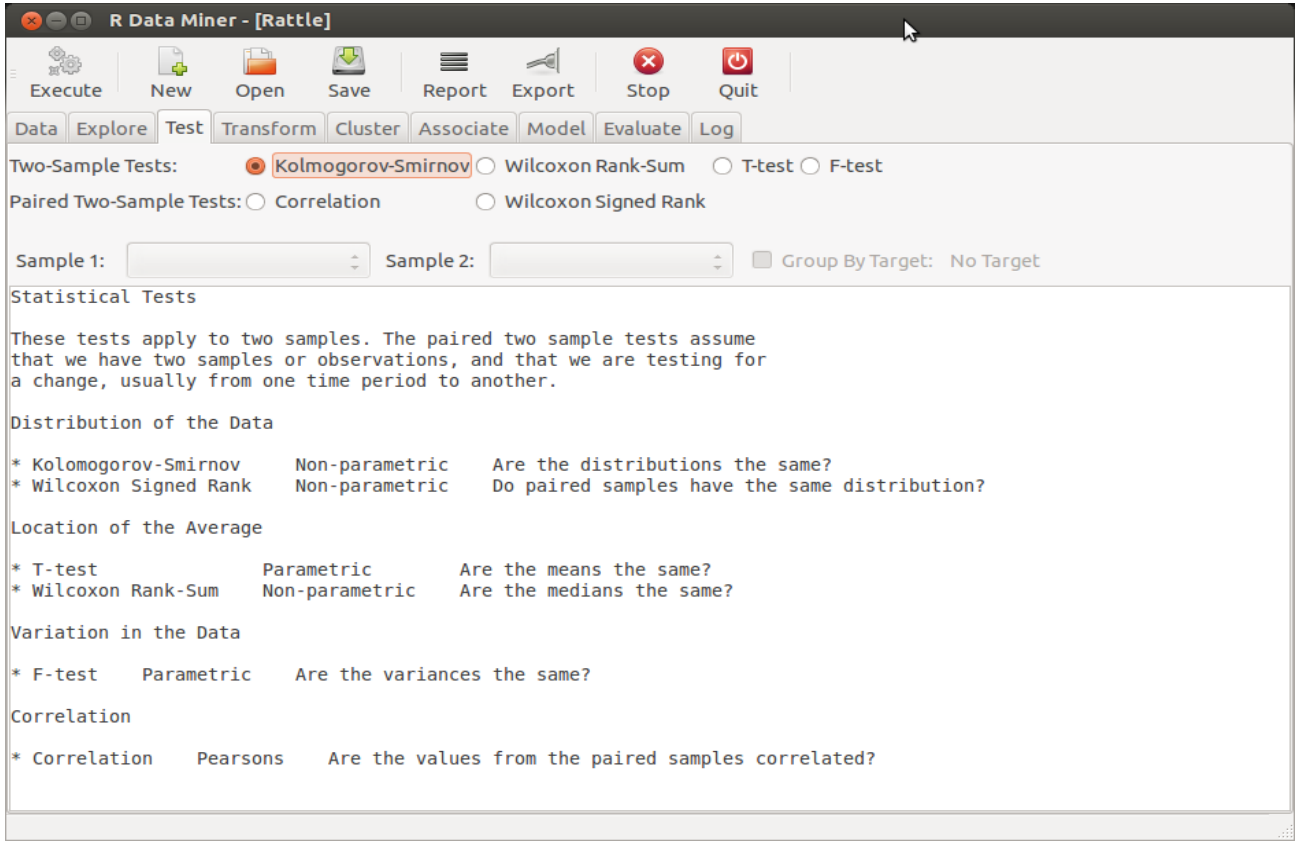
Hypothesis testing



Hypothesis testing

Comparison of MEANS	Degrees of Freedom	Application	Assumptions	Test Statistic
One Sample Z-Test	Not Applicable	Testing the difference of a sample mean, x-bar, with a known population mean, μ (fixed mean, historical mean, or targeted mean)	Normal distribution Known population σ.	$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
One Sample t-test	n-1	Testing the difference of one sample mean, x-bar, with a known population mean, μ (fixed mean, historical mean, or targeted mean)	Normal distribution Population standard deviation, σ, is unknown.	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
Two Sample t-test	$n_1 + n_2 - 2$	Testing difference of two sample means when population variances unknown but considered equal	Normal Distribution Requires standard pooled deviation calculation, s_p	$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Paired t-test	n - 1	Testing two sample means when their respective population standard deviations are unknown but considered equal. Data recorded in pairs and each pair has a difference, d.	Normal Distribution Two dependent samples Always two-tailed test S_d = standard deviation of the differences of all samples	$t = \frac{\bar{d}\sqrt{n}}{s_d}$
One-Way ANOVA	$n_1 - 1$ & $n_2 - 1$	Testing the difference of three or more population means	Normal Distribution s_1^2 and s_2^2 represent sample variances	$F = \frac{(s_1)^2}{(s_2)^2}$

Hypothesis testing



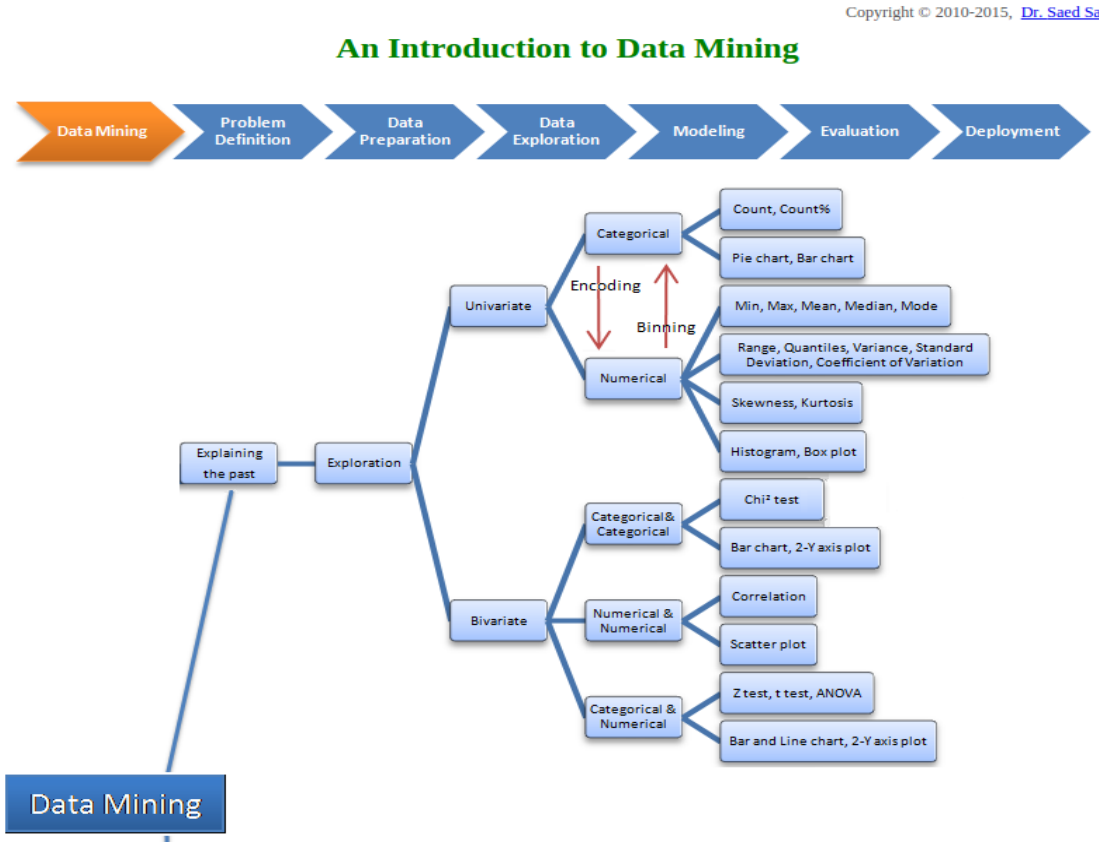
Data Mining

Data Mining is an analytic process designed to explore **data** (usually large amounts of **data** - typically business or market related - also known as "big **data**") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns

Data Mining Map

Source-

http://www.saedsayad.com/



Data Mining Map

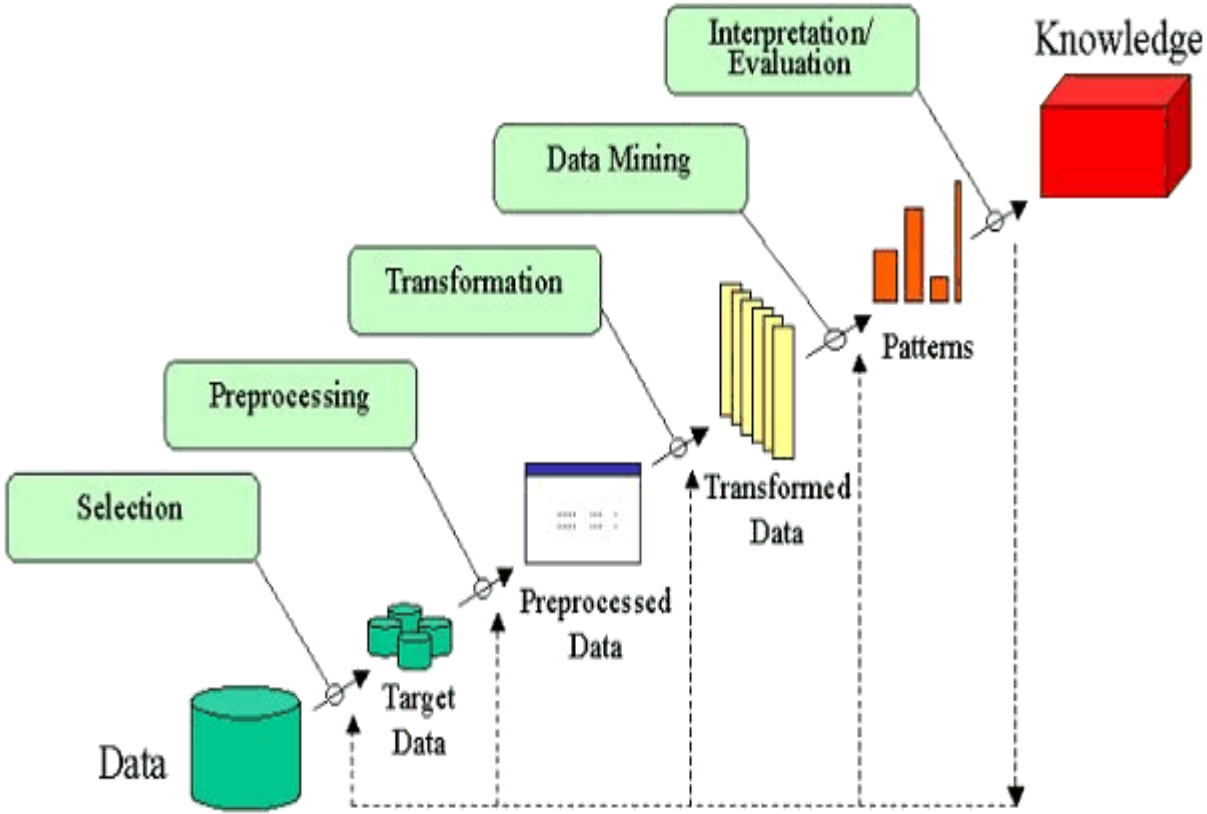
Source-

http://www.saedsayad.com/

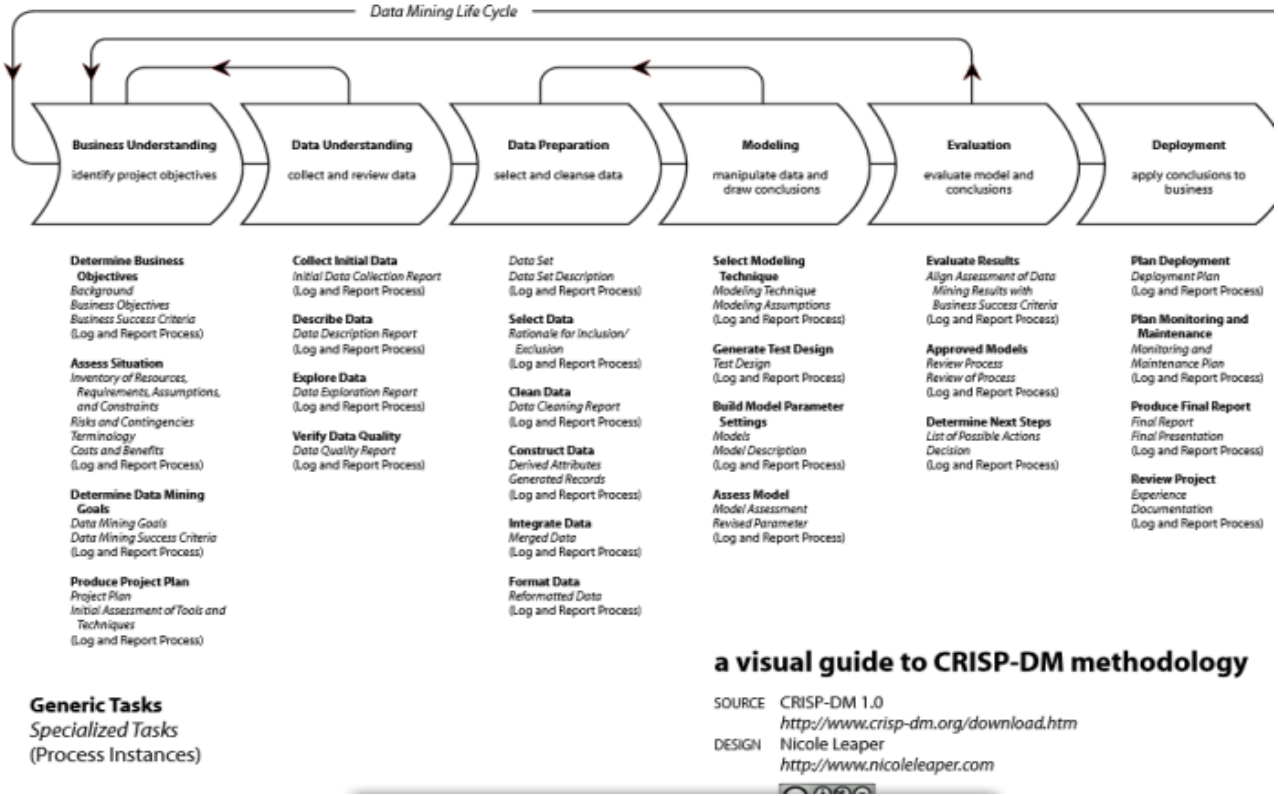
Examples of Data Mining

- which items sell well together in retail (market basket)
- which products sell well together on a website (association analysis)
- which customers are likely to buy a new credit card (regression)

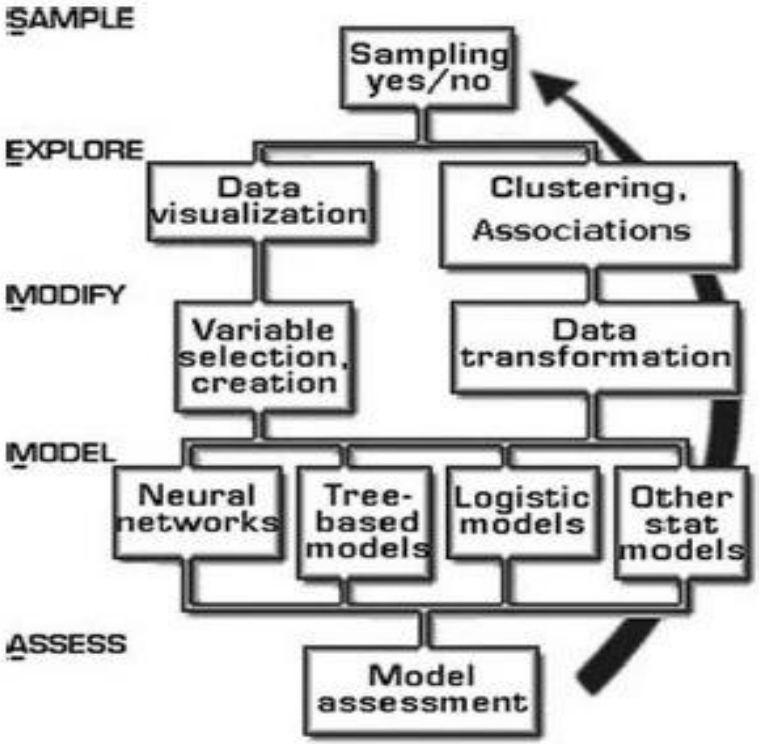
KDD


Interpretation/
Evaluation
Knowledge
Data Mining
Transformation
Patterns
Preprocessing
Selection
Transformed
Data
Preprocessed
Data
Target
Data
Data

CRISP DM


Data Mining Life Cycle
Business Understanding
identify project objectives
Data Understanding
collect and review data
Data Preparation
select and cleanse data
Modeling
manipulate data and draw conclusions
Evaluation
evaluate model and conclusions
Deployment
apply conclusions to business
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)
Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)
Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)
Collect Initial Data
Initial Data Collection Report
(Log and Report Process)
Describe Data
Data Description Report
(Log and Report Process)
Explore Data
Data Exploration Report
(Log and Report Process)
Verify Data Quality
Data Quality Report
(Log and Report Process)
Data Set
Data Set Description
(Log and Report Process)
Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)
Clean Data
Data Cleaning Report
(Log and Report Process)
Construct Data
Derived Attributes
Generated Records
(Log and Report Process)
Integrate Data
Merged Data
(Log and Report Process)
Format Data
Reformatted Data
(Log and Report Process)
Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)
Generate Test Design
Test Design
(Log and Report Process)
Build Model Parameter Settings
Models
Model Description
(Log and Report Process)
Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)
Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)
Approved Models
Review Process
Review of Process
(Log and Report Process)
Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)
Plan Deployment
Deployment Plan
(Log and Report Process)
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)
Produce Final Report
Final Report
Final Presentation
(Log and Report Process)
Review Project
Experience Documentation
(Log and Report Process)
Generic Tasks
Specialized Tasks
(Process Instances)
a visual guide to CRISP-DM methodology
SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

SEMMA


SAMPLE
Sampling yes/no
EXPLORE
Data visualization
Clustering, Associations
MODIFY
Variable selection, creation
Data transformation
MODEL
Neural networks
Tree-based models
Logistic models
Other stat models
ASSESS
Model assessment

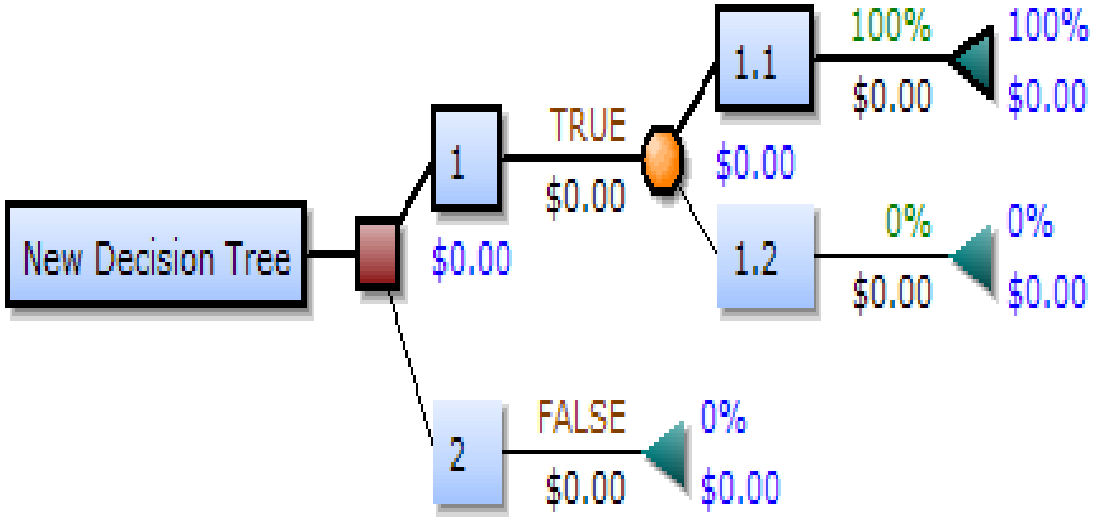
Machine Learning

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational **learning** theory in artificial intelligence.

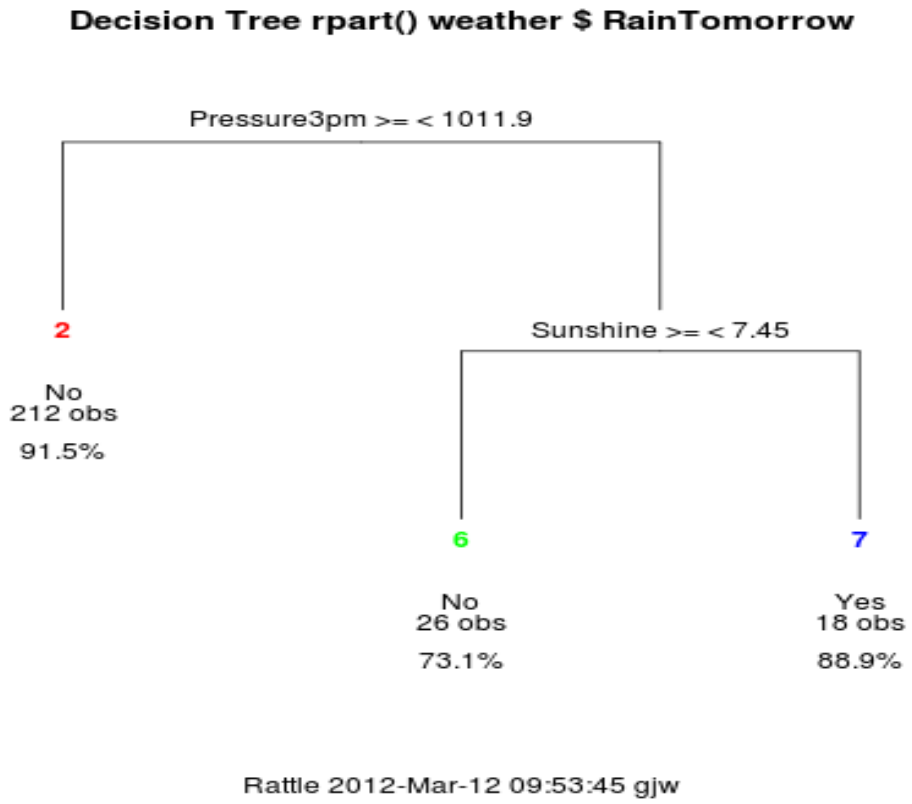
Machine learning explores the construction and study of algorithms that can learn from and make predictions on data.

