Summer School

INTRODUCTION

Know your Instructor

DECISION STATS
BETTER DECISIONS === FASTER STATS

- 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)

http://linkedin.com/in/ajayohri

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

DECISION STATS
BETTER DECISIONS === FASTER STATS

Name
Education Degree from Institute
Work Ex in Years in Domain

Introduce yourself

DECISION STATS
BETTER DECISIONS === FASTER STATS

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

Basics of Data Science
Basics of Analytics
LTV Analysis

LTV Analysis Quiz

RFM Analysis

RFM Analysis Quiz

Basic Stats
Introduction to Modeling
Introduction to Google Analytics
Blogging

Web Analytics Quiz

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

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

Basics of Data Science

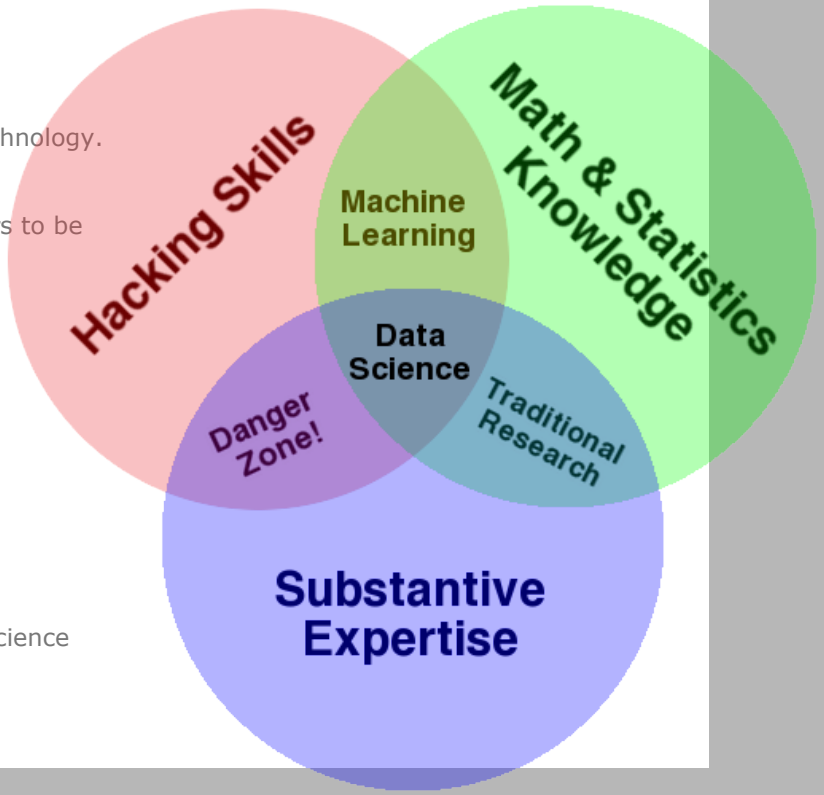
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

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



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

What is Business Analytics

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

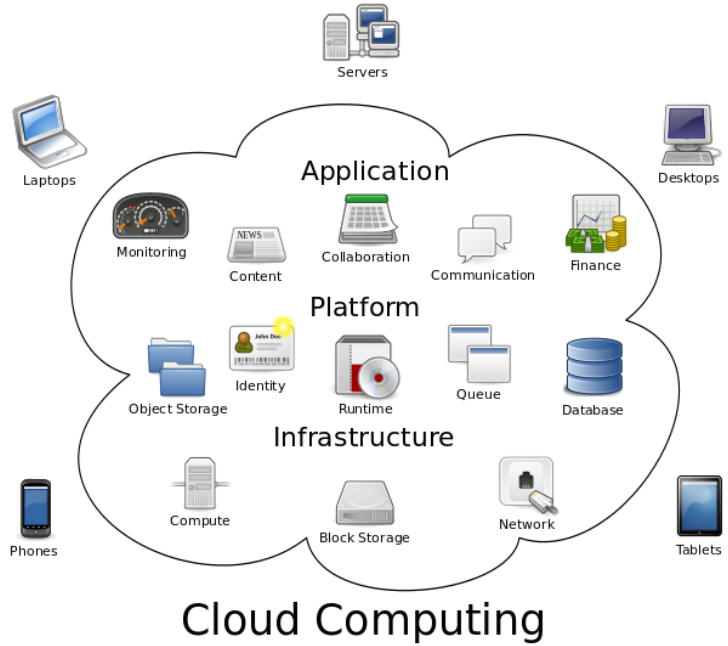
So once again

- 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.