- Supervised learning. The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- Unsupervised learning, no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
- In classification, a supervised way, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) Spam filtering, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
- In regression, also a supervised problem, the outputs are continuous rather than discrete.
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

Decision Trees



Decision Trees



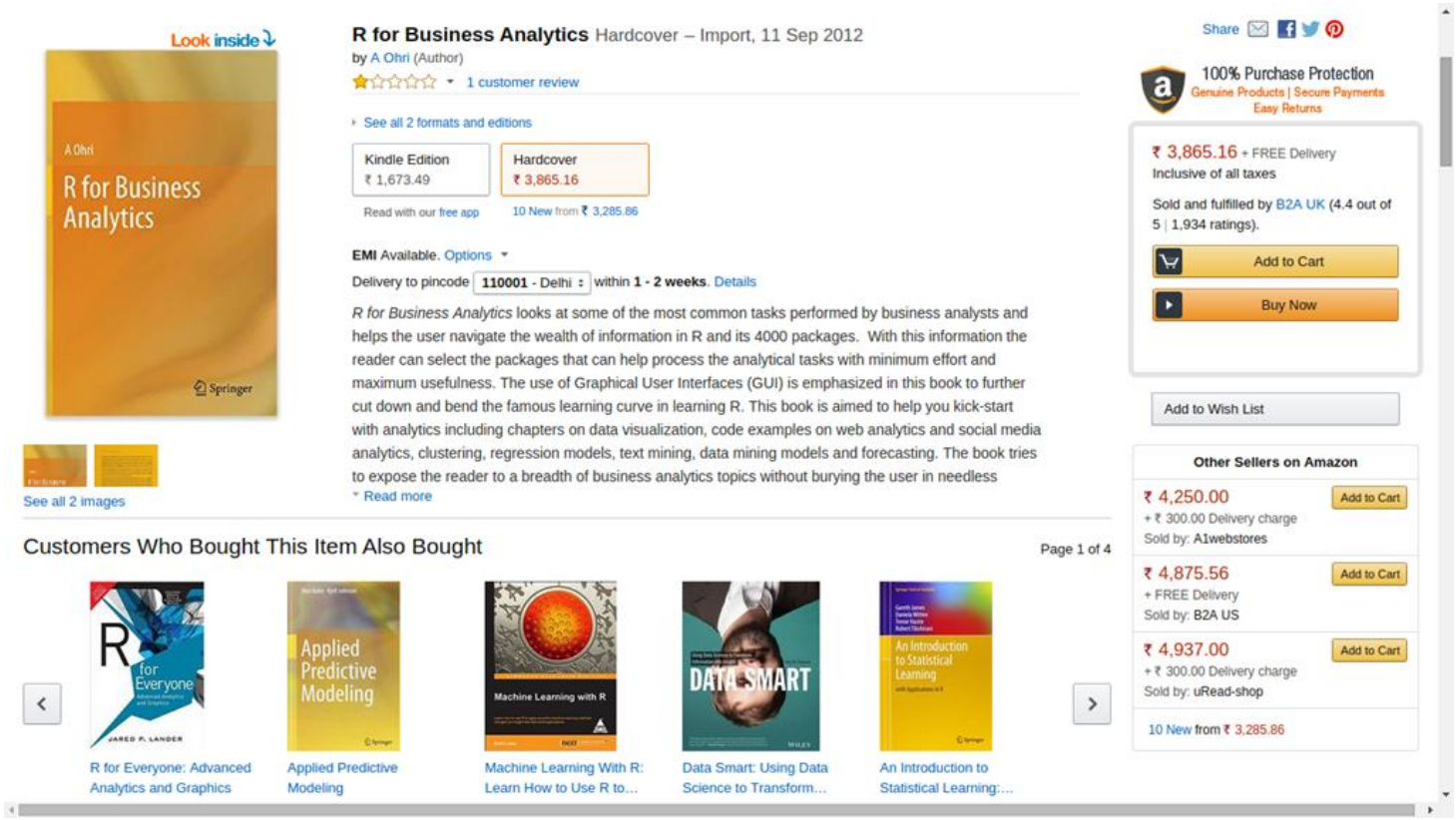
Association Analysis

As an unsupervised learning technique it has delivered considerable benefit in areas ranging from the traditional shopping basket analysis to the analysis of who bought what other books or who watched what other videos, and in areas including health care, telecommunications, and so on

from

http://handsondatascience.com/ARulesO.pdf

An Example of Data Mining



An Example of Data Mining

Examples

https://en.wikipedia.org/wiki/Apriori_algorithm

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

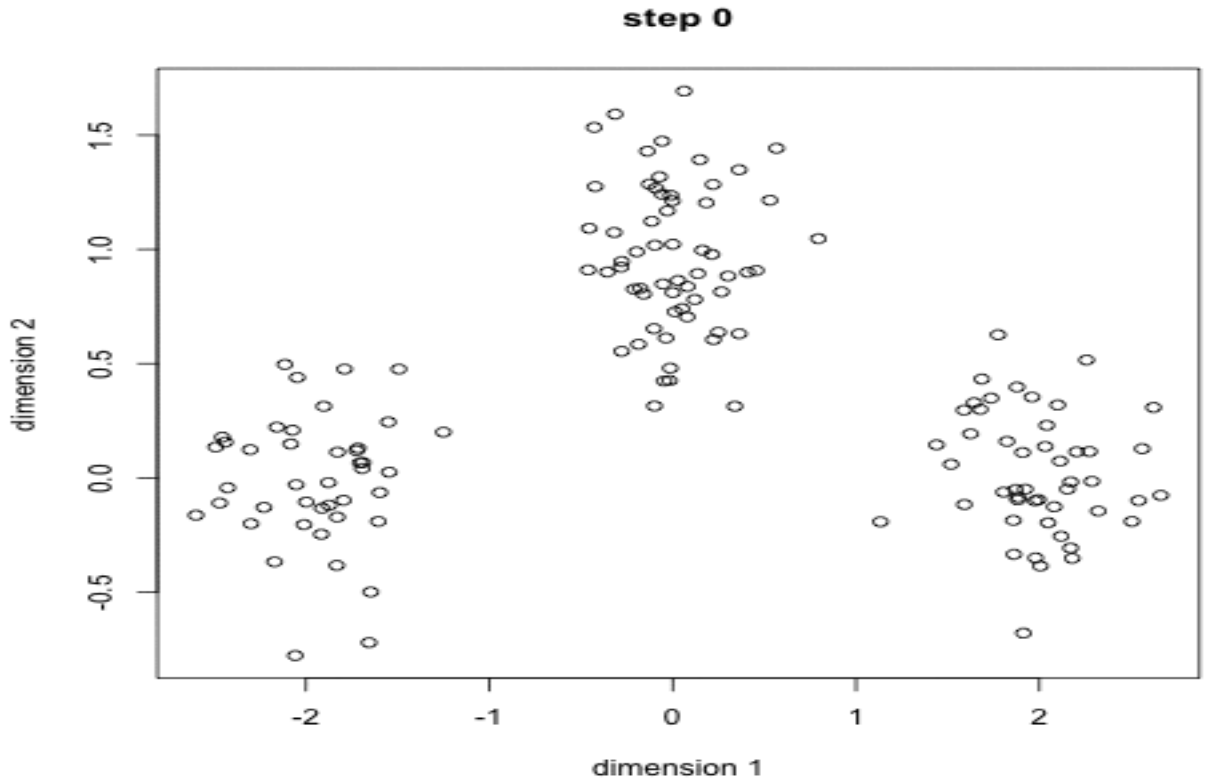
Clustering

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (**clusters**).

k-means clustering aims to partition n observations into **k clusters** in which each observation belongs to the **cluster** with the nearest **mean**, serving as a prototype of the**cluster**. This results in a partitioning of the data space into Voronoi cells

http://shabal.in/visuals/kmeans/1.html

Clustering

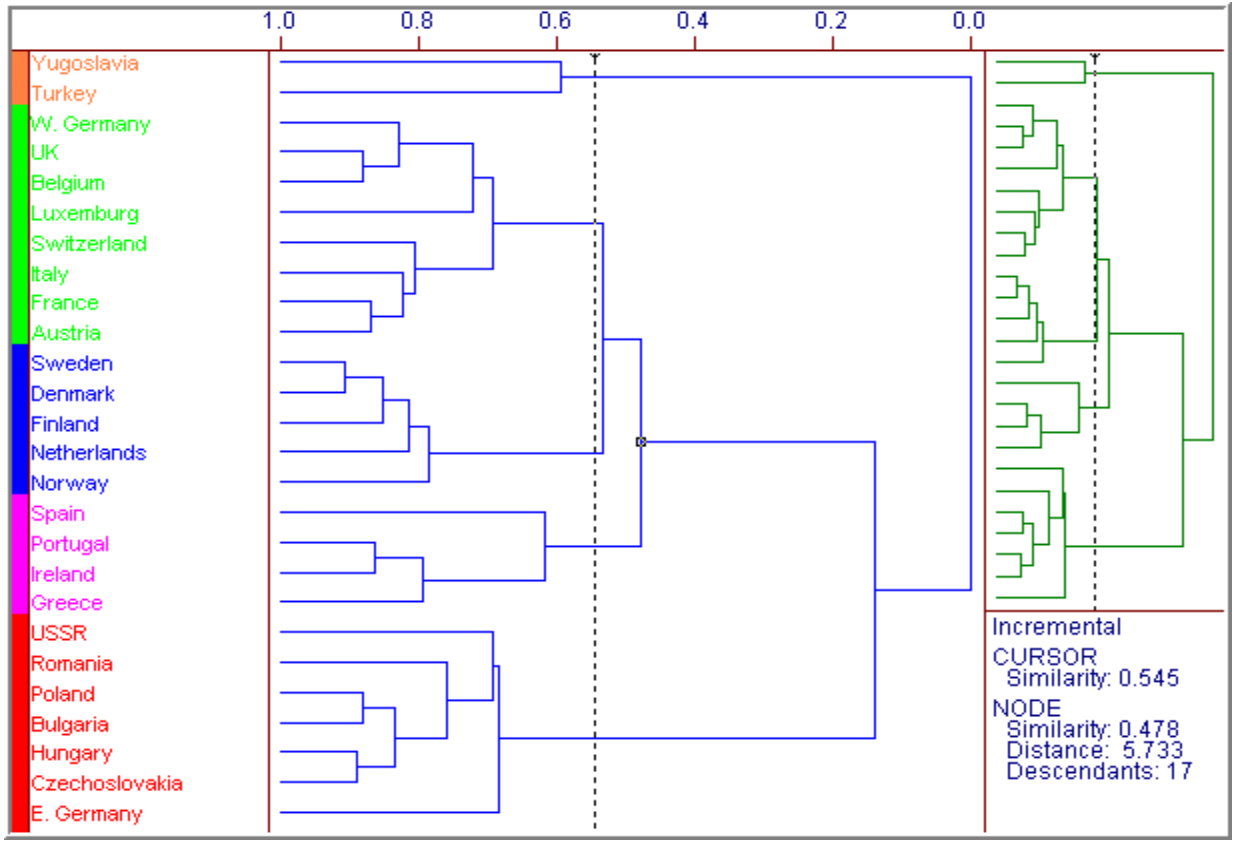


Clustering

hierarchical clustering (also called **hierarchical cluster analysis** or **HCA**) is a method of cluster analysis which seeks to build ahierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: [1]

- **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Clustering


1.0
0.8
0.6
0.4
0.2
0.0
Yugoslavia
Turkey
W. Germany
UK
Belgium
Luxemburg
Switzerland
Italy
France
Austria
Sweden
Denmark
Finland
Netherlands
Norway
Spain
Portugal
Ireland
Greece
USSR
Romania
Poland
Bulgaria
Hungary
Czechoslovakia
E. Germany
Incremental
CURSOR
Similarity: 0.545
NODE
Similarity: 0.478
Distance: 5.733
Descendants: 17

Regression

regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

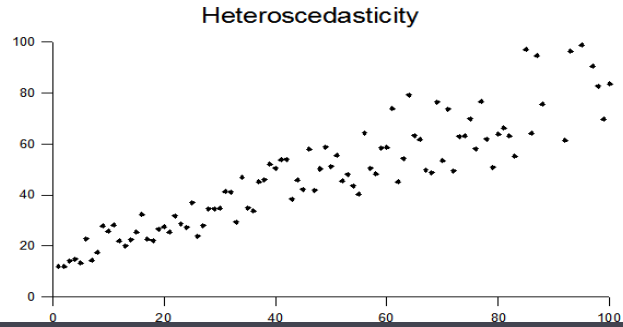
$$y = a + bx$$

$$y = a + bx + cy$$

$$\ln (p /1-p) = a + bx$$

Regression

1. In statistics, **multicollinearity** (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a non-trivial degree of accuracy. A multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor
2. **heteroscedasticity**(also spelled heteroskedasticity) refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.
3. The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< **0.05**) indicates that you can reject the null hypothesis.



Text Mining

1. Retrieving Text
2. Transforming Text to corpus
3. Cleaning Text (lowercase, punctuation, numbers, commonly used words (stop words))
4. Stemming Words
5. Building a Document-Term Matrix
6. Frequent Terms and Associations
7. Word Cloud

http://www.rdatamining.com/examples/text-mining

Sentiment Analysis

Sentiment analysis (also known as **opinion mining**) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document

example- http://www.slideshare.net/ajayohri/twitter-analysis-by-kaify-rais

Sentiment Analysis

A sentiment analysis model is used to analyze a text string and classify it with one of the labels that you provide; for example, you could analyze a tweet to determine whether it is positive or negative, or analyze an email to determine whether it is happy, frustrated, or sad.

R package "sentiment"

Another interesting option that we can use to do our sentiment analysis is by utilizing the R package **sentiment**by Timothy Jurka. This package contains two handy functions serving our purposes:

classify_emotion

This function helps us to analyze some text and classify it in different types of emotion: anger, disgust, fear, joy, sadness, and surprise.

classify_polarity

In contrast to the classification of emotions, the classify_polarity function allows us to classify some text as positive or negative.

example- https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

Social Network Analysis

Social network analysis (SNA) is a strategy for investigating **social** structures through the use of**network** and graph theories. It characterizes networked structures in terms of nodes (individual actors, people, or things within the **network**) and the ties or edges (relationships or interactions) that connect them.

The NSA has been performing social network analysis on Call Detail Records (CDRs), also known as metadata, since shortly after the September 11 Attacks

Social Network Analysis to Optimize Tax Enforcement Effort -The South African Revenue Service

http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1579&context=amcis2012

Irish Tax & Customs Authority

http://www.sas.com/en_ie/customers/irish-tax-and-customers.html

Social Network Analysis

Bridge: An individual whose weak ties fill a structural hole, providing the only link between two individuals or clusters. It also includes the shortest route when a longer one is unfeasible due to a high risk of message distortion or delivery failure.[18]

Centrality: Centrality refers to a group of metrics that aim to quantify the "importance" or "influence" (in a variety of senses) of a particular node (or group) within a network.

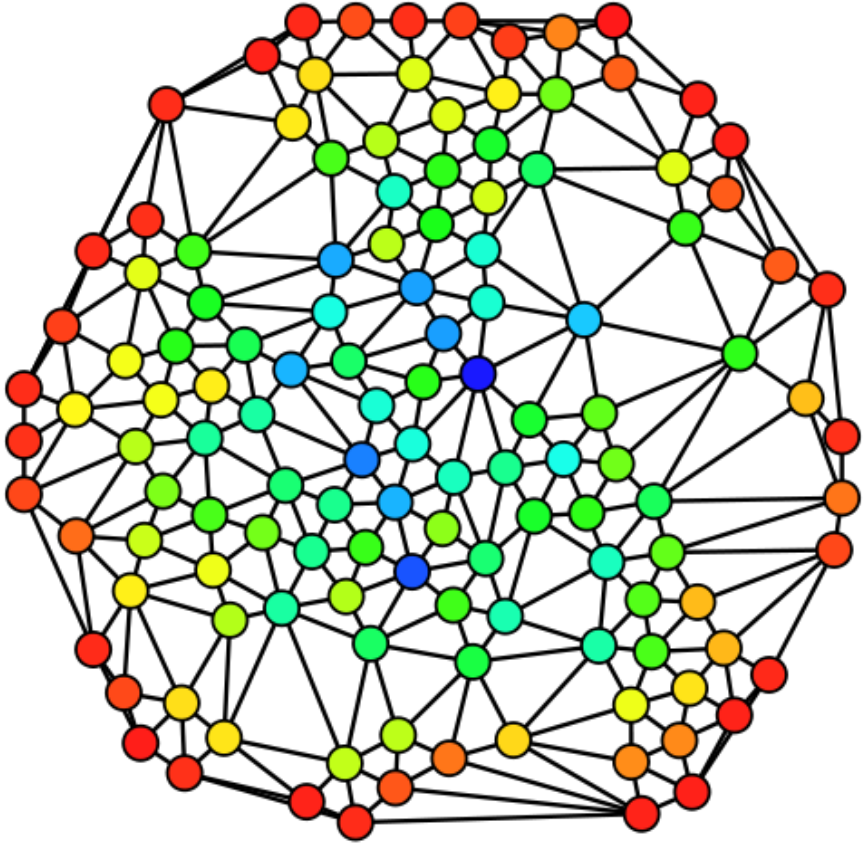
Density: The proportion of direct ties in a network relative to the total number possible.[25][26]

Distance: The minimum number of ties required to connect two particular actors, as popularized by Stanley Milgram's small world experiment and the idea of 'six degrees of separation'.

Mutuality/Reciprocity: The extent to which two actors reciprocate each other's friendship or other interaction.[16]

Network Closure: A measure of the completeness of relational triads.

Social Network Analysis



Hue (from red=0 to blue=max) indicates each node's betweenness centrality.

Social Network Analysis

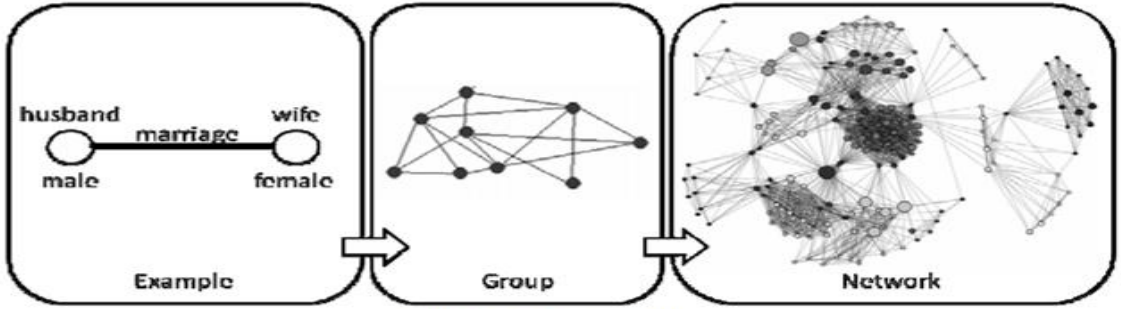


Figure 1. Social Network Analysis

Managing Tax Compliance through Decision Support Systems

Much like any other organization, tax compliance can be managed at a strategic, operational and tactical level, as presented in Figure 2 (OECD, 2008:10). Strategic compliance management considers the tax system in its entirety whereas operational compliance management focuses on whole taxpayer segments. Tactical compliance management considers targeted individuals, or groups of which social structures such as marriage, employee and employer relationships and tax consultant and taxpayer relationships are but a few examples. The different DSS defined by Power (2002:13-16) can be associated with the types of compliance-management levels. Knowledge driven DSS are associated with strategic management, data driven DSS with operational management, and model driven DSS with tactical management. Model driven DSS is often used to conduct SNA, and DSS tools such as Analyst Notebook and SAS are widely recognized as industry leaders in this domain.

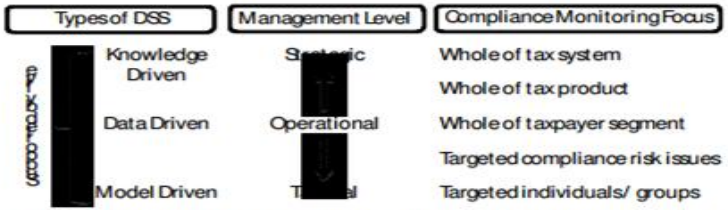
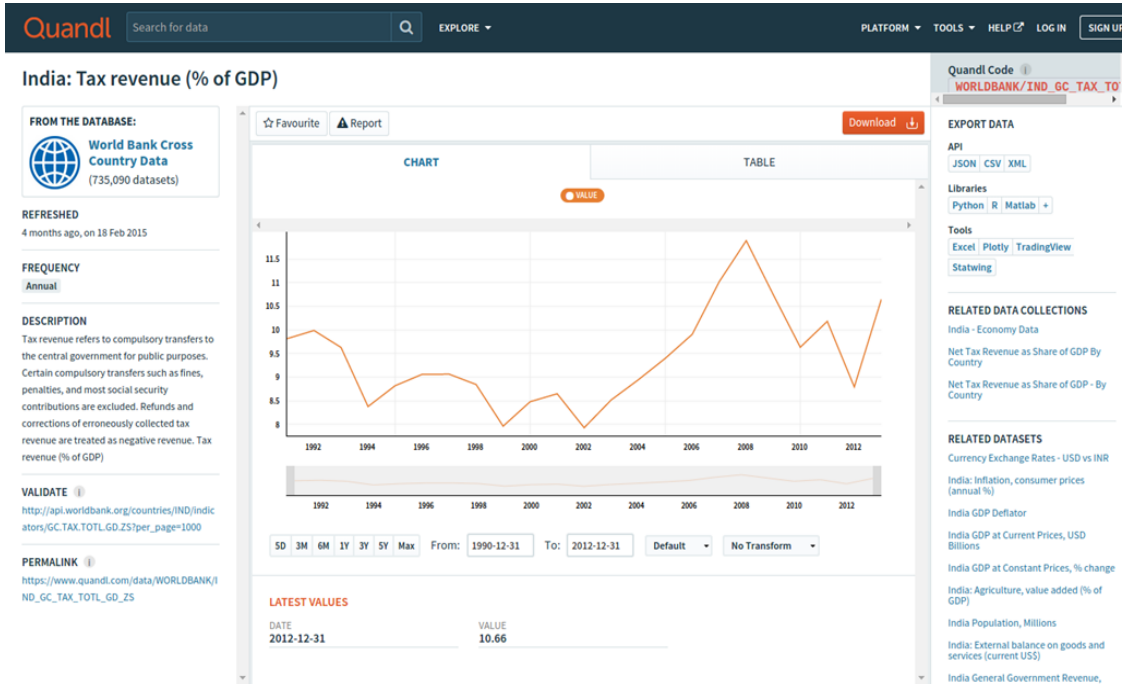


Figure 2. Compliance Management and Decision Support Systems (Derived from OECD, 2008:10; Power, 2002:13-16)

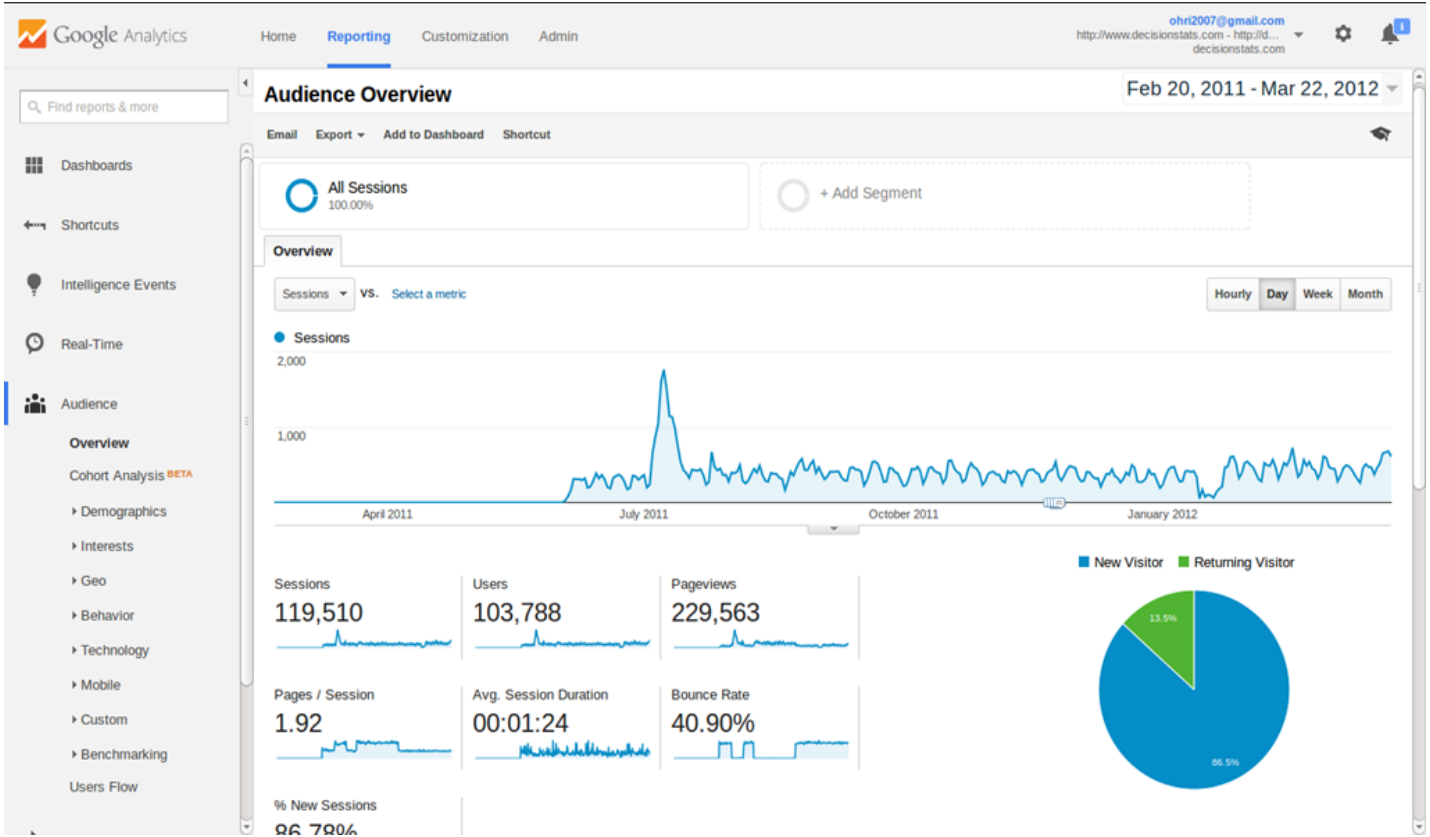
RESEARCH OBJECTIVE

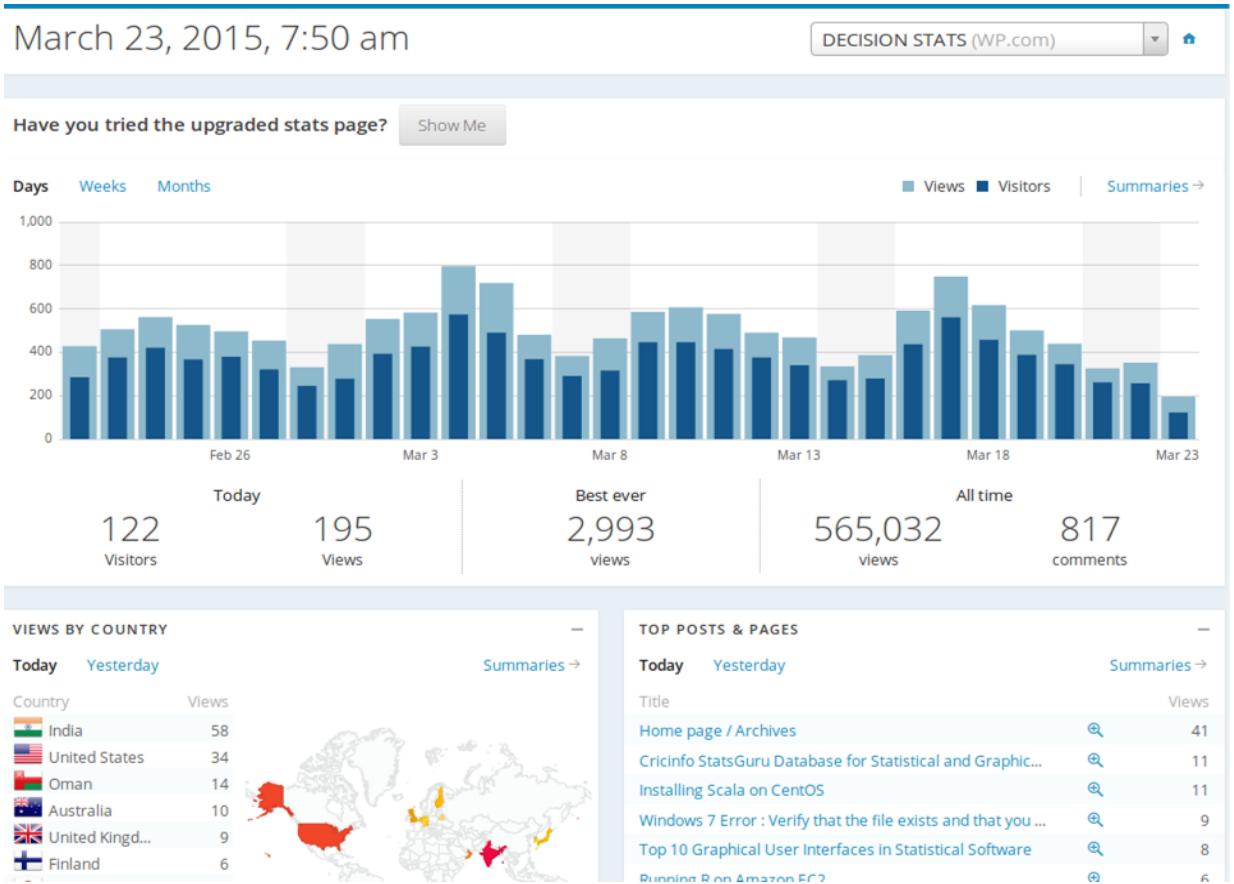
Web Data for Time Series



https://www.quandl.com/data/WORLDBANK/IND_GC_TAX_TOTL_GD_ZS-India-Tax-revenue-of-GDP

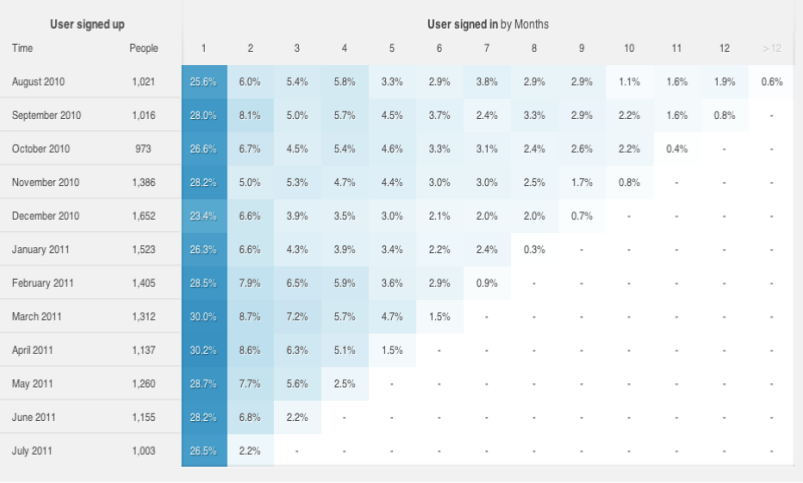
Introduction to Web Analytics


Google Analytics
Home
Reporting
Customization
Admin
ohri2007@gmail.com
http://www.decisionstats.com - http://d...
decisionstats.com
Find reports & more
Dashboards
Shortcuts
Intelligence Events
Real-Time
Audience
Overview
Cohort Analysis BETA
Demographics
Interests
Geo
Behavior
Technology
Mobile
Custom
Benchmarking
Users Flow
Audience Overview
Feb 20, 2011 - Mar 22, 2012
Email
Export
Add to Dashboard
Shortcut
All Sessions
100.00%
+ Add Segment
Overview
Sessions
VS.
Select a metric
Hourly
Day
Week
Month
Sessions
2,000
1,000
April 2011
July 2011
October 2011
January 2012
Sessions
119,510
Users
103,788
Pageviews
229,563
Pages / Session
1.92
Avg. Session Duration
00:01:24
Bounce Rate
40.90%
% New Sessions
86.78%
New Visitor
Returning Visitor
13.5%
86.5%


March 23, 2015, 7:50 am
DECISION STATS (WP.com)
Have you tried the upgraded stats page?
Show Me
Days
Weeks
Months
Views
Visitors
Summaries
1,000
800
600
400
200
0
Feb 26
Mar 3
Mar 8
Mar 13
Mar 18
Mar 23
Today
122
Visitors
195
Views
Best ever
2,993
views
All time
565,032
views
817
comments
VIEWS BY COUNTRY
Today
Yesterday
Summaries
Country
Views
India 58
United States 34
Oman 14
Australia 10
United Kingd... 9
Finland 6
TOP POSTS & PAGES
Today
Yesterday
Summaries
Title
Views
Home page / Archives 41
Cricinfo StatsGuru Database for Statistical and Graphic... 11
Installing Scala on CentOS 11
Windows 7 Error : Verify that the file exists and that you ... 9
Top 10 Graphical User Interfaces in Statistical Software 8

New: Cohort Analysis

Cohort analysis is a subset of behavioral analytics that takes the data from a given eCommerce platform, web application, or online game and rather than looking at all users as one unit, it breaks them into related groups for analysis. These related groups, or cohorts, usually share common characteristics or experiences within a defined timespan.

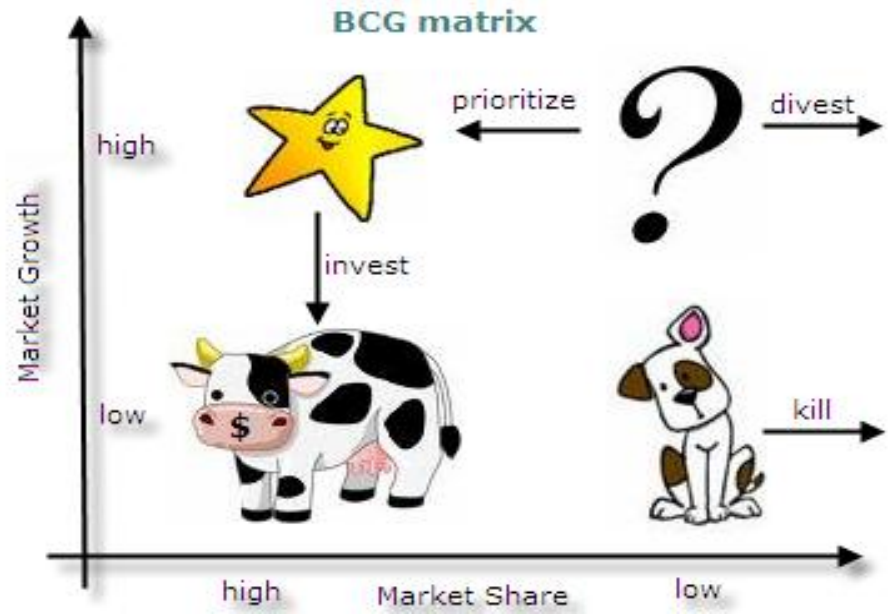


User signed up		User signed in by Months												
Time	People	1	2	3	4	5	6	7	8	9	10	11	12	>12
August 2010	1,021	25.6%	6.0%	5.4%	5.8%	3.3%	2.9%	3.8%	2.9%	2.9%	1.1%	1.6%	1.9%	0.6%
September 2010	1,016	28.0%	8.1%	5.0%	5.7%	4.5%	3.7%	2.4%	3.3%	2.9%	2.2%	1.6%	0.8%	-
October 2010	973	26.6%	6.7%	4.5%	5.4%	4.6%	3.3%	3.1%	2.4%	2.6%	2.2%	0.4%	-	-
November 2010	1,386	28.2%	5.0%	5.3%	4.7%	4.4%	3.0%	3.0%	2.5%	1.7%	0.8%	-	-	-
December 2010	1,652	23.4%	6.6%	3.9%	3.5%	3.0%	2.1%	2.0%	2.0%	0.7%	-	-	-	-
January 2011	1,523	26.3%	6.6%	4.3%	3.9%	3.4%	2.2%	2.4%	0.3%	-	-	-	-	-
February 2011	1,405	28.5%	7.9%	6.5%	5.9%	3.6%	2.9%	0.9%	-	-	-	-	-	-
March 2011	1,312	30.0%	8.7%	7.2%	5.7%	4.7%	1.5%	-	-	-	-	-	-	-
April 2011	1,137	30.2%	8.6%	6.3%	5.1%	1.5%	-	-	-	-	-	-	-	-
May 2011	1,260	28.7%	7.7%	5.6%	2.5%	-	-	-	-	-	-	-	-	-
June 2011	1,155	28.2%	6.8%	2.2%	-	-	-	-	-	-	-	-	-	-
July 2011	1,003	26.5%	2.2%	-	-	-	-	-	-	-	-	-	-	-

People		Weeks later											
		1	2	3	4	5	6	7	8	9	10	11	12
Oct 7, 2013	44	27.27%	20.45%	22.73%	18.18%	15.91%	11.36%	6.82%	13.64%	13.64%	9.09%	6.82%	2.27%
Oct 14, 2013	50	24.00%	14.00%	24.00%	14.00%	6.00%	14.00%	14.00%	12.00%	6.00%	2.00%	0.00%	
Oct 21, 2013	49	26.53%	20.41%	16.33%	8.16%	6.12%	12.24%	12.24%	8.16%	6.12%	0.00%		
Oct 28, 2013	43	16.28%	11.63%	11.63%	11.63%	11.63%	11.63%	11.63%	2.33%	0.00%			
Nov 4, 2013	69	21.74%	11.59%	7.25%	11.59%	13.04%	5.80%	2.90%	0.00%				
Nov 11, 2013	62	20.97%	14.52%	16.13%	11.29%	4.84%	3.23%	0.00%					
Nov 18, 2013	83	13.25%	13.25%	13.25%	8.43%	1.20%	1.20%						
Nov 25, 2013	74	17.57%	13.51%	8.11%	2.70%	0.00%							
Dec 2, 2013	97	17.53%	12.37%	1.03%	1.03%								
Dec 9, 2013	62	24.19%	6.45%	1.61%									
Dec 16, 2013	40	10.00%	5.00%										
Dec 23, 2013	16	6.25%											

Strategy Refresher- Business Strategy Models

BCG Matrix for Product Lines



Strategy Refresher- Business Strategy Models

Porter's Model for Industries

Strategy Refresher- Business Strategy Models

Grenier's Theory

More Strategy Anyone?

1. Porters 5 forces Model-To analyze industries
2. Business Canvas
3. BCG Matrix- To analyze Product Portfolios
4. Porters Diamond Model- To analyze locations
5. McKinsey 7 S Model-To analyze teams
6. Gernier Theory- To analyze growth of organization
7. Herzberg Hygiene Theory- To analyze soft aspects of individuals
8. Marketing Mix Model- To analyze marketing mix.

http://decisionstats.com/2013/12/19/business-strategy-models/

Introduction

The R language is now the leading language for analytics and statistics on this planet.This R training starts with R language basics and covers basics of analytics
This course provides hands-on experience to execute analytics using the R language.
There will be many challenging tasks and focused practicals for the learners.

Requirements

Installations

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

http://www.rstudio.com/products/rstudio/download/ RStudio

R Packages rattle Rcmdr Deducer

Episode 1

Learning Objectives

- learn about R
- install R and it’s packages

What will you learn from this lesson

- Installation of R, Rtools, R Studio, R packages and GUIs

- Using RStudio and Using GUIs

Data Driven Decision Making

- using data and trending historical data
- validating assumptions if any
- using champion challenger to test scenarios
- using experiments
- use baselines
- continuous improvement
    - customer experiences
    - costs
    - revenues

If you can't measure it, you can't manage it -Peter Drucker

What is R

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

http://www.r-project.org/about.html

Statistical Software Landscape

SAS
Python (Pandas)
IBM SPSS
R
Julia
Clojure
Octave

Matlab
JMP
E views

Using R with other software

https://rforanalytics.wordpress.com/useful-links-for-r/using-r-from-other-software/

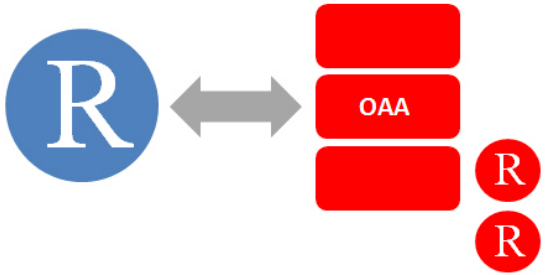


Tableau http://www.tableausoftware.com/new-features/r-integration

Qlik http://qliksolutions.ru/qlikview/add-ons/r-connector-eng/

Oracle R http://www.oracle.com/technetwork/database/database-technologies/r/r-enterprise/overview/index.html

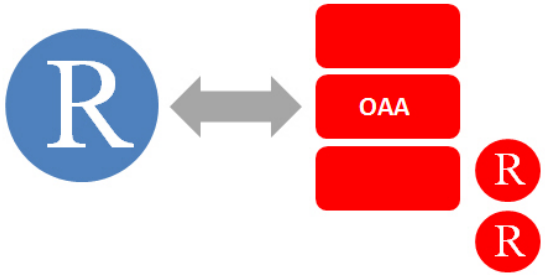
Rapid Miner https://rapid-i.com/content/view/202/206/lang,en/#r

JMP http://blogs.sas.com/jmp/index.php?/archives/298-JMP-Into-R!.html

Using R with other software

https://rforanalytics.wordpress.com/useful-links-for-r/using-r-from-other-software/

SAS/IML http://www.sas.com/technologies/analytics/statistics/iml/index.html



Teradata http://developer.teradata.com/applications/articles/in-database-analytics-with-teradata-r

Pentaho http://bigdatatechworld.blogspot.in/2013/10/integration-of-rweka-with-pentaho-data.html

IBM SPSS https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=ibm-analytics&S_PKG=ov18855&S_TACT=M161003W&dynform=127&lang=en_US

TIBCO TERR http://spotfire.tibco.com/discover-spotfire/what-does-spotfire-do/predictive-analytics/tibco-enterprise-runtime-for-r-terr

Some Advantages of R

open source
free
large number of algorithms and packages esp for statistics
flexible
very good for data visualization
superb community
rapidly growing
can be used with other software

Some Disadvantages of R

in memory (RAM) usage
steep learning curve
some IT departments frown on open source
verbose documentation
tech support
evolving ecosystem for corporates

Solutions for Disadvantages of R

in memory (RAM) usage ⇒ specialized packages, in database computing
steep learning curve ⇒ TRAINING !!!
some IT departments frown on open source ⇒ TRAINING and education!
verbose documentation ⇒ CRAN View , R Documentation
tech support ⇒ expanding pool of resources
evolving ecosystem for corporates ⇒ getting better with MS et al

R used by Government

- In the early days of the Deepwater Horizon disaster, NIST used uncertainty analysis in R to harmonize spill estimates from various sources, and to provide ranges of estimates to other agencies and the media.

- Before new drugs are allowed on the market, the FDA works with pharmaceutical companies to verify safety and efficacy through clinical trials. Despite a false perception that only commercial software may be used, many pharmaceutical companies are now using open-source R to analyze data from clinical trials.

- The National Weather Service uses R for research and development of models to predict river flooding.

- The newly-formed Consumer Financial Protection Bureau -- freed from the restrictions of a legacy IT infrastructure -- is championing the use of open-source technologies in government.

- Local governments are also building data-based applications. The SF Estuary Institute uses R and Google Maps to provide a tool to track pollution in the San Francisco Bay area.

http://gsnmagazine.com/node/26483?c=cyber_security

R used by Telecom

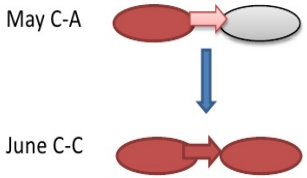
- Churn using

Social Network Analysis

http://www.slideshare.net/dataspora/social-network-analysis-for-telecoms

Results: A Customer With a Canceller in Their Network Churns at Twice the Rate

Types of Connections (Edges)



reality	expected by chance	delta
X	Y	**2.0**

In essence, we are asking whether being connected to another canceller has any effect on one's rate of cancellation. It turns out that it does.

And if we only look at voluntary port-outs, we see that customers churn at 3x the rate.

R used by Insurance

a few more insurance related packages:

- ChainLadder – Reserving methods in R. The package provides Mack-, Munich-, Bootstrap, and Multivariate-chain-ladder methods, as well as the LDF Curve Fitting methods of Dave Clark and GLM-based reserving models.
- cplm – Monte Carlo EM algorithms and Bayesian methods for fitting Tweedie compound Poisson linear models
- lossDev – A Bayesian time series loss development model. Features include skewed-t distribution with time-varying scale parameter, Reversible Jump MCMC for determining the functional form of the consumption path, and a structural break in this path; by Christopher W. Laws and Frank A. Schmid
- actuar: Loss distributions modelling, risk theory (including ruin theory), simulation of compound hierarchical models and credibility theory check out the actuar package by C. Dutang, V. Goulet and M. Pigeon.
- favir: Formatted Actuarial Vignettes in R. FAViR lowers the learning curve of the R environment. It is a series of peer-reviewed Sweave papers that use a consistent style.
- mondate: R packackge to keep track of dates in terms of months
- lifecontingencies – Package to perform actuarial evaluation of life contingencies

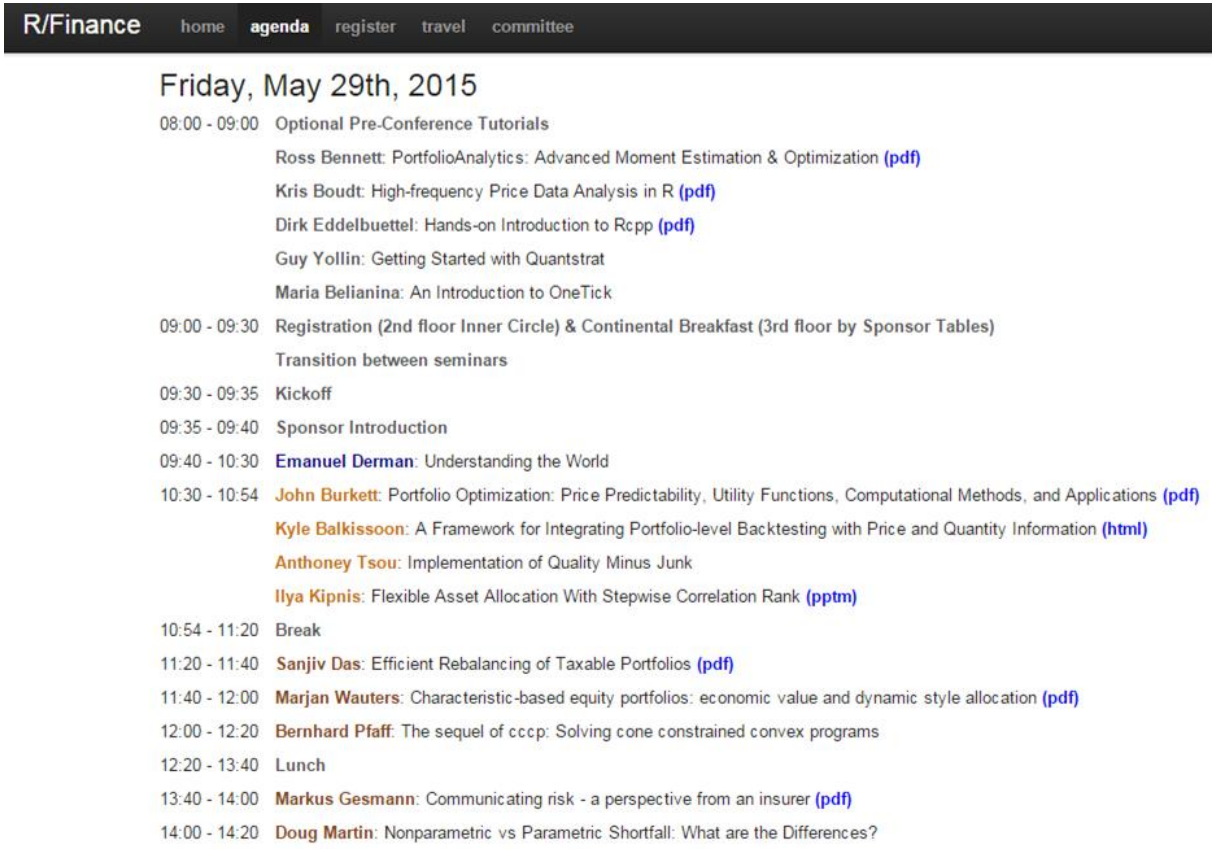
and

Introduction to R for Actuaries by Nigel de Silva

and http://www.rininsurance.com/

R in Finance

http://www.rinfinance.com/



R/Finance home agenda register travel committee

Friday, May 29th, 2015

08:00 - 09:00 Optional Pre-Conference Tutorials
- Ross Bennett: PortfolioAnalytics: Advanced Moment Estimation & Optimization (pdf)
- Kris Boudt: High-frequency Price Data Analysis in R (pdf)
- Dirk Eddelbuettel: Hands-on Introduction to Rcpp (pdf)
- Guy Yollin: Getting Started with Quantstrat
- Maria Belianina: An Introduction to OneTick

09:00 - 09:30 Registration (2nd floor Inner Circle) & Continental Breakfast (3rd floor by Sponsor Tables)
- Transition between seminars

09:30 - 09:35 Kickoff

09:35 - 09:40 Sponsor Introduction

09:40 - 10:30 Emanuel Derman: Understanding the World

10:30 - 10:54 John Burkett: Portfolio Optimization: Price Predictability, Utility Functions, Computational Methods, and Applications (pdf)
- Kyle Balkissoon: A Framework for Integrating Portfolio-level Backtesting with Price and Quantity Information (html)
- Anthoney Tsou: Implementation of Quality Minus Junk
- Ilya Kipnis: Flexible Asset Allocation With Stepwise Correlation Rank (pptm)

10:54 - 11:20 Break

11:20 - 11:40 Sanjiv Das: Efficient Rebalancing of Taxable Portfolios (pdf)

11:40 - 12:00 Marjan Wauters: Characteristic-based equity portfolios: economic value and dynamic style allocation (pdf)

12:00 - 12:20 Bernhard Pfaff: The sequel of cccp: Solving cone constrained convex programs

12:20 - 13:40 Lunch

13:40 - 14:00 Markus Gesmann: Communicating risk - a perspective from an insurer (pdf)

14:00 - 14:20 Doug Martin: Nonparametric vs Parametric Shortfall: What are the Differences?

R in Finance

http://cran.r-project.org/web/views/Finance.html
This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic.

- The Rmetrics suite of packages comprises fArma, fAsianOptions, fAssets, fBasics, fBonds, timeDate (formerly: fCalendar), fCopulae, fExoticOptions, fExtremes, fGarch, fImport,fNonlinear, fOptions, fPortfolio, fRegression, timeSeries (formerly: fSeries), fTrading, fUnitRoots and contains a very large number of relevant functions for different aspect of empirical and computational finance.
- The RQuantLib package provides several option-pricing functions as well as some fixed-income functionality from the QuantLib project to R.
- The quantmod package offers a number of functions for quantitative modelling in finance as well as data acqusition, plotting and other utilities.
- The portfolio package contains classes for equity portfolio management; the portfolioSim builds a related simulation framework. The backtest offers tools to explore portfolio-based hypotheses about financial instruments. The stockPortfolio package provides functions for single index, constant correlation and multigroup models. The pa package offers performance attribution functionality for equity portfolios.
- The PerformanceAnalytics package contains a large number of functions for portfolio performance calculations and risk management.

R in Finance

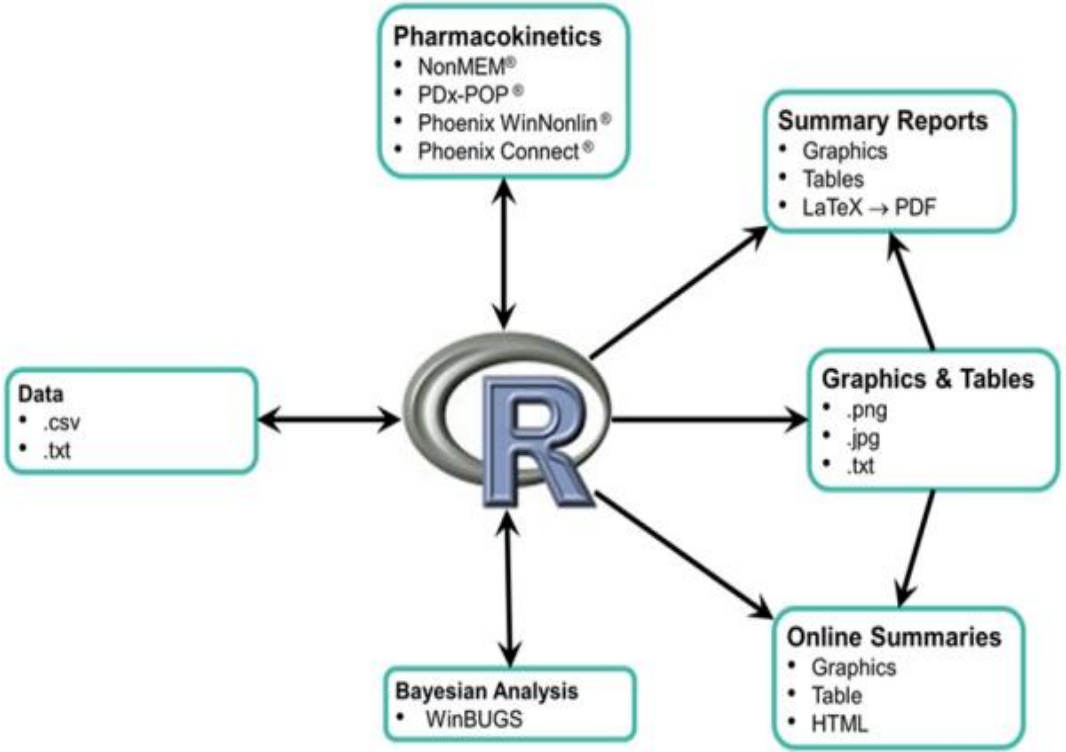
http://blog.revolutionanalytics.com/2013/08/r-drug-development-and-the-fda.html

Opening the Doors to Open Source Programming in Drug Development.

R: Regulatory Compliance and Validation Issues A Guidance Document for the Use of R in Regulated Clinical Trial Environments in which he concluded that useR 2012 FDA statistician Jea Brodsky presented a poster described how FDA scientists "use R on a daily basis" and have themselves written R packages for use at various stages in the drug submission process.

Open Source Software in the Biopharma Industry: Challenges and Opportunities,

R in Pharma

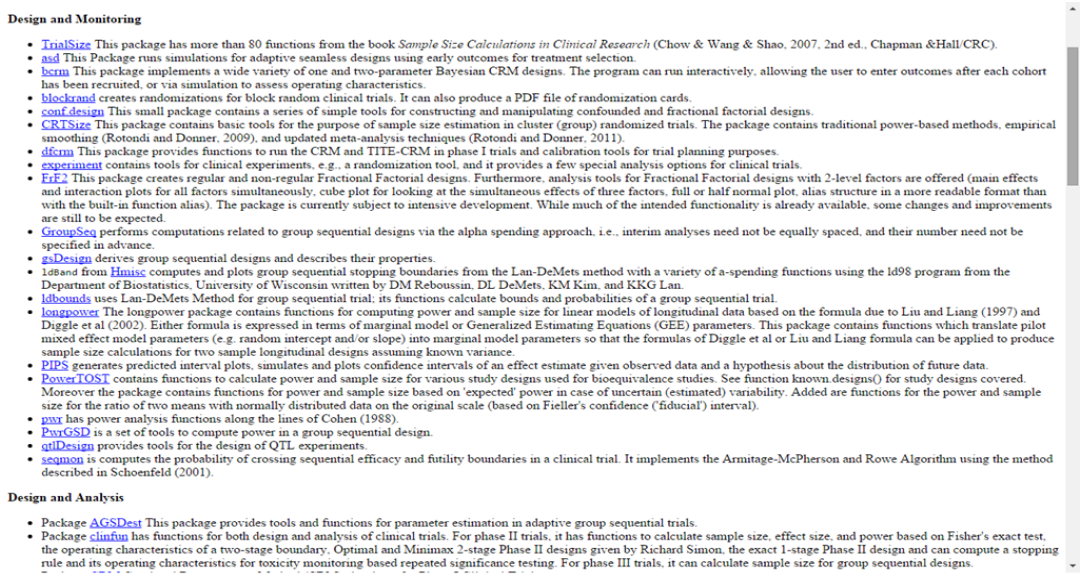


http://web.quanticate.com/bid/102741/Using-the-Statistical-Programming-Language-R-in-the-Pharma-Industry

R in Pharma

http://cran.r-project.org/web/views/ClinicalTrials.html

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including packages for clinical trial design and monitoring in general plus data analysis packages for a specific type of design.



Design and Monitoring

- TrialSize This package has more than 80 functions from the book *Sample Size Calculations in Clinical Research* (Chow & Wang & Shao, 2007, 2nd ed., Chapman &Hall/CRC).
- asd This Package runs simulations for adaptive seamless designs using early outcomes for treatment selection.
- bcrm This package implements a wide variety of one and two-parameter Bayesian CRM designs. The program can run interactively, allowing the user to enter outcomes after each cohort has been recruited, or via simulation to assess operating characteristics.
- blockrand creates randomizations for block random clinical trials. It can also produce a PDF file of randomization cards.
- conf.design This small package contains a series of simple tools for constructing and manipulating confounded and fractional factorial designs.
- CRTSize This package contains basic tools for the purpose of sample size estimation in cluster (group) randomized trials. The package contains traditional power-based methods, empirical smoothing (Rotondi and Donner, 2009), and updated meta-analysis techniques (Rotondi and Donner, 2011).
- dfcrm This package provides functions to run the CRM and TITE-CRM in phase I trials and calibration tools for trial planning purposes.
- experiment contains tools for clinical experiments, e.g., a randomization tool, and it provides a few special analysis options for clinical trials.
- FrF2 This package creates regular and non-regular Fractional Factorial designs. Furthermore, analysis tools for Fractional Factorial designs with 2-level factors are offered (main effects and interaction plots for all factors simultaneously, cube plot for looking at the simultaneous effects of three factors, full or half normal plot, alias structure in a more readable format than with the built-in function alias). The package is currently subject to intensive development. While much of the intended functionality is already available, some changes and improvements are still to be expected.
- GroupSeq performs computations related to group sequential designs via the alpha spending approach, i.e., interim analyses need not be equally spaced, and their number need not be specified in advance.
- gsDesign derives group sequential designs and describes their properties.
- ldBand from Hmisc computes and plots group sequential stopping boundaries from the Lan-DeMets method with a variety of a-spending functions using the ld98 program from the Department of Biostatistics, University of Wisconsin written by DM Reboussin, DL DeMets, KM Kim, and KKG Lan.
- ldbounds uses Lan-DeMets Method for group sequential trial; its functions calculate bounds and probabilities of a group sequential trial.
- longpower The longpower package contains functions for computing power and sample size for linear models of longitudinal data based on the formula due to Liu and Liang (1997) and Diggle et al (2002). Either formula is expressed in terms of marginal model or Generalized Estimating Equations (GEE) parameters. This package contains functions which translate pilot mixed effect model parameters (e.g. random intercept and/or slope) into marginal model parameters so that the formulas of Diggle et al or Liu and Liang formula can be applied to produce sample size calculations for two sample longitudinal designs assuming known variance.
- PIPS generates predicted interval plots, simulates and plots confidence intervals of an effect estimate given observed data and a hypothesis about the distribution of future data.
- PowerTOST contains functions to calculate power and sample size for various study designs used for bioequivalence studies. See function known.designs() for study designs covered. Moreover the package contains functions for power and sample size based on 'expected' power in case of uncertain (estimated) variability. Added are functions for the power and sample size for the ratio of two means with normally distributed data on the original scale (based on Fieller's confidence ('fiducial') interval).
- pwr has power analysis functions along the lines of Cohen (1988).
- PwrGSD is a set of tools to compute power in a group sequential design.
- qtlDesign provides tools for the design of QTL experiments.
- seqmon is computes the probability of crossing sequential efficacy and futility boundaries in a clinical trial. It implements the Armitage-McPherson and Rowe Algorithm using the method described in Schoenfeld (2001).

Design and Analysis

- Package AGSDest This package provides tools and functions for parameter estimation in adaptive group sequential trials.
- Package clinfun has functions for both design and analysis of clinical trials. For phase II trials, it has functions to calculate sample size, effect size, and power based on Fisher's exact test, the operating characteristics of a two-stage boundary, Optimal and Minimax 2-stage Phase II designs given by Richard Simon, the exact 1-stage Phase II design and can compute a stopping rule and its operating characteristics for toxicity monitoring based repeated significance testing. For phase III trials, it can calculate sample size for group sequential designs.

Companies using R

from http://www.revolutionanalytics.com/companies-using-r

ANZ, the fourth largest bank in Australia, using R for credit risk analysis

Bank of America uses R for reporting.

The Consumer Financial Protection Bureau uses R for data analysis.

Facebook
Facebook and R:

- Analysis of Facebook Status Updates
- Facebook's Social Network Graph
- How Google and Facebook are using R
- Predicting Colleague Interactions with R

Pre Requisites

- Installation of R
  http://cran.rstudio.com/bin/windows/base/

- R Studio

- R Packages

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

R-3.1.1 for Windows (32/64 bit)

Download R 3.1.1 for Windows (54 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- How do I install R when using Windows Vista?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

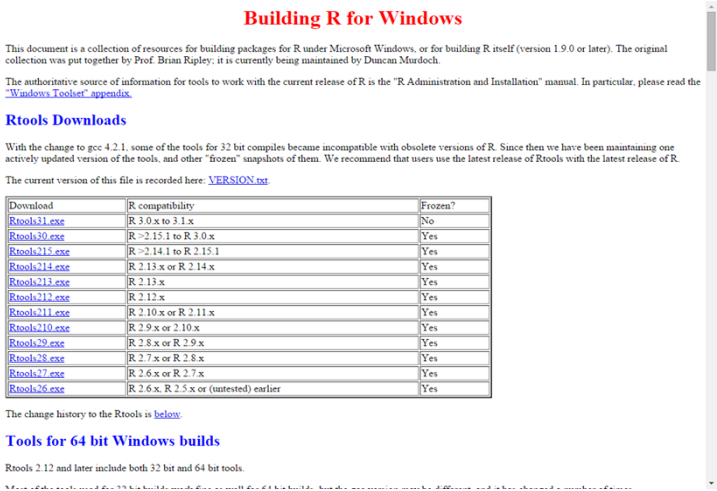
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2014-07-10, by Duncan Murdoch

Pre Requisites

- Installation of R
  - Rtools

    http://cran.rstudio.com/bin/windows/Rtools/



Building R for Windows

This document is a collection of resources for building packages for R under Microsoft Windows, or for building R itself (version 1.9.0 or later). The original collection was put together by Prof. Brian Ripley; it is currently being maintained by Duncan Murdoch.

The authoritative source of information for tools to work with the current release of R is the "R Administration and Installation" manual. In particular, please read the "Windows Toolset" appendix.

Rtools Downloads

With the change to gcc 4.2.1, some of the tools for 32 bit compiles became incompatible with obsolete versions of R. Since then we have been maintaining one actively updated version of the tools, and other "frozen" snapshots of them. We recommend that users use the latest release of Rtools with the latest release of R.

The current version of this file is recorded here: VERSION.txt.

Download	R compatibility	Frozen?
Rtools31.exe	R 3.0.x to 3.1.x	No
Rtools30.exe	R >2.15.1 to R 3.0.x	Yes
Rtools215.exe	R >2.14.1 to R 2.15.1	Yes
Rtools214.exe	R 2.13.x or R 2.14.x	Yes
Rtools213.exe	R 2.13.x	Yes
Rtools212.exe	R 2.12.x	Yes
Rtools211.exe	R 2.10.x or R 2.11.x	Yes
Rtools210.exe	R 2.9.x or 2.10.x	Yes
Rtools29.exe	R 2.8.x or R 2.9.x	Yes
Rtools28.exe	R 2.7.x or R 2.8.x	Yes
Rtools27.exe	R 2.6.x or R 2.7.x	Yes
Rtools26.exe	R 2.6.x, R 2.5.x or (untested) earlier	Yes

The change history to the Rtools is below.

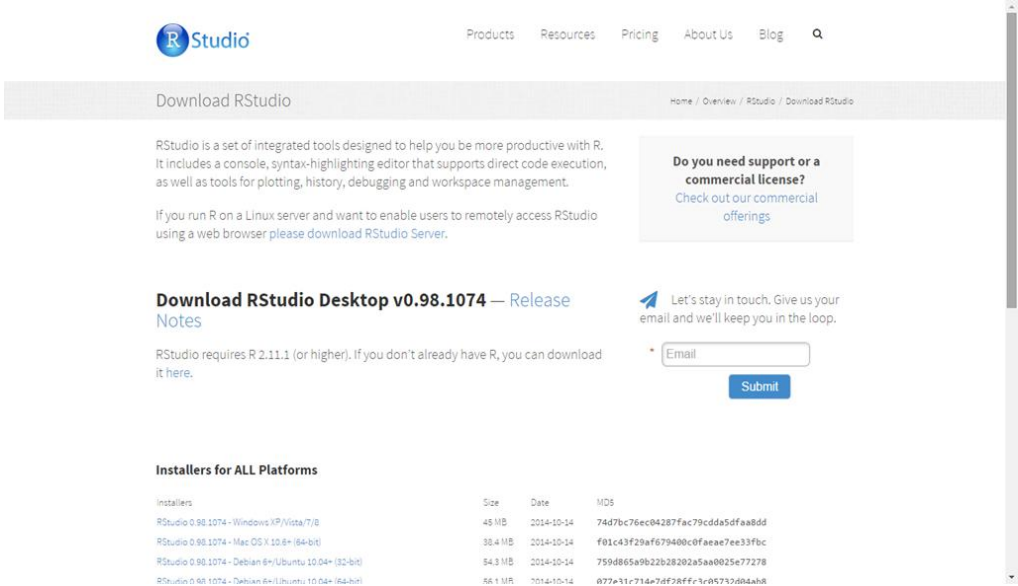
Tools for 64 bit Windows builds

Rtools 2.12 and later include both 32 bit and 64 bit tools.

- R Studio
- R Packages

Pre Requisites

- Installation of R
  - RTools
- R Studio
  http://www.rstudio.com/products/rstudio/download/
- R Packages

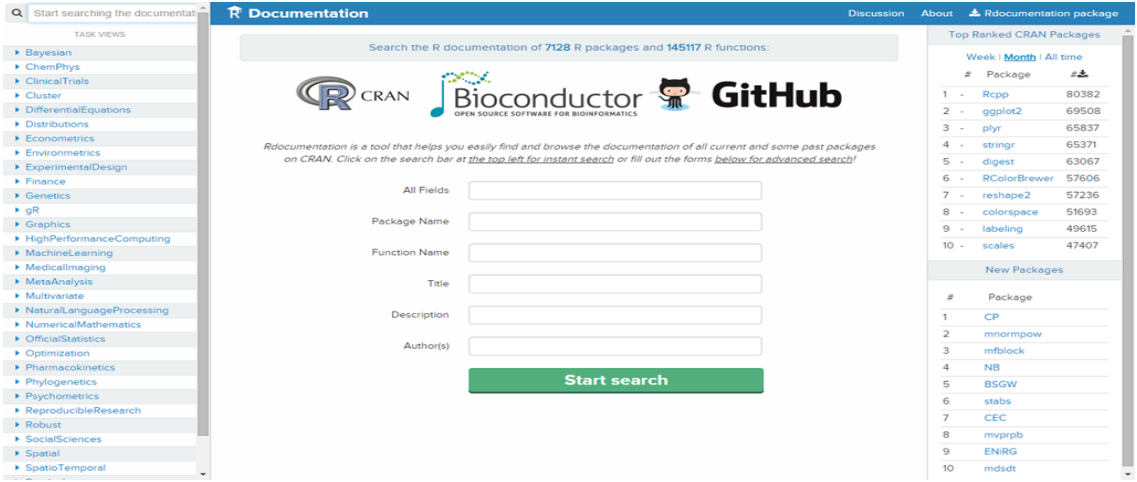


Pre Requisites

- Installation of R
  - Rtools

- R Studio

http://www.rstudio.com/products/rstudio/download/



- R Packages

about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

CRAN

107 sites in 49 regions

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

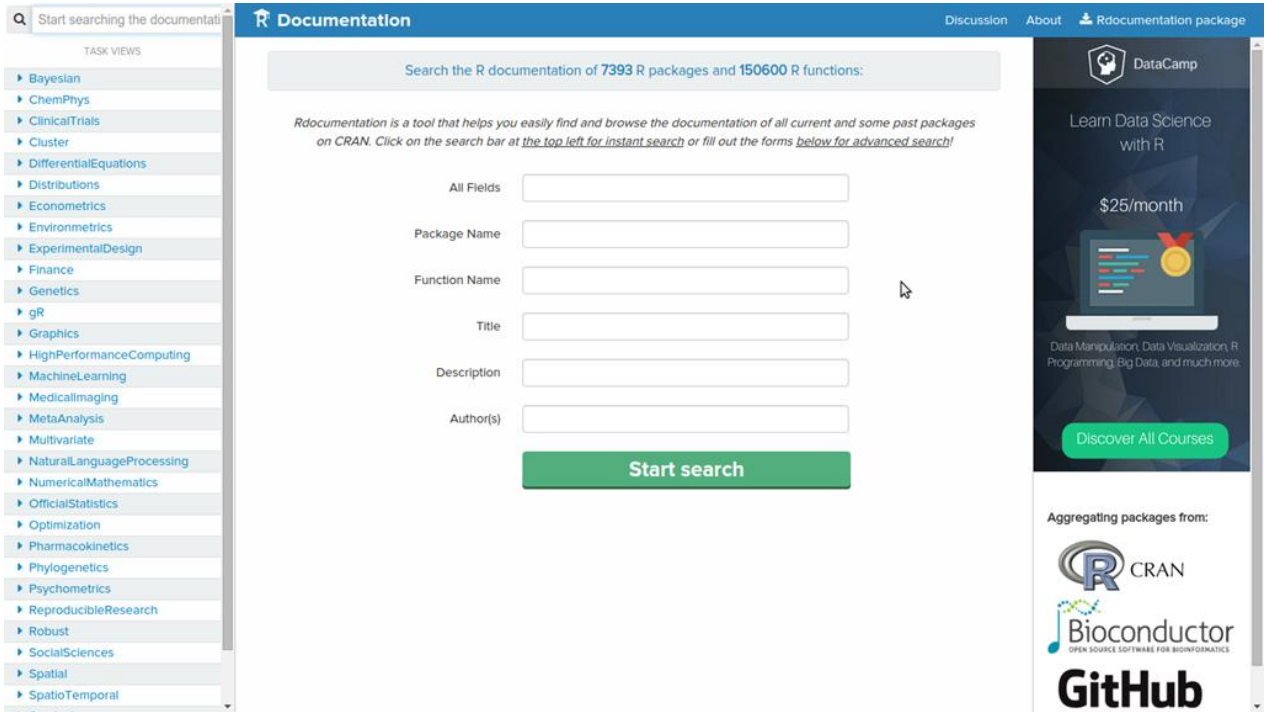
CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

0-Cloud	
http://cran.rstudio.com/	Rstudio, automatic redirection to servers worldwide
Algeria	
http://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
http://cran.csiro.au/	CSIRO
http://cran.ms.unimelb.edu.au/	University of Melbourne
Austria	
http://cran.at.r-project.org/	Wirtschaftsuniversitaet Wien
Belgium	
http://www.freestatistics.org/cran/	K.U.Leuven Association
Brazil	
http://nbcgib.uesc.br/mirrors/cran/	Center for Comp. Biol. at Universidade Estadual de Santa Cruz
http://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
http://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
http://www.vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
http://brieger.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba
Canada	
http://cran.stat.sfu.ca/	Simon Fraser University, Burnaby
http://mirror.its.dal.ca/cran/	Dalhousie University, Halifax
http://cran.utstat.utoronto.ca/	University of Toronto
http://cran.skazkaforyou.com/	iWeb, Montreal
http://cran.parentingamerica.com/	iWeb, Montreal
Chile	
http://dirichlet.mat.puc.cl/	Pontificia Universidad Catolica de Chile, Santiago
China	
http://ftp.ctex.org/mirrors/CRAN/	CTEX.ORG
http://mirror.bjtu.edu.cn/cran/	Beijing Jiaotong University, Beijing
http://mirrors.opencas.cn/cran/	Chinese Academy of Sciences, Beijing

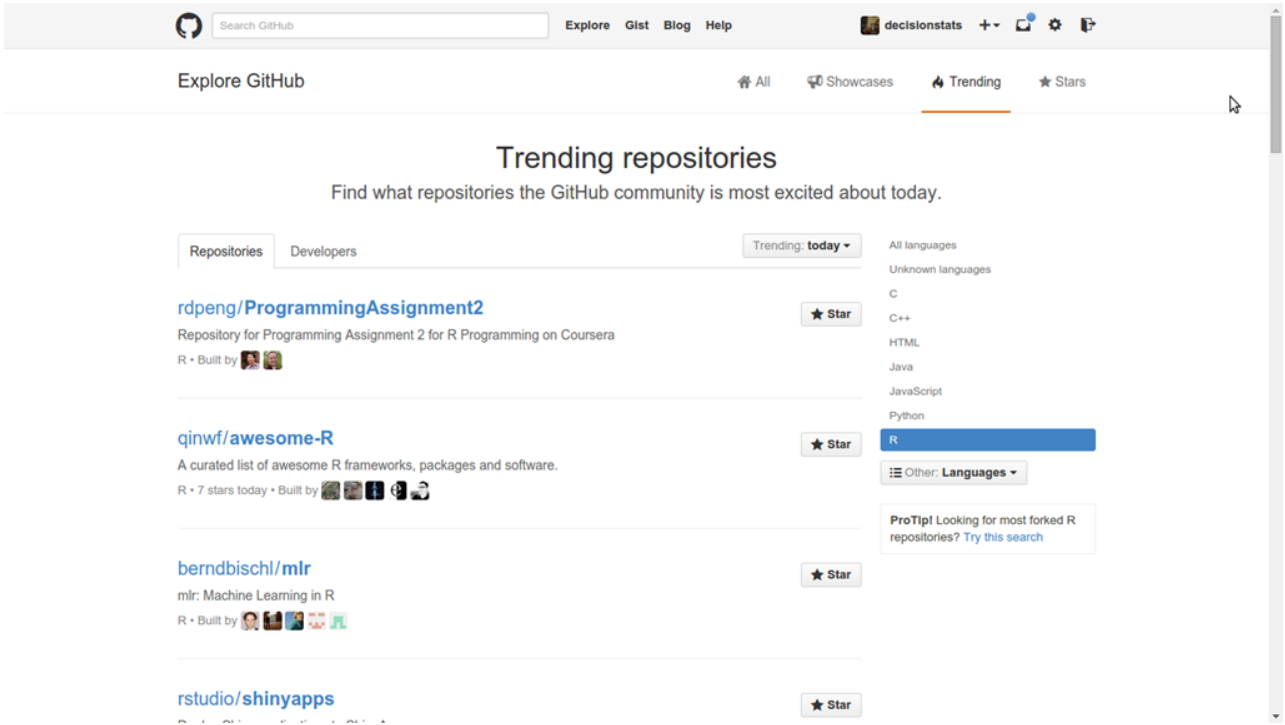
Non CRAN Repositories

http://www.rdocumentation.org/



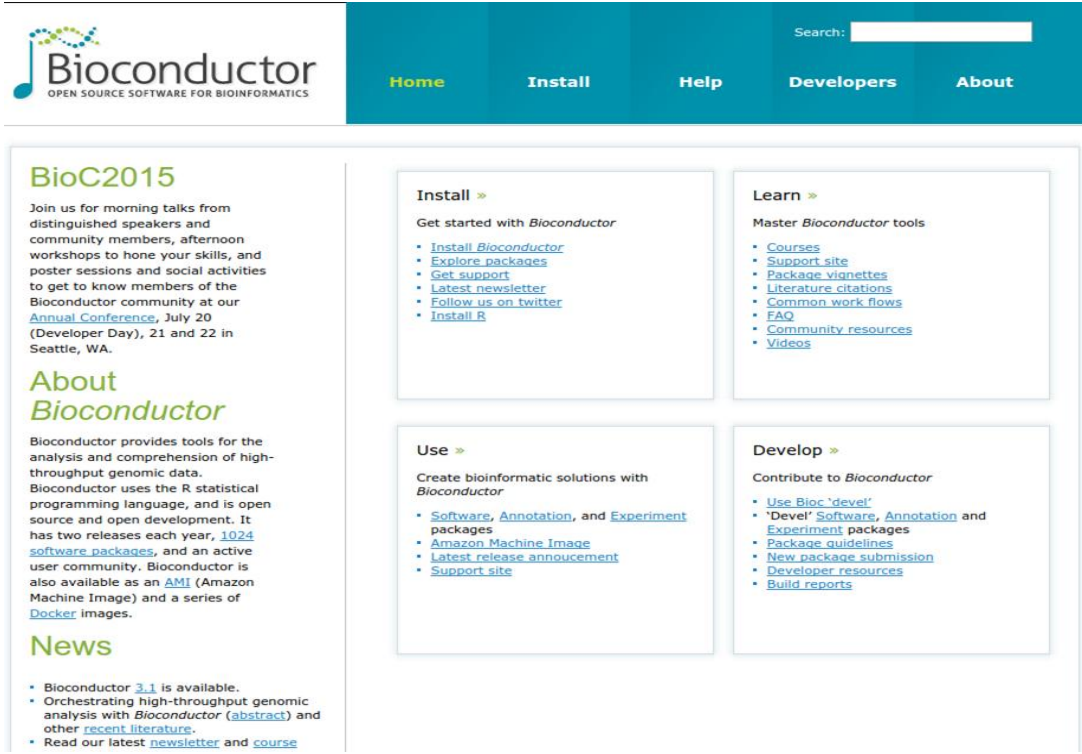
github

https://github.com/trending?l=R



bioconductor

http://www.bioconductor.org/



Pre Requisites

- R Packages

```
install.packages()  INSTALLS
update.packages()  UPDATES
library() LOADS
```

- Packages are **installed** once, updated periodically, but **loaded** every time

Interfaces to R

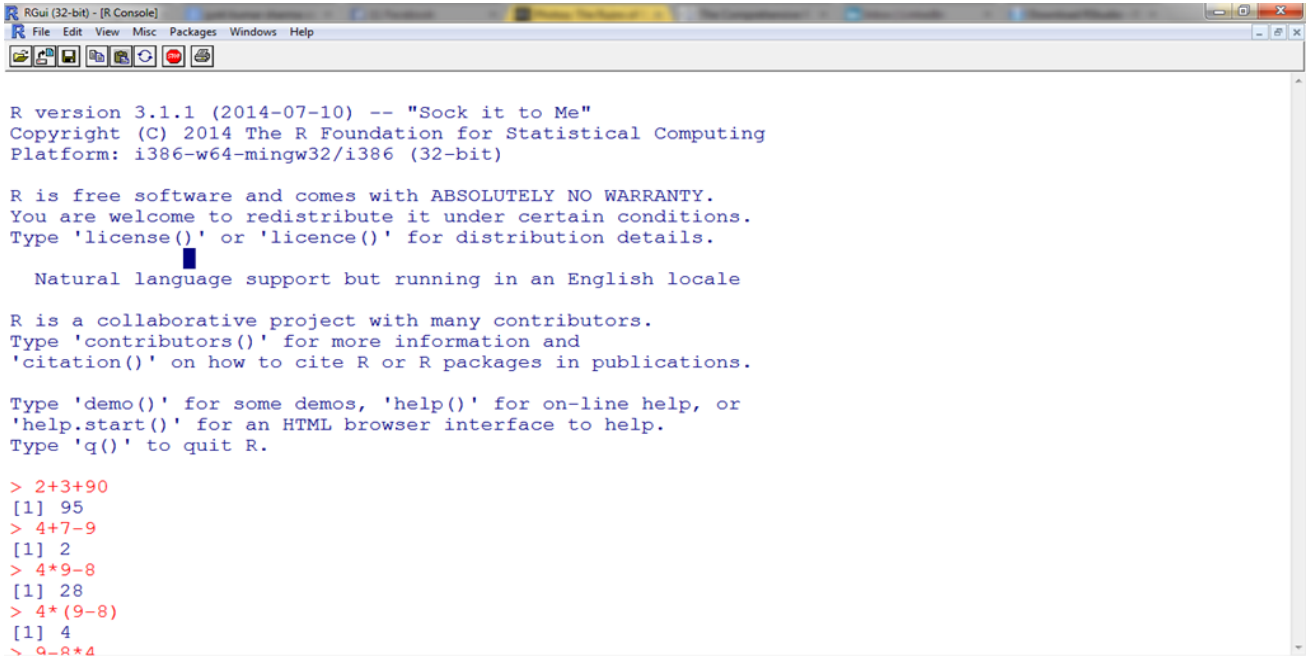
- Console

Default

Customization

- IDE

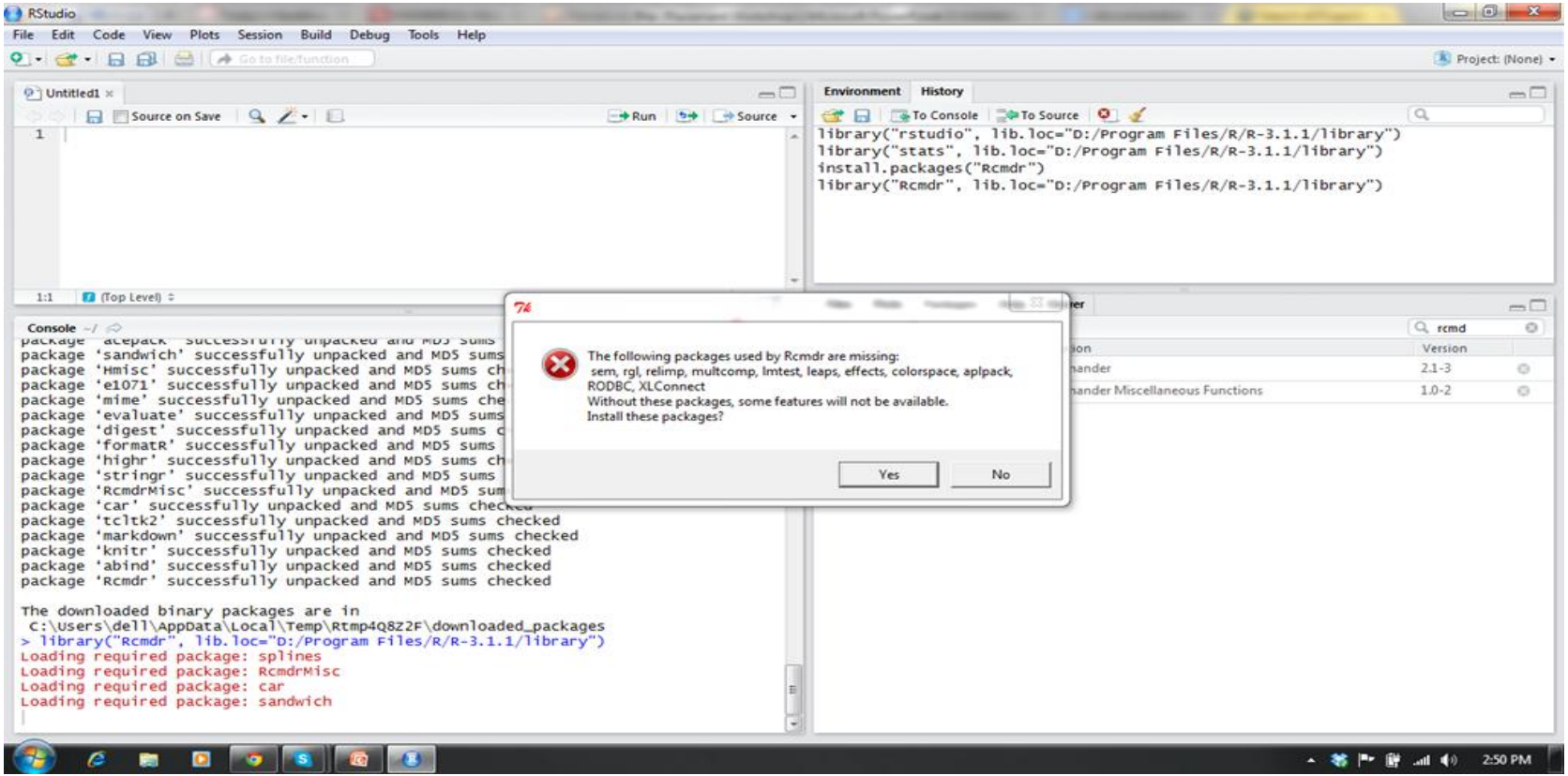
- GUI



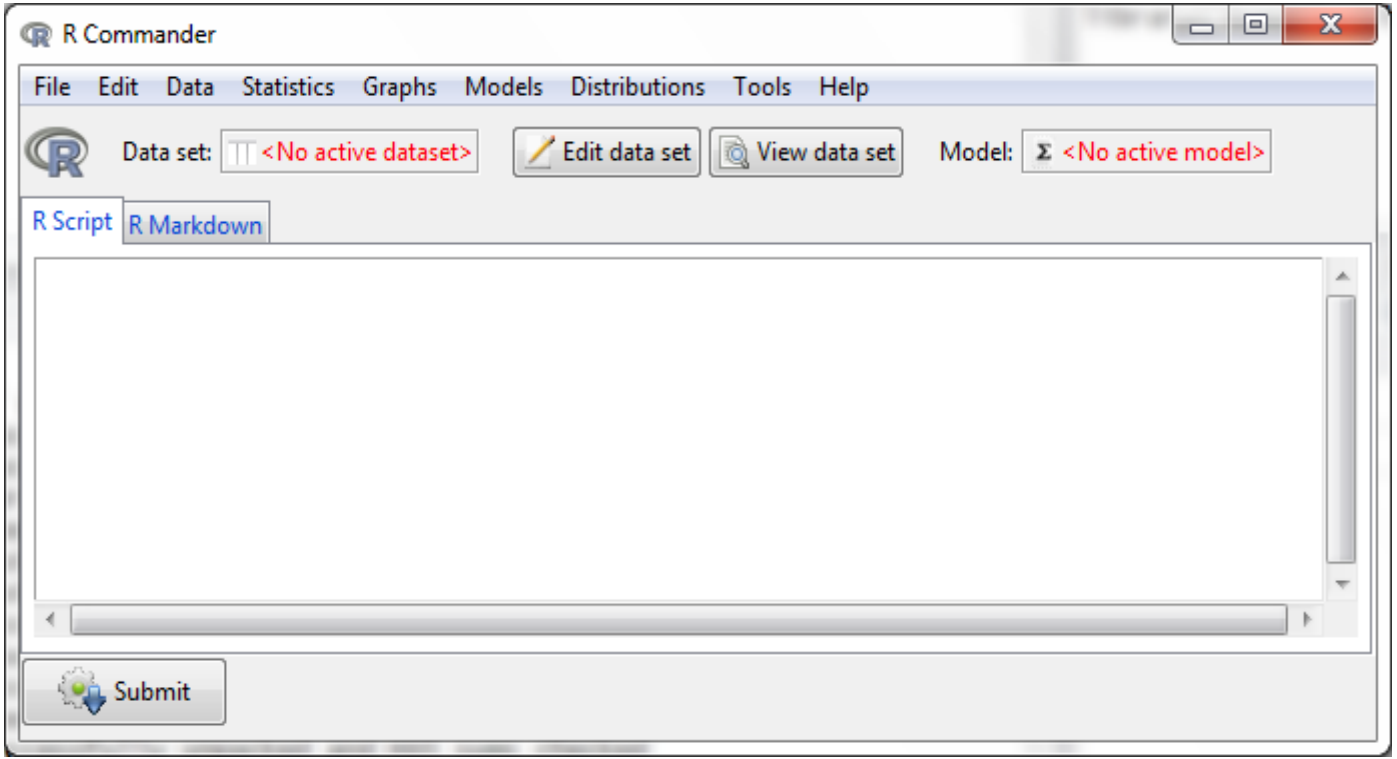
Graphical Interfaces to R

- R Commander
- Rattle
- Deducer

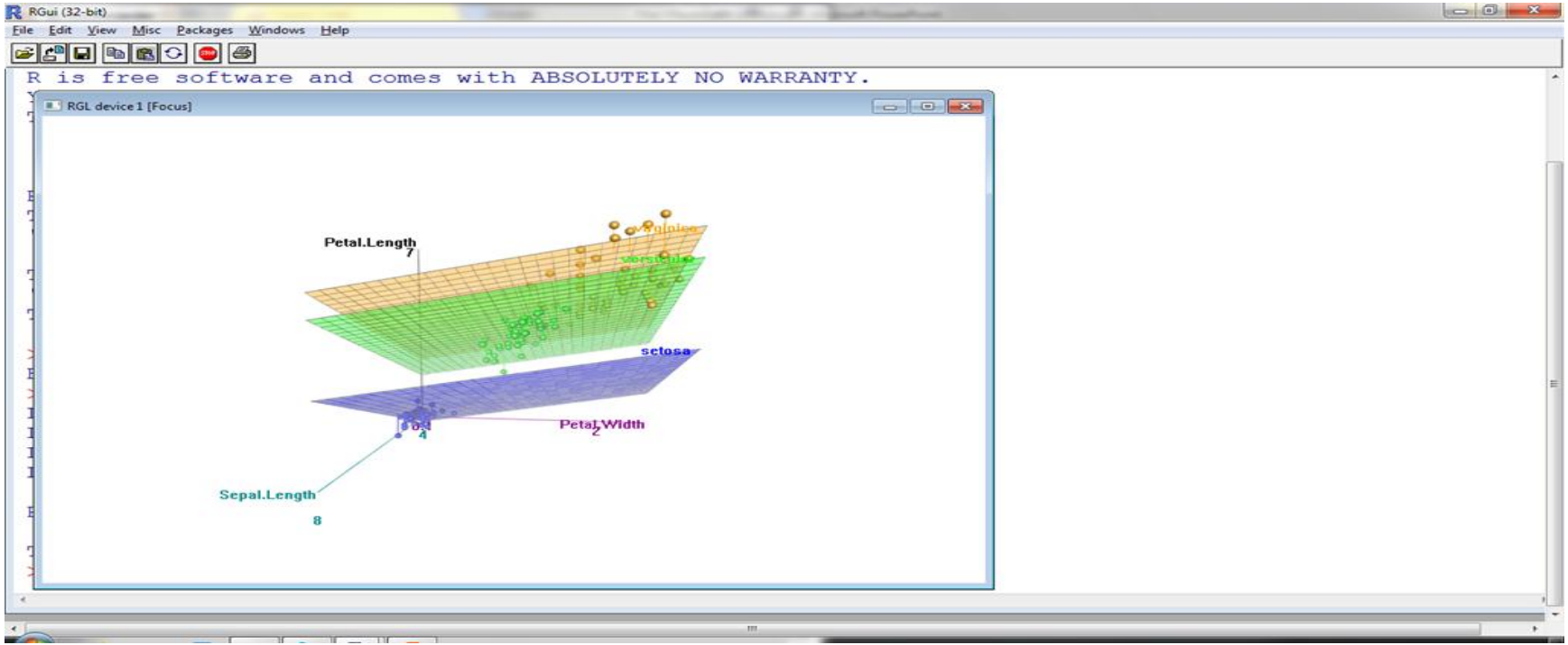
Installation of R Commander



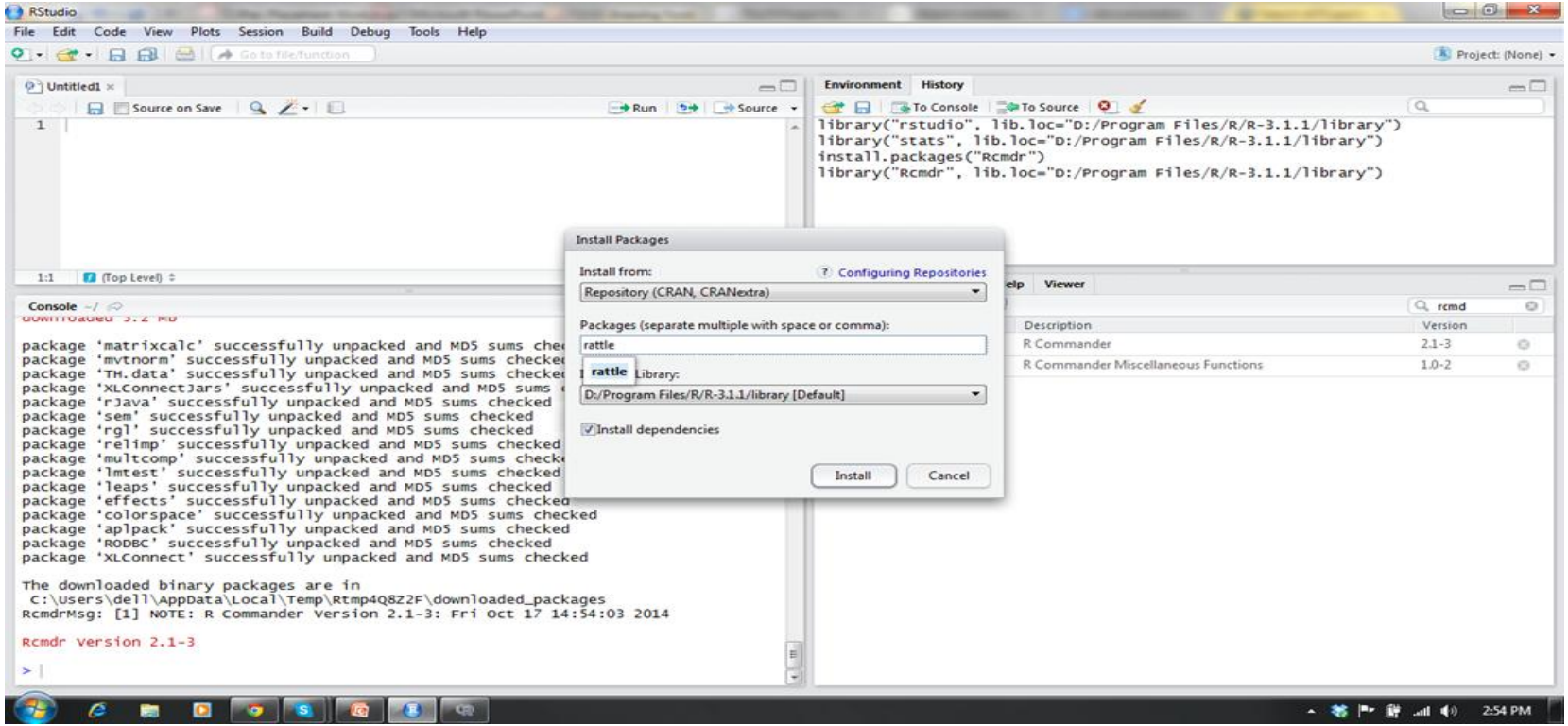
Overview of R Commander



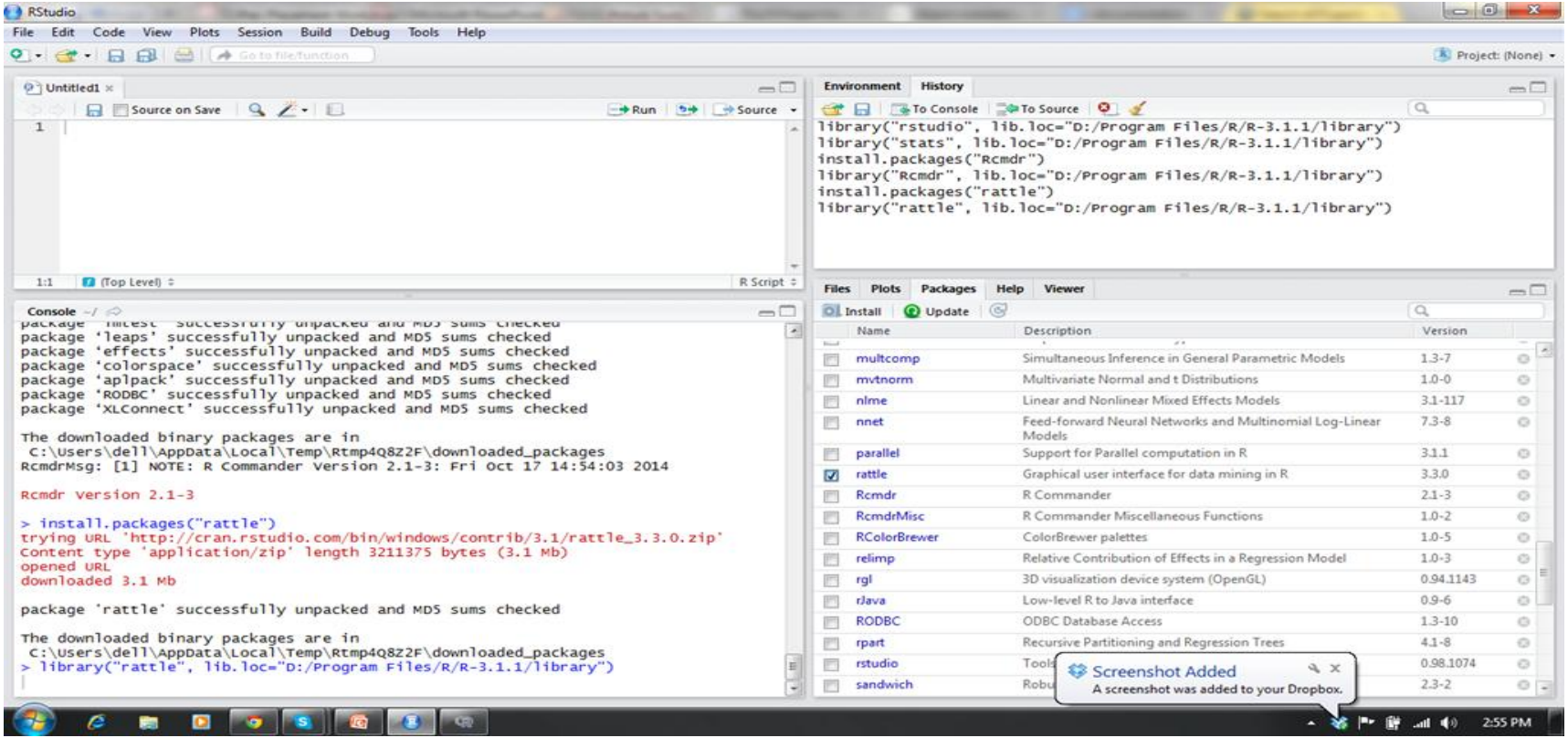
Demo
R Commander – 3D Graphs



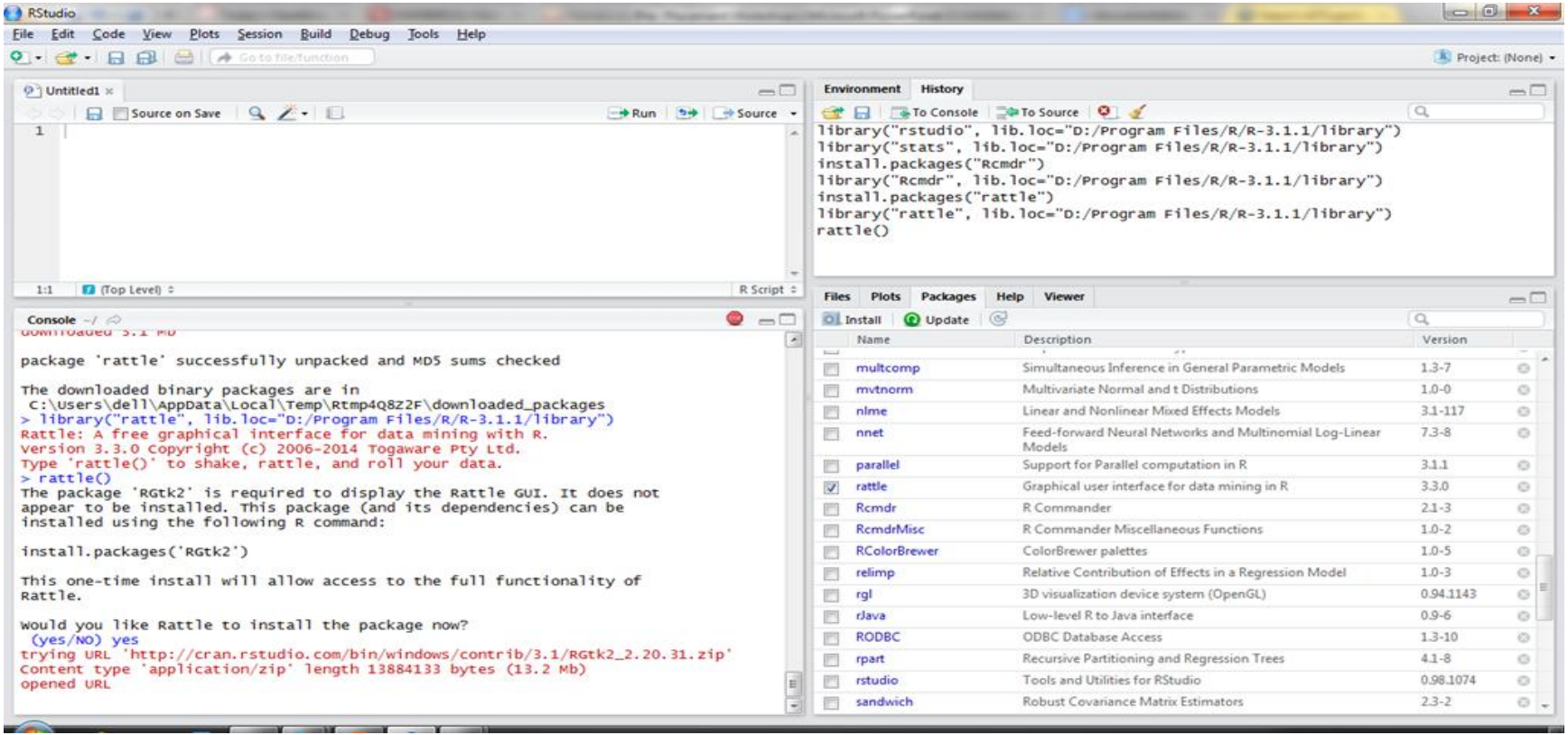
Installation of Rattle



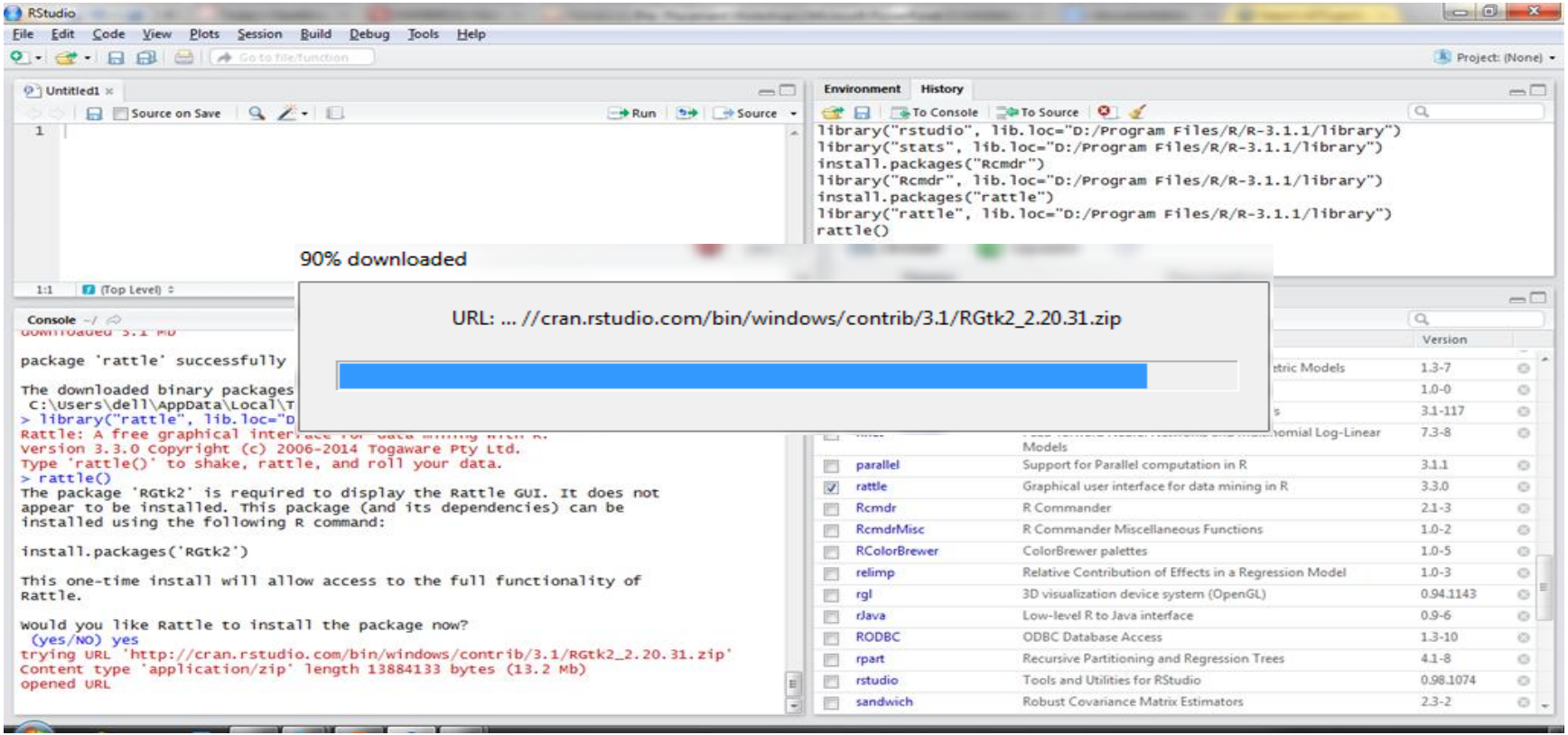
Installation of Rattle



Installation of Rattle

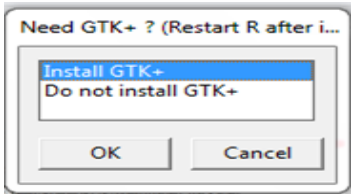


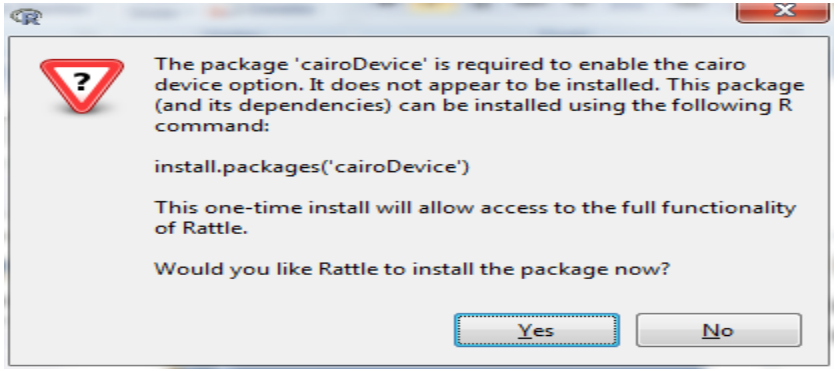
Installation of Rattle

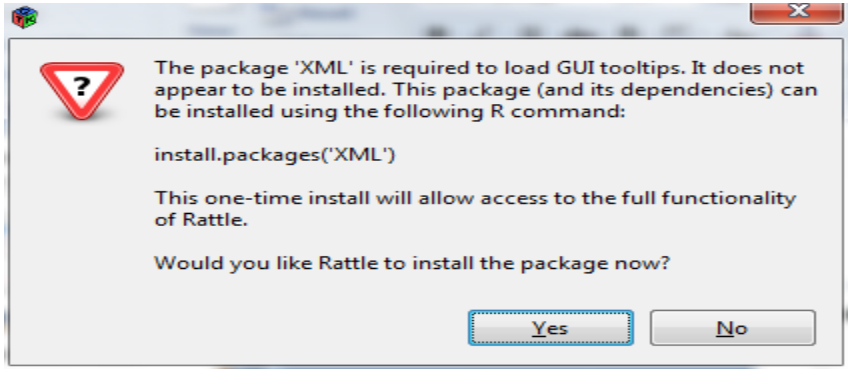


Installation of Rattle

- GTK+ Installation Necessary
- Install other packages when prompted

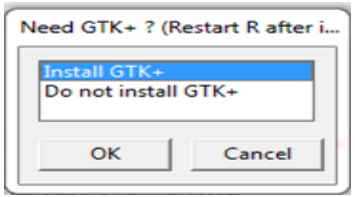




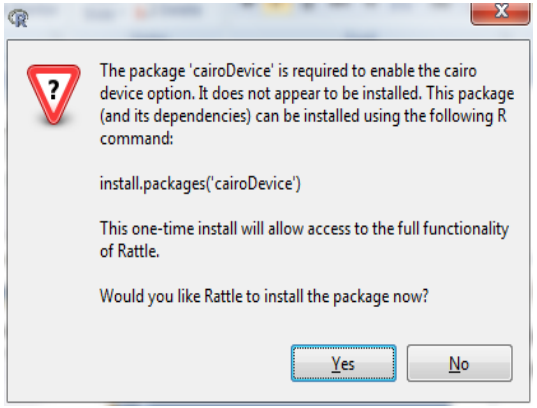


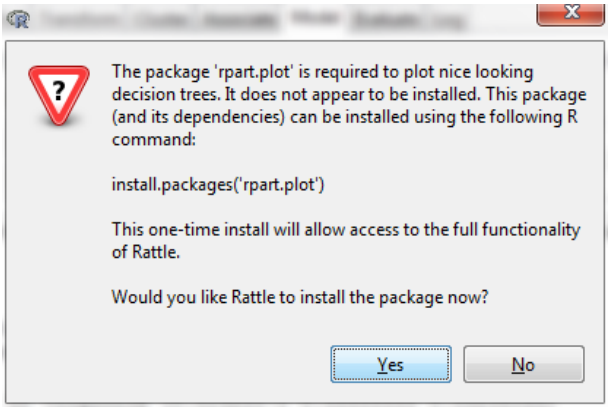
Installation of Rattle

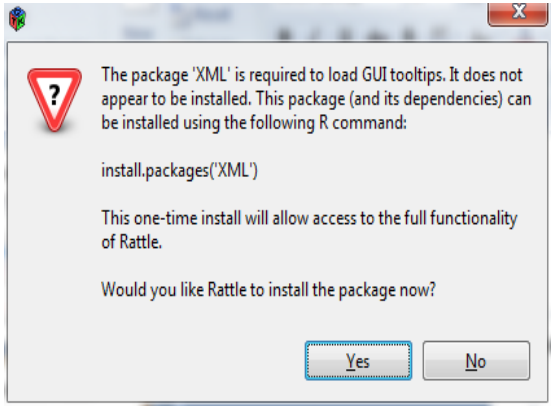
- GTK+ Installation Necessary



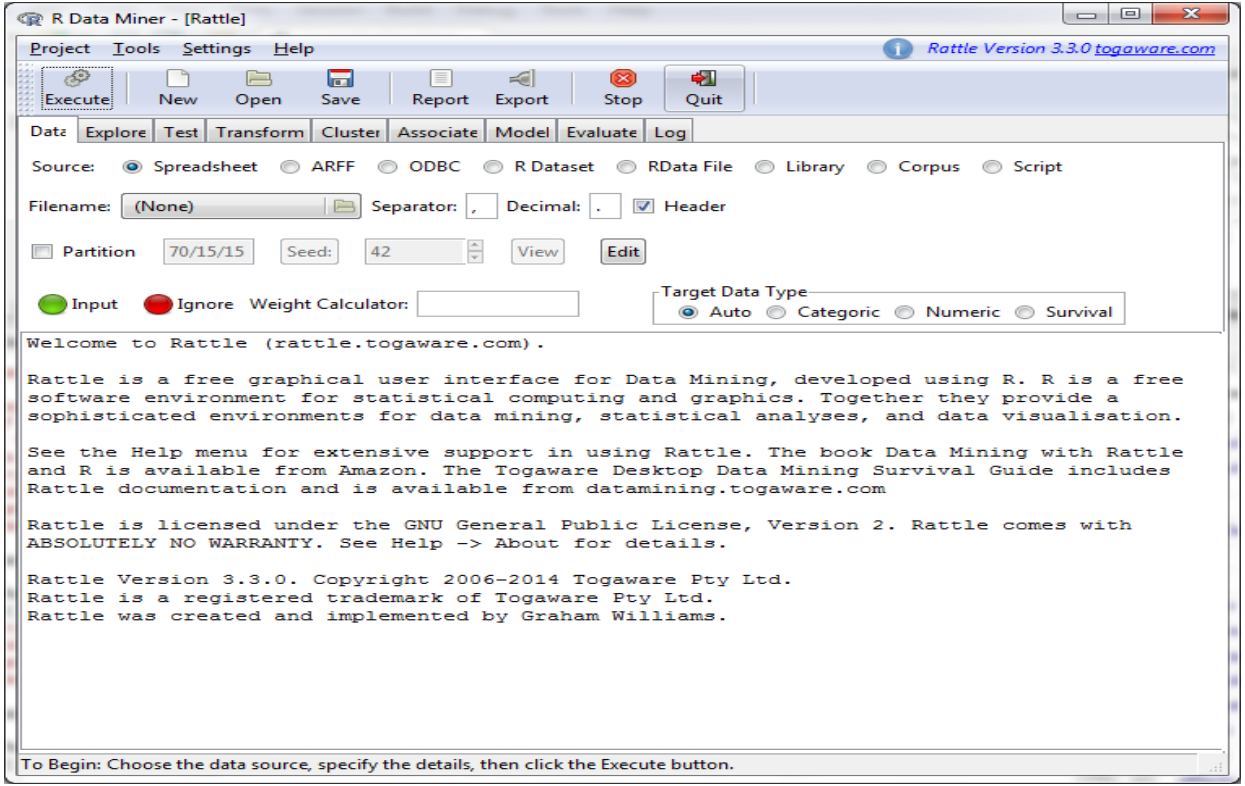
- Install other packages when prompted



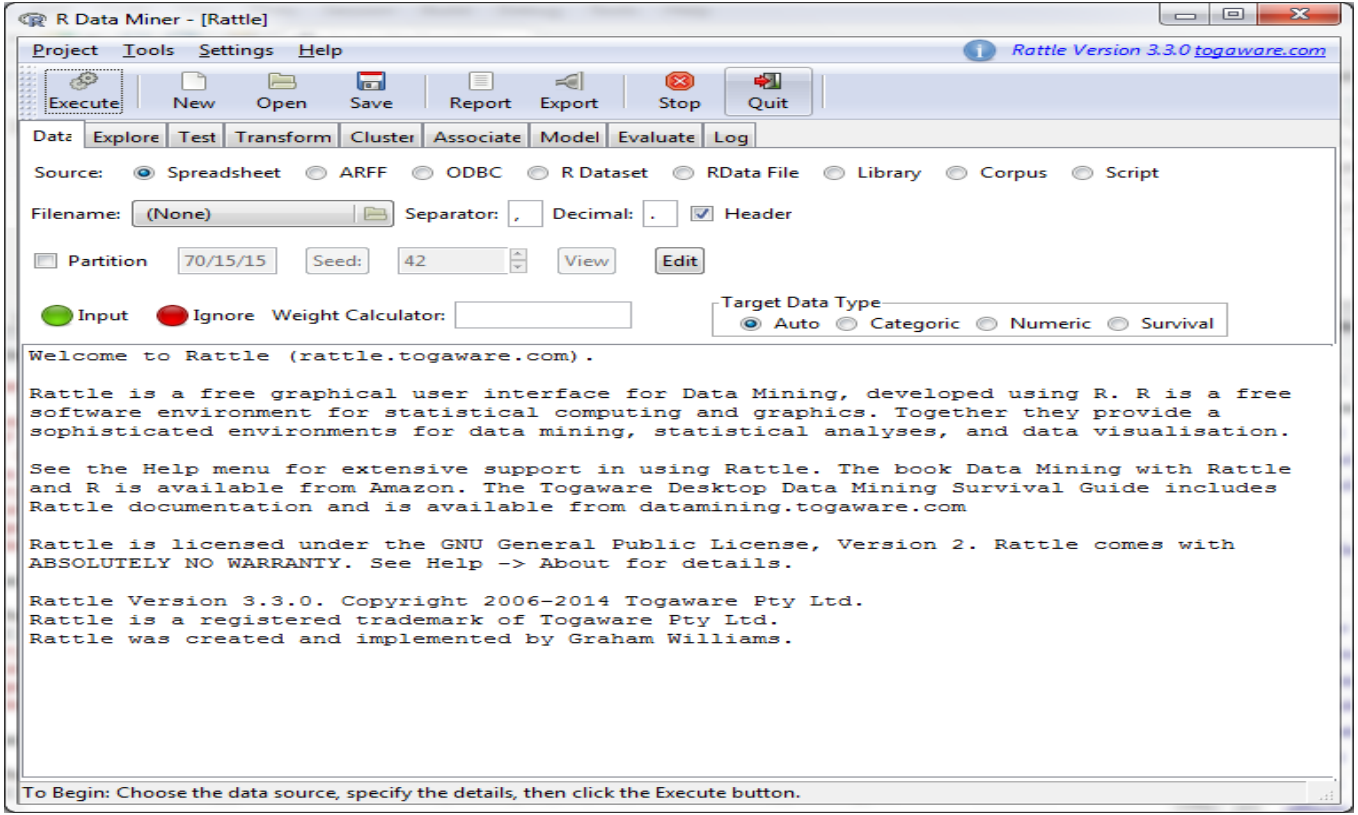




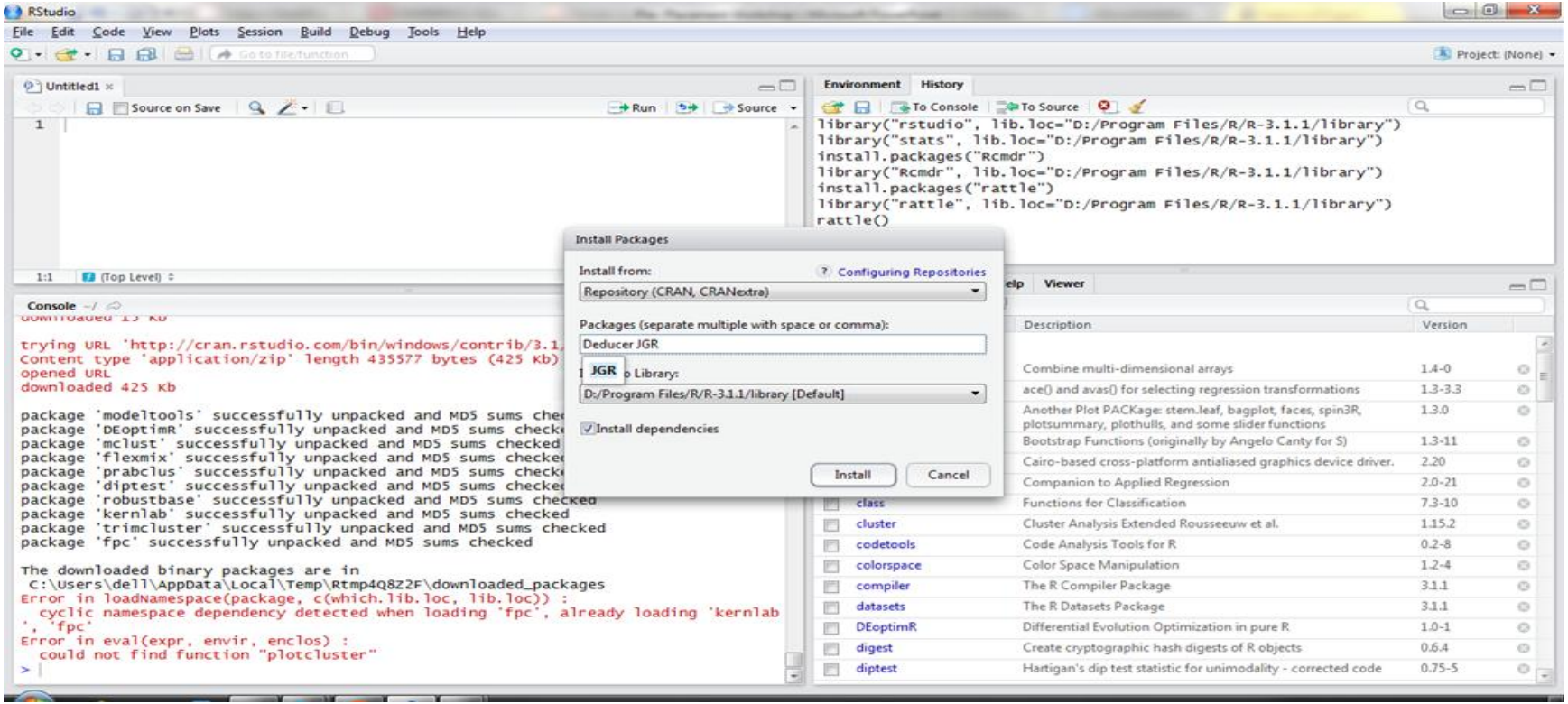
Overview of Rattle



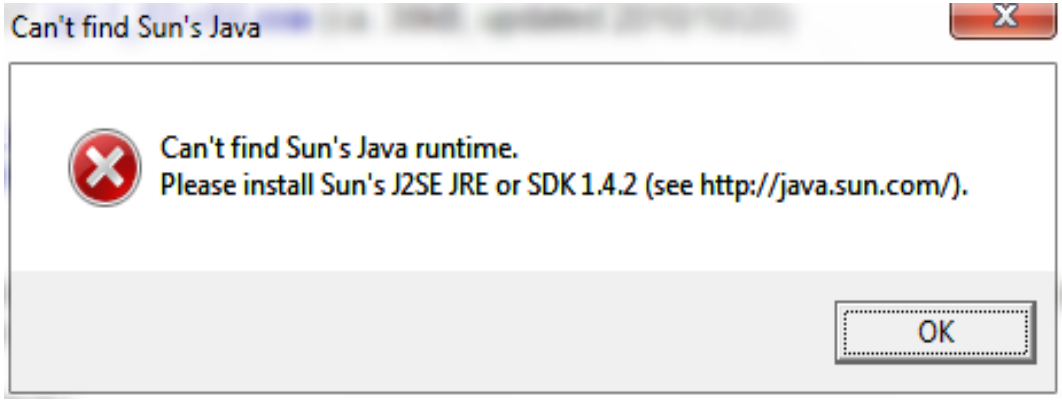
Demo Rattle


R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 3.3.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script
Filename: (None) Separator: , Decimal: . Header
Partition 70/15/15 Seed: 42 View Edit
Input Ignore Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
Welcome to Rattle (rattle.togaware.com).
Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environments for data mining, statistical analyses, and data visualisation.
See the Help menu for extensive support in using Rattle. The book Data Mining with Rattle and R is available from Amazon. The Togaware Desktop Data Mining Survival Guide includes Rattle documentation and is available from datamining.togaware.com
Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.
Rattle Version 3.3.0. Copyright 2006-2014 Togaware Pty Ltd.
Rattle is a registered trademark of Togaware Pty Ltd.
Rattle was created and implemented by Graham Williams.
To Begin: Choose the data source, specify the details, then click the Execute button.

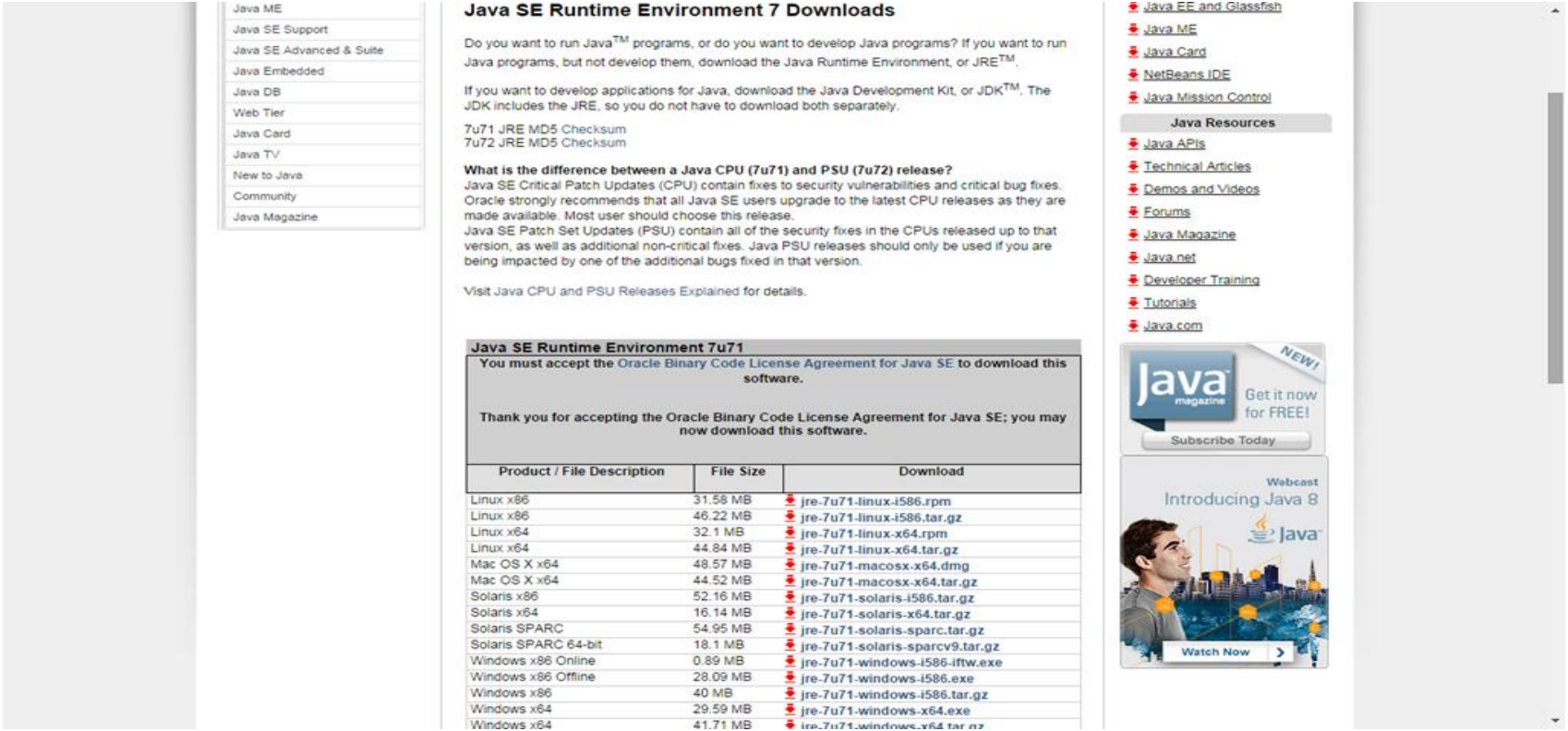
Installation Deducer (with JGR)



Installation Deducer (with JGR)



Installation Deducer (with JGR)



Java ME
Java SE Support
Java SE Advanced & Suite
Java Embedded
Java DB
Web Tier
Java Card
Java TV
New to Java
Community
Java Magazine

Java SE Runtime Environment 7 Downloads

Do you want to run Java™ programs, or do you want to develop Java programs? If you want to run Java programs, but not develop them, download the Java Runtime Environment, or JRE™.

If you want to develop applications for Java, download the Java Development Kit, or JDK™. The JDK includes the JRE, so you do not have to download both separately.

7u71 JRE MD5 Checksum
7u72 JRE MD5 Checksum

What is the difference between a Java CPU (7u71) and PSU (7u72) release?
Java SE Critical Patch Updates (CPU) contain fixes to security vulnerabilities and critical bug fixes. Oracle strongly recommends that all Java SE users upgrade to the latest CPU releases as they are made available. Most user should choose this release.
Java SE Patch Set Updates (PSU) contain all of the security fixes in the CPUs released up to that version, as well as additional non-critical fixes. Java PSU releases should only be used if you are being impacted by one of the additional bugs fixed in that version.

Visit Java CPU and PSU Releases Explained for details.

Java SE Runtime Environment 7u71

You must accept the Oracle Binary Code License Agreement for Java SE to download this software.

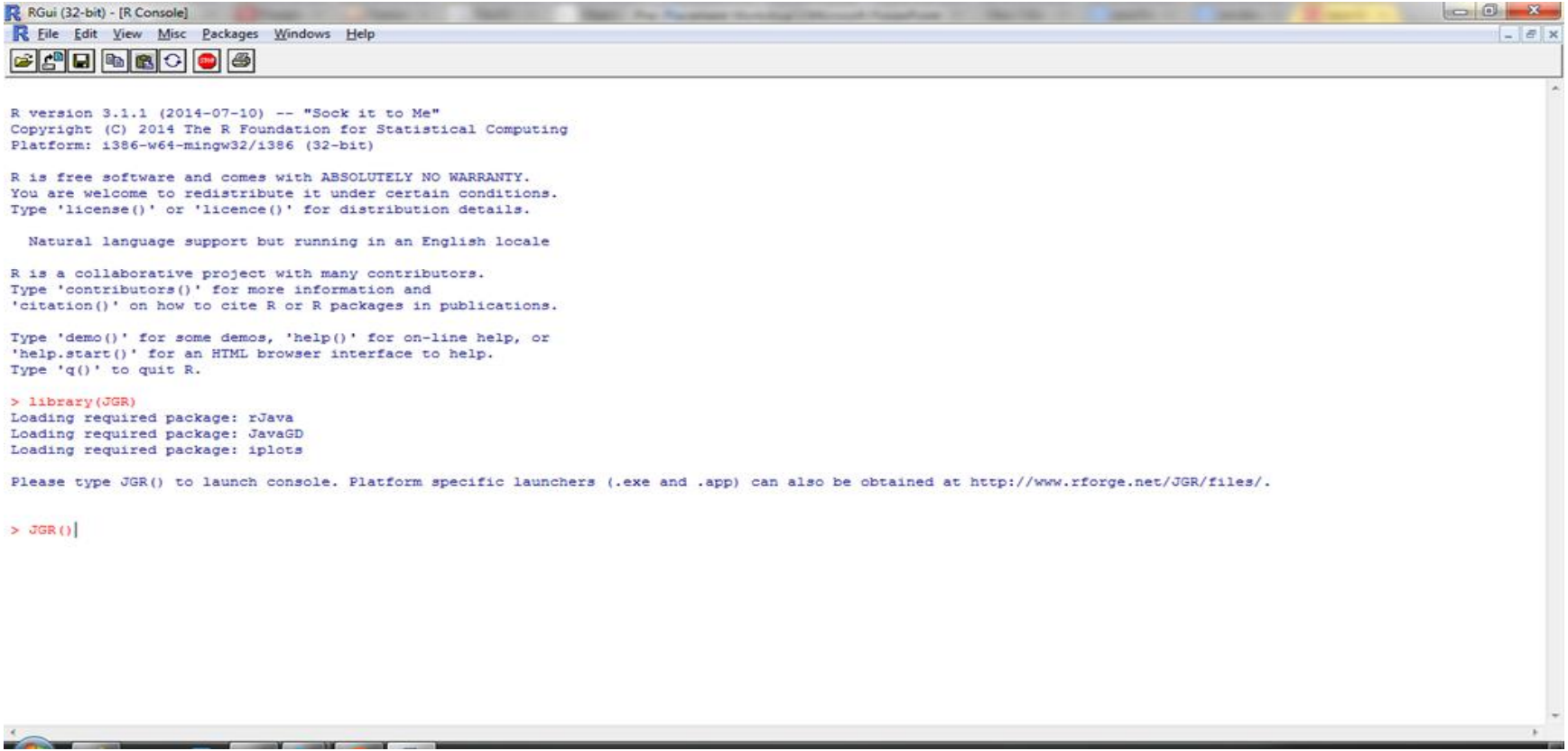
Thank you for accepting the Oracle Binary Code License Agreement for Java SE; you may now download this software.

Product / File Description	File Size	Download
Linux x86	31.58 MB	jre-7u71-linux-i586.rpm
Linux x86	46.22 MB	jre-7u71-linux-i586.tar.gz
Linux x64	32.1 MB	jre-7u71-linux-x64.rpm
Linux x64	44.84 MB	jre-7u71-linux-x64.tar.gz
Mac OS X x64	48.57 MB	jre-7u71-macosx-x64.dmg
Mac OS X x64	44.52 MB	jre-7u71-macosx-x64.tar.gz
Solaris x86	52.16 MB	jre-7u71-solaris-i586.tar.gz
Solaris x64	16.14 MB	jre-7u71-solaris-x64.tar.gz
Solaris SPARC	54.95 MB	jre-7u71-solaris-sparc.tar.gz
Solaris SPARC 64-bit	18.1 MB	jre-7u71-solaris-sparcv9.tar.gz
Windows x86 Online	0.89 MB	jre-7u71-windows-i586-iftw.exe
Windows x86 Offline	28.09 MB	jre-7u71-windows-i586.exe
Windows x86	40 MB	jre-7u71-windows-i586.tar.gz
Windows x64	29.59 MB	jre-7u71-windows-x64.exe
Windows x64	41.71 MB	jre-7u71-windows-x64.tar.gz

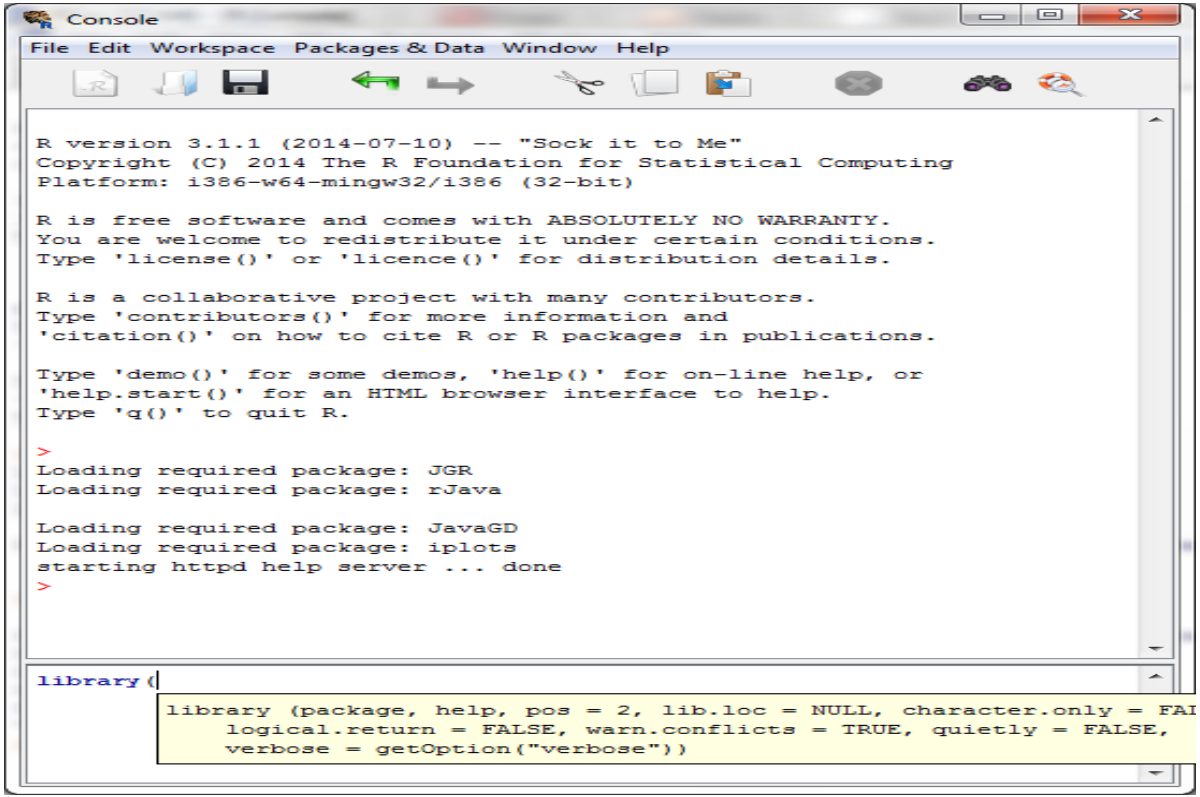
Java EE and Glassfish
Java ME
Java Card
NetBeans IDE
Java Mission Control

Java Resources
Java APIs
Technical Articles
Demos and Videos
Forums
Java Magazine
Java.net
Developer Training
Tutorials
Java.com

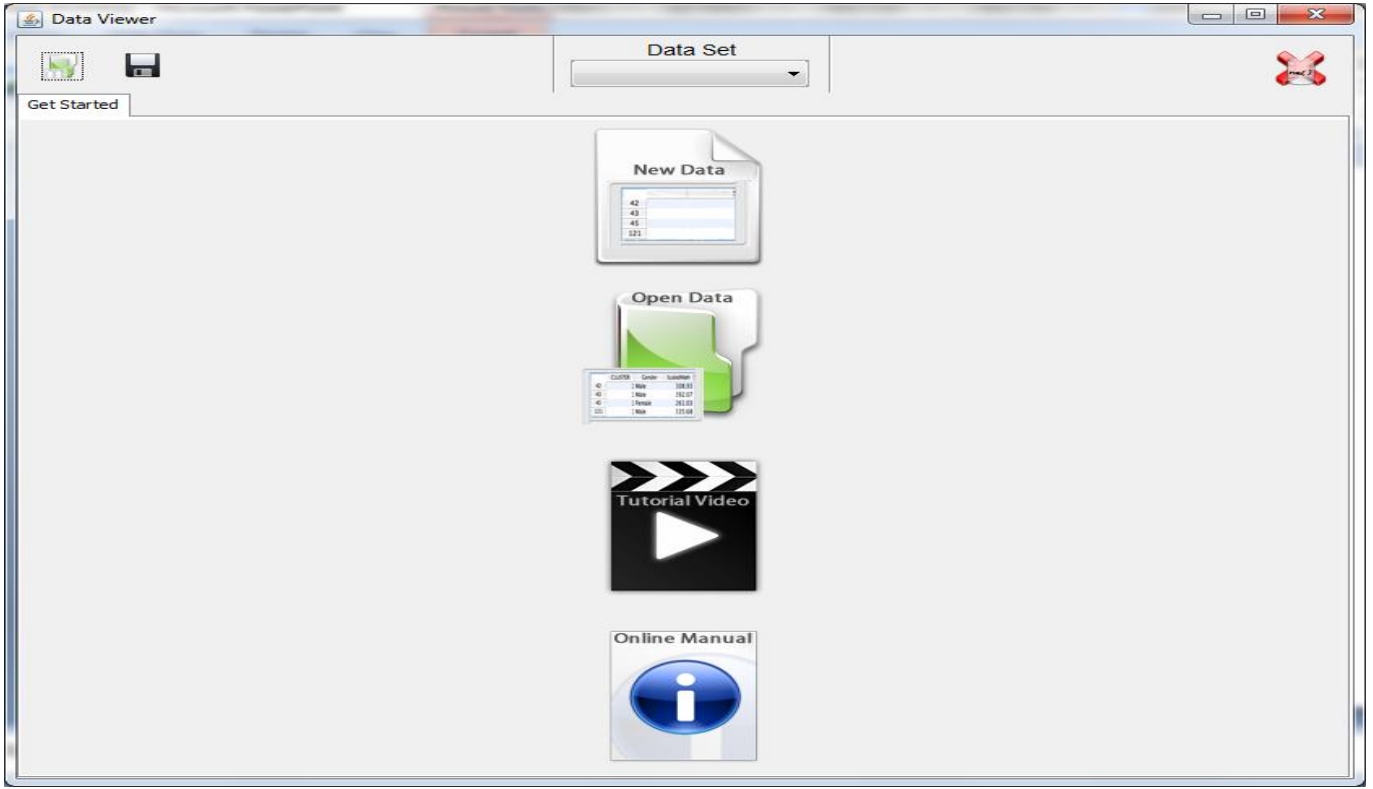
Installation Deducer (with JGR)



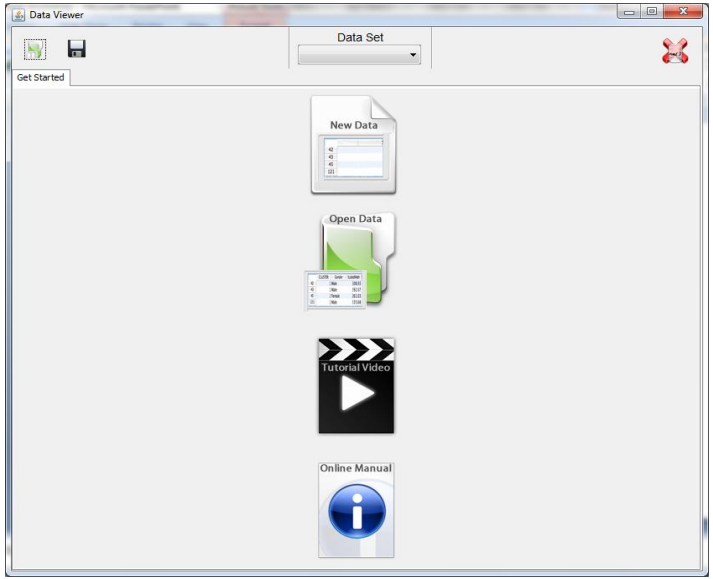
Installation Deducer (with JGR)

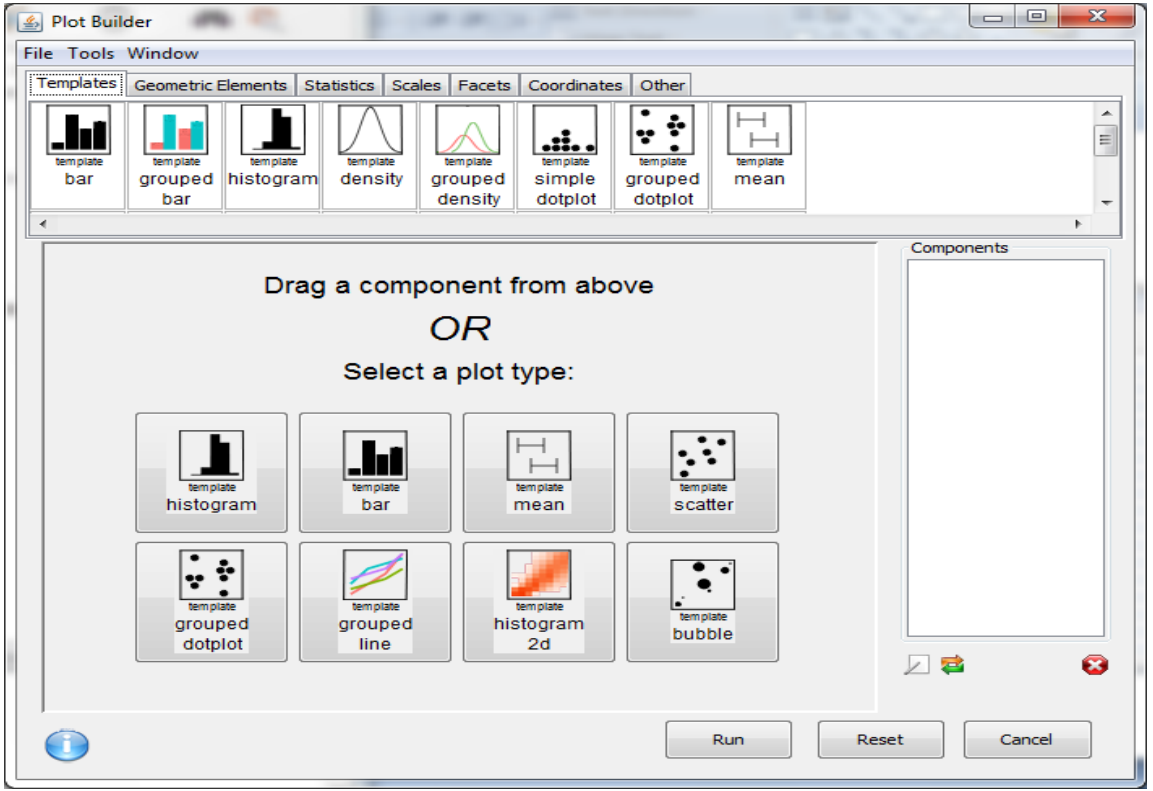


Installation Deducer (with JGR)



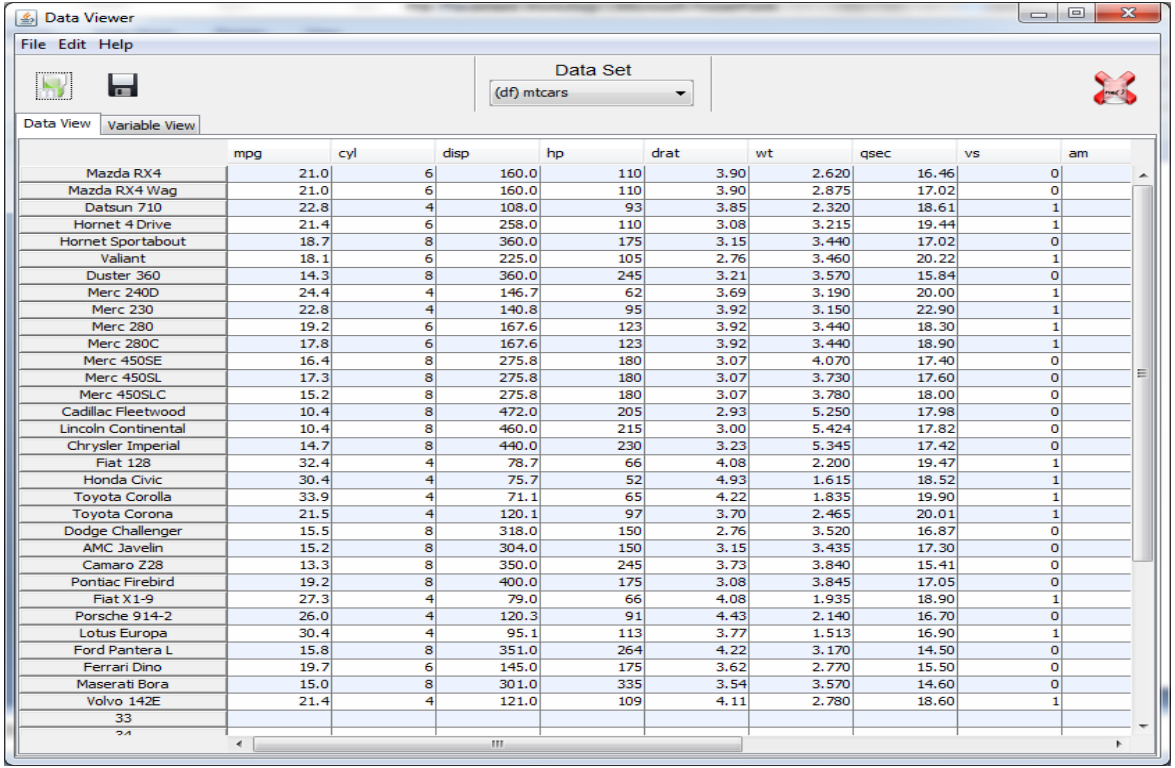
Overview of Deducer (with JGR)





Demo Deducer

- data()
- data(mtcars)



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	
33									

RStudio

RStudio Desktop enables you with following advantages of native R console

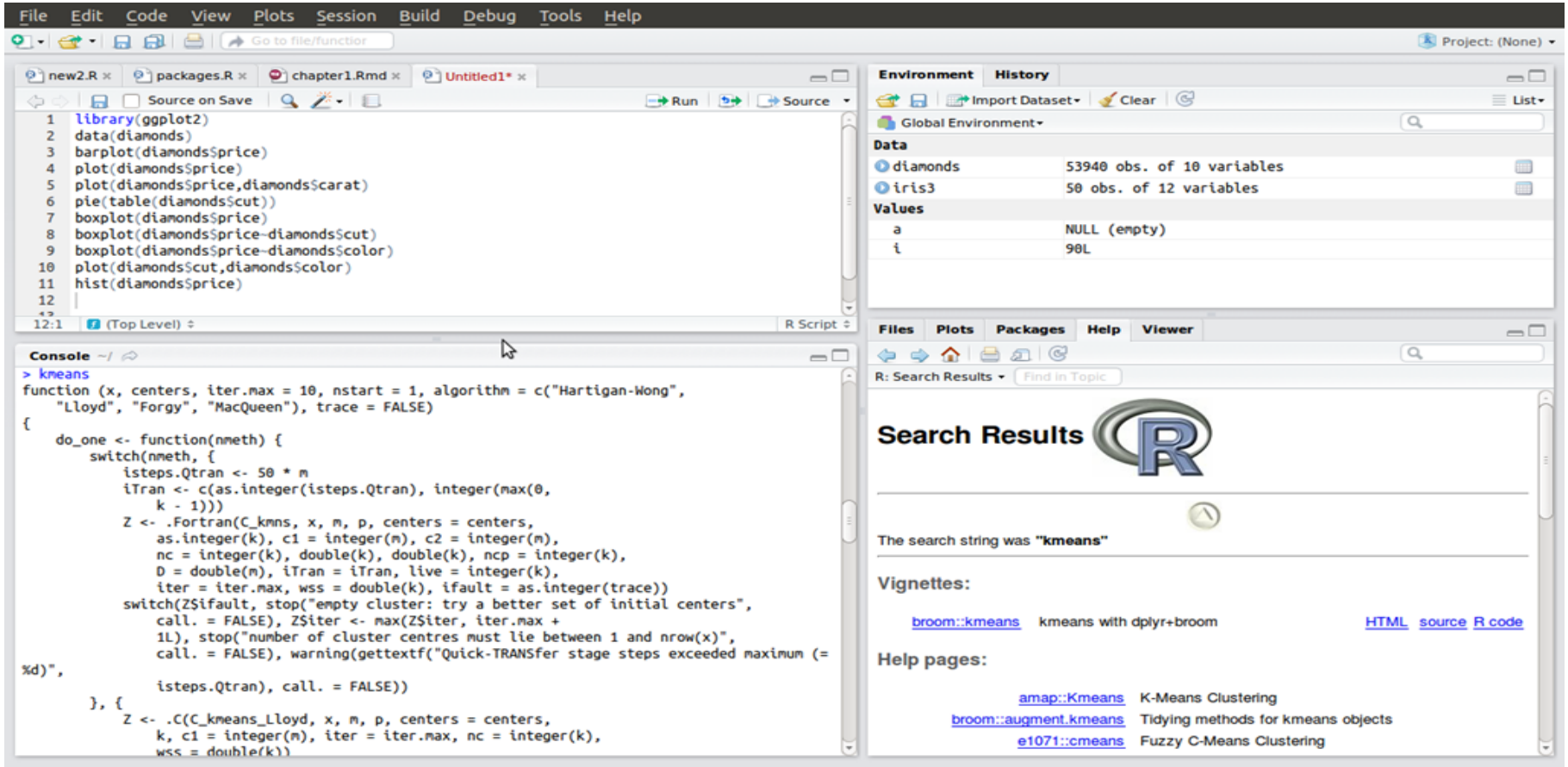
- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions
- Easily manage multiple working directories using projects
- Integrated R help and documentation
- Interactive debugger to diagnose and fix errors quickly
- Extensive package development tools

http://www.rstudio.com/products/

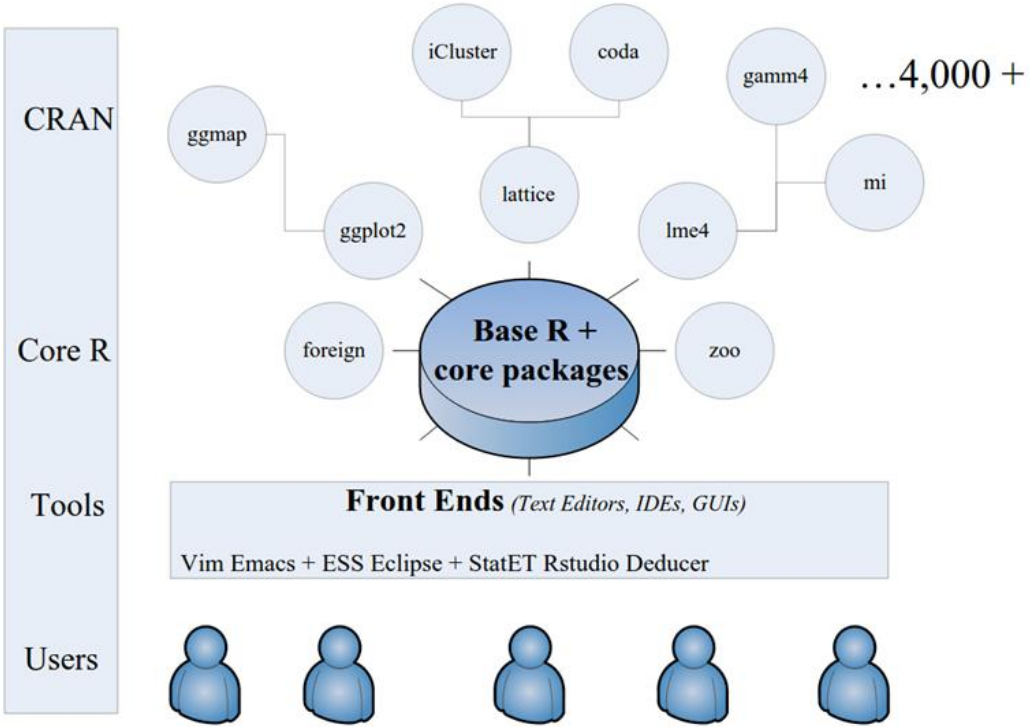
RStudio

RStudio Server enables you to provide a browser based interface (the RStudio IDE) to a version of R running on a remote Linux server. Deploying R and RStudio on a server has a number of benefits, including:

- The ability to access your R workspace from any computer in any location;
- Easy sharing of code, data, and other files with colleagues;
- Allowing multiple users to share access to the more powerful compute resources (memory, processors, etc.) available on a well equipped server; and
- Centralized installation and configuration of R, R packages, TeX, and other supporting libraries.


File Edit Code View Plots Session Build Debug Tools Help
Project: (None)
new2.R
packages.R
chapter1.Rmd
Untitled1*
Source on Save
Run
Source
1 library(ggplot2)
2 data(diamonds)
3 barplot(diamonds$price)
4 plot(diamonds$price)
5 plot(diamonds$price,diamonds$carat)
6 pie(table(diamonds$cut))
7 boxplot(diamonds$price)
8 boxplot(diamonds$price~diamonds$cut)
9 boxplot(diamonds$price~diamonds$color)
10 plot(diamonds$cut,diamonds$color)
11 hist(diamonds$price)
12
12:1 (Top Level) R Script
Console ~/
> kmeans
function (x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong",
"Lloyd", "Forgy", "MacQueen"), trace = FALSE)
{
do_one <- function(nmeth) {
switch(nmeth, {
isteps.Qtran <- 50 * m
iTran <- c(as.integer(isteps.Qtran), integer(max(0,
k - 1)))
Z <- .Fortran(C_kmns, x, m, p, centers = centers,
as.integer(k), c1 = integer(m), c2 = integer(m),
nc = integer(k), double(k), double(k), ncp = integer(k),
D = double(m), iTran = iTran, live = integer(k),
iter = iter.max, wss = double(k), ifault = as.integer(trace))
switch(Z$ifault, stop("empty cluster: try a better set of initial centers",
call. = FALSE), Z$iter <- max(Z$iter, iter.max +
1L), stop("number of cluster centres must lie between 1 and nrow(x)",
call. = FALSE), warning(gettextf("Quick-TRANSfer stage steps exceeded maximum (=
%d)",
isteps.Qtran), call. = FALSE))
}, {
Z <- .C(C_kmeans_Lloyd, x, m, p, centers = centers,
k, c1 = integer(m), iter = iter.max, nc = integer(k),
wss = double(k))
Environment History
Import Dataset
Clear
List
Global Environment
Data
diamonds 53940 obs. of 10 variables
iris3 50 obs. of 12 variables
Values
a NULL (empty)
i 90L
Files Plots Packages Help Viewer
R: Search Results
Find in Topic
Search Results
The search string was "kmeans"
Vignettes:
broom::kmeans kmeans with dplyr+broom HTML source R code
Help pages:
amap::Kmeans K-Means Clustering
broom::augment.kmeans Tidying methods for kmeans objects
e1071::cmeans Fuzzy C-Means Clustering

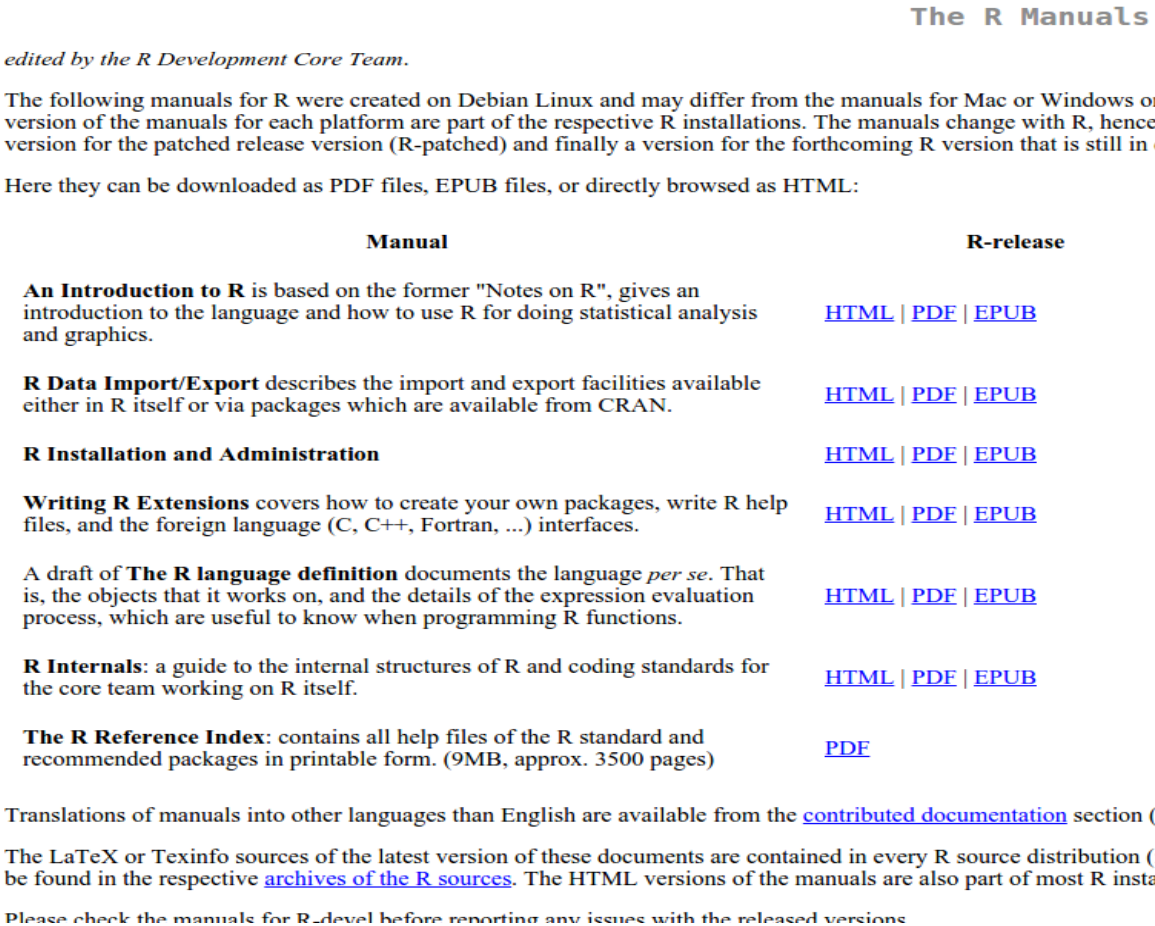
R Landscape


CRAN
Core R
Tools
Users
iCluster
coda
gamm4
…4,000 +
ggmap
lattice
mi
ggplot2
lme4
foreign
Base R +
core packages
zoo
Front Ends (Text Editors, IDEs, GUIs)
Vim Emacs + ESS Eclipse + StatET Rstudio Deducer

R Documentation

Manuals

http://cran.r-project.org/manuals.html



The R Manuals

edited by the R Development Core Team.

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platf
version of the manuals for each platform are part of the respective R installations. The manuals change with R, hence we pr
version for the patched release version (R-patched) and finally a version for the forthcoming R version that is still in develo

Here they can be downloaded as PDF files, EPUB files, or directly browsed as HTML:

Manual	R-release
An Introduction to R is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.	HTML | PDF | EPUB
R Data Import/Export describes the import and export facilities available either in R itself or via packages which are available from CRAN.	HTML | PDF | EPUB
R Installation and Administration	HTML | PDF | EPUB
Writing R Extensions covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.	HTML | PDF | EPUB
A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions.	HTML | PDF | EPUB
R Internals: a guide to the internal structures of R and coding standards for the core team working on R itself.	HTML | PDF | EPUB
The R Reference Index: contains all help files of the R standard and recommended packages in printable form. (9MB, approx. 3500 pages)	PDF

Translations of manuals into other languages than English are available from the contributed documentation section (only a

The LaTeX or Texinfo sources of the latest version of these documents are contained in every R source distribution (in the
be found in the respective archives of the R sources. The HTML versions of the manuals are also part of most R installation

Please check the manuals for R-devel before reporting any issues with the released versions

R Documentation

- Vignettes

ggplot2: An Implementation of the Grammar of Graphics

An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: by step from multiple data sources. It also implements a sophisticated multidimensional conditioning system and a information, documentation and examples.

Version:	1.0.1
Depends:	R (≥ 2.14), stats, methods
Imports:	plyr (≥ 1.7.1), digest, grid, gtable (≥ 0.1.1), reshape2, scales (≥ 0.2.3), proto, MASS
Suggests:	quantreg, Hmisc, mapproj, maps, hexbin, maptools, multcomp, nlme, testthat, knitr, mgcv
Enhances:	sp
Published:	2015-03-17
Author:	Hadley Wickham [aut, cre], Winston Chang [aut]
Maintainer:	Hadley Wickham <h.wickham at gmail.com>
BugReports:	https://github.com/hadley/ggplot2/issues
License:	GPL-2
URL:	http://ggplot2.org, https://github.com/hadley/ggplot2
NeedsCompilation:	no
Citation:	ggplot2 citation info
Materials:	README NEWS
In views:	Graphics, Phylogenetics
CRAN checks:	ggplot2 results

Downloads:

Reference manual:	ggplot2.pdf
Vignettes:	Contributing to ggplot2 development ggplot2 release process
Package source:	ggplot2_1.0.1.tar.gz
Windows binaries:	r-devel: ggplot2_1.0.1.zip, r-release: ggplot2_1.0.1.zip, r-oldrel: ggplot2_1.0.1.zip
OS X Snow Leopard binaries:	r-release: not available, r-oldrel: ggplot2_1.0.1.tgz
OS X Mavericks binaries:	r-release: ggplot2_1.0.1.tgz
Old sources:	ggplot2 archive

Reverse dependencies:

Reverse depends: alphahull, AmpliconDuo, aoristic, apsimr, bcrm, bde, benchmark, biomod2, bootnet, brms, car

CRAN Views

http://cran.r-project.org/web/views/

CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

R Community

- email groups http://www.r-project.org/mail.html

R-announce

R-help

R-package-devel

R-devel

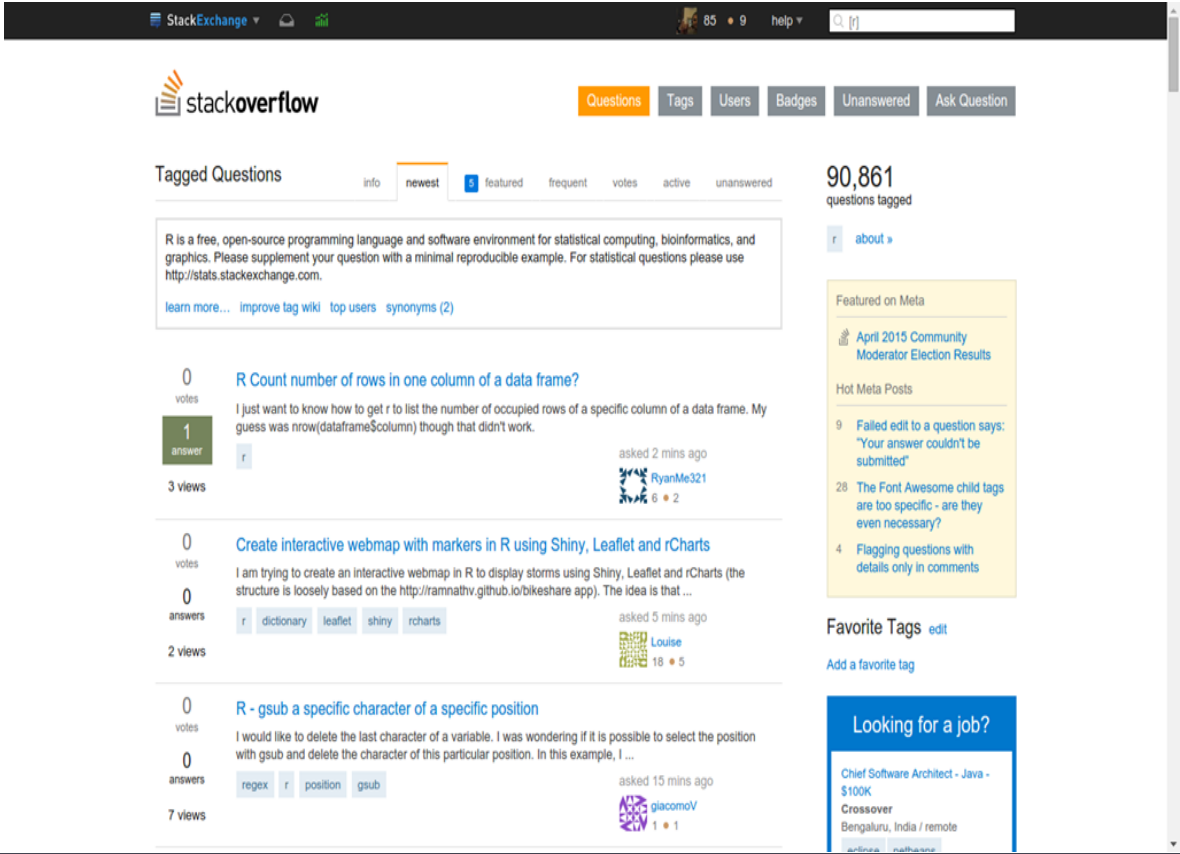
R-packages

Special Interest Groups

- Stack Overflow [r]
- Twitter #rstats
- Blogs at http://www.r-bloggers.com/ (573 blogs)

Stack Overflow

http://stackoverflow.com/questions/tagged/r



Twitter

https://twitter.com/search?q=rstats&src=sprv

Help within R

?”keyword”

??”keyword”

Example-

> ?kmeans

> ??kmeans

Introductory R

```
> Sys.Date()

[1] "2015-05-10„

> Sys.time()

[1] "2015-05-10 18:28:32 IST"
```

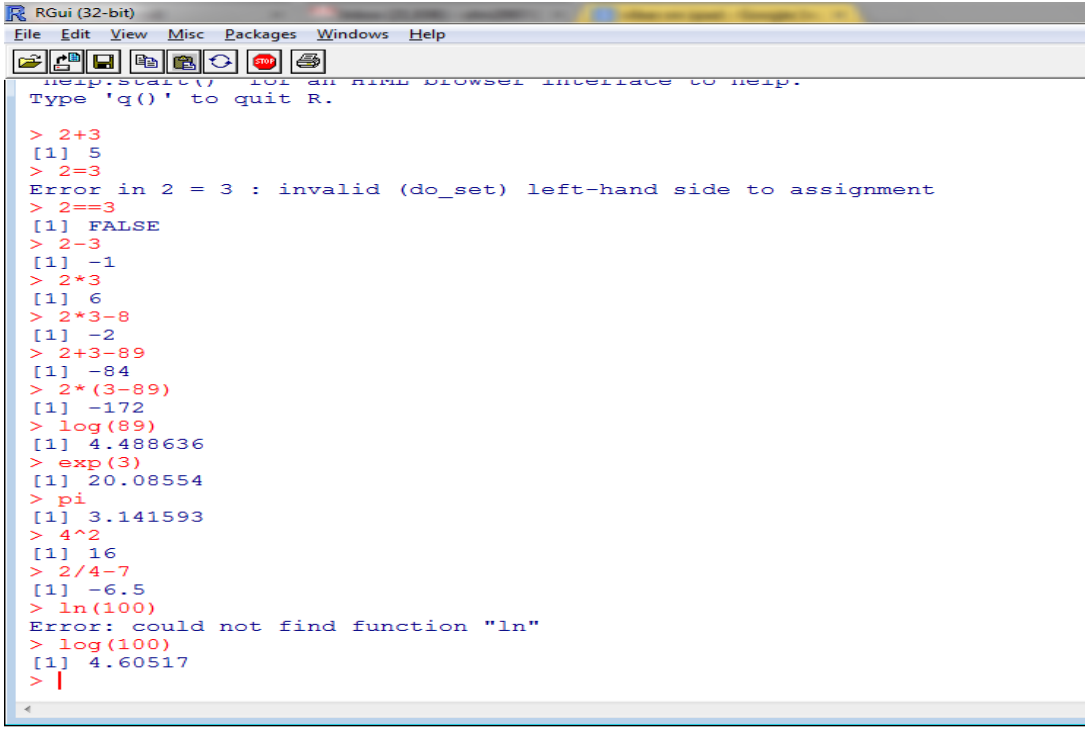
R as a Calculator
Basic Math on R Console

- +
- -
- Log
- Exp
- *
- /
- ()

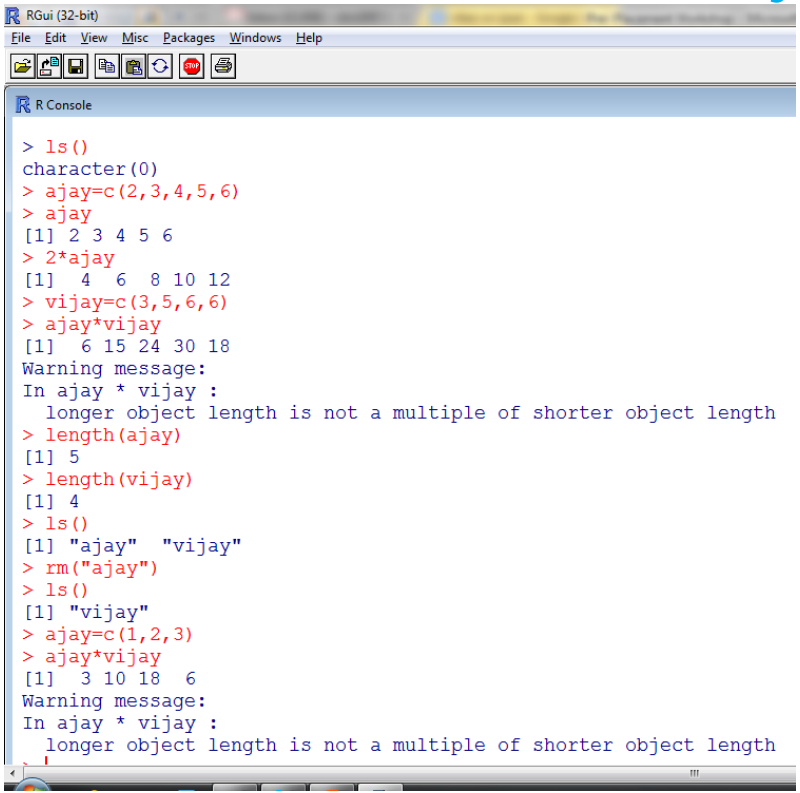
- mean
- sum
- sd
- log
- median
- exp

Demo-
Basic Math on R Console

Hint- Ctrl +L clears screen



Demo-
Basic Objects on R Console



Functions-
ls() – what objects are here
rm("foo") removes object named foo

Assignment
Using = or -> assigns object names to values

Hint- Up arrow gives you last typed command

Functions and Loops

- Loops

for (number in 1:5){ print (number) }

```
> for (number in 1:5){ print (number) }
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
> for (i in 1:5){ print (i) }
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
> for (i in 1:5){ rnorm(i,10,10) }
> for (i in 1:5){ print(rnorm(i,10,10)) }
[1] 1.090406
[1]  8.611727 16.670168
[1] 10.84623 13.13938 11.56230
[1]   6.068250 -18.723389  33.174107  -1.320091
[1] 13.939702 -9.037375 13.755986  9.459680  9.625309
>
```

Functions and Loops

- Function

functionajay=function(a)(a^2+2*a+1)

```
> functionajay=function(a)(a^2+2*a+1)
> functionajay(1)
[1] 4
> functionajay(2)
[1] 9
> for (i in 1:5){ print(rnorm(i) }
Error: unexpected '}' in "for (i in 1:5){ print(rnorm(i) }"
>
> for (i in 1:5){ print(functionajay(i)) }
[1] 4
[1] 9
[1] 16
[1] 25
[1] 36
>
```

Hint: Always match brackets

Each (deserves a)

Each { deserves a }

Each [deserves a]

Other sources to learn R

swirlstats

http://swirlstats.com/

datacamp

https://www.datacamp.com/

codeschool

http://tryr.codeschool.com/

coursera

https://www.coursera.org/course/compdata

https://www.coursera.org/course/rprog

Good Coding Practices

- Use # for comment
- Use git for version control
- Use Rstudio for multiple lines of code

Functions in R

- custom functions
- source code for a function
- Understanding help ? , ??

```
Console ~/
> kmeans
function (x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong",
    "Lloyd", "Forgy", "MacQueen"), trace = FALSE)
{
    do_one <- function(nmeth) {
        switch(nmeth, {
            isteps.Qtran <- 50 * m
            iTran <- c(as.integer(isteps.Qtran), integer(max(0,
                k - 1)))
            Z <- .Fortran(C_kmns, x, m, p, centers = centers,
                as.integer(k), c1 = integer(m), c2 = integer(m),
                nc = integer(k), double(k), double(k), ncp = integer(k),
                D = double(m), iTran = iTran, live = integer(k),
                iter = iter.max, wss = double(k), ifault = as.integer(trace))
            switch(Z$ifault, stop("empty cluster: try a better set of initial centers",
                call. = FALSE), Z$iter <- max(Z$iter, iter.max +
                1L), stop("number of cluster centres must lie between 1 and nrow(x)",
                call. = FALSE), warning(gettextf("Quick-TRANSfer stage steps exceeded maximum (=
%d)",
                isteps.Qtran), call. = FALSE))
        }, {
            Z <- .C(C_kmeans_Lloyd, x, m, p, centers = centers,
                k, c1 = integer(m), iter = iter.max, nc = integer(k),
                wss = double(k))
        }, {
            Z <- .C(C_kmeans_MacQueen, x, m, p, centers = as.double(centers),
                k, c1 = integer(m), iter = iter.max, nc = integer(k),
                wss = double(k))
        })
        if (m23 <- any(nmeth == c(2L, 3L))) {
            if (any(Z$nc == 0))
                warning("empty cluster: try a better set of initial centers",
                  call. = FALSE)
```

HOMEWORK TIME !

Functions Used in this Lesson

- function(x)
- for
- library
- install.packages
- update.packages
- ls
- rm
- print

Citations and References

```
> citation()
```

To cite R in publications use:

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.


Collabera®
Value. Accelerated.
Collabera