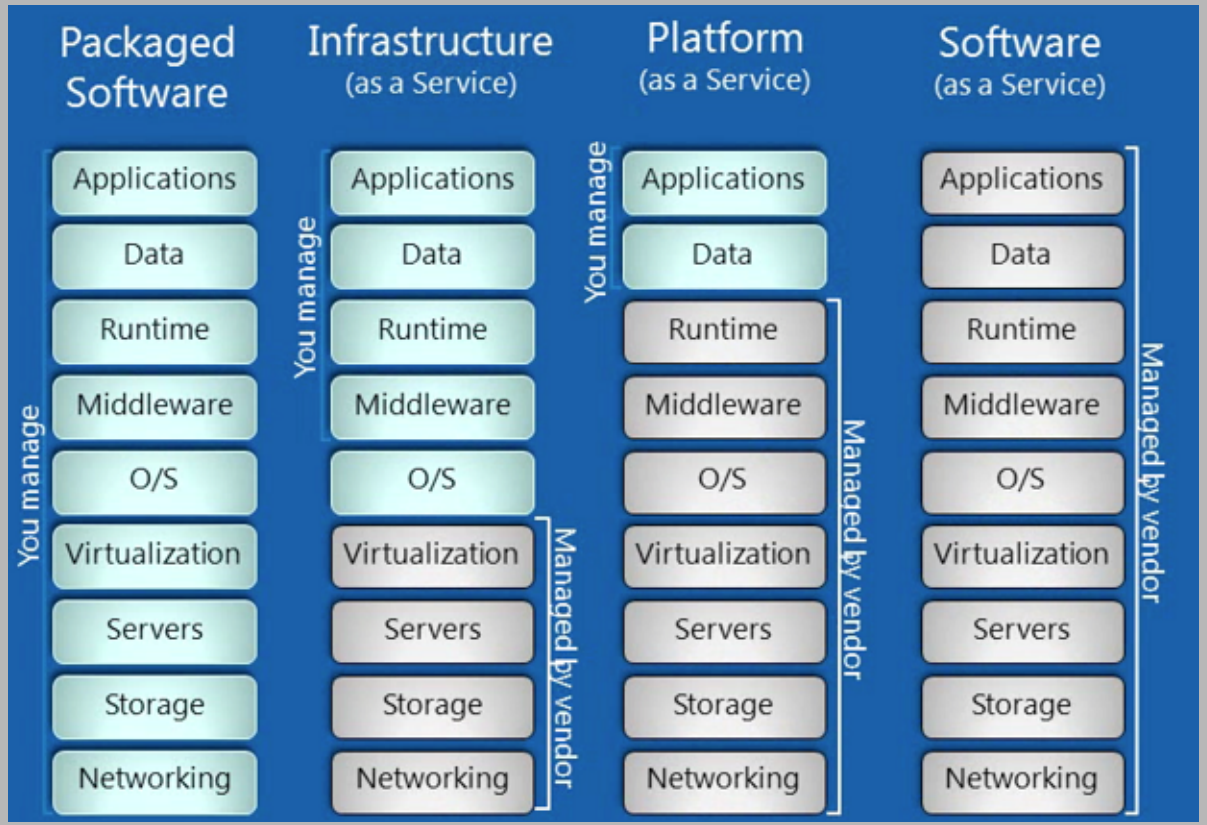
http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

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

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.

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?

LTV Analytics

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

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

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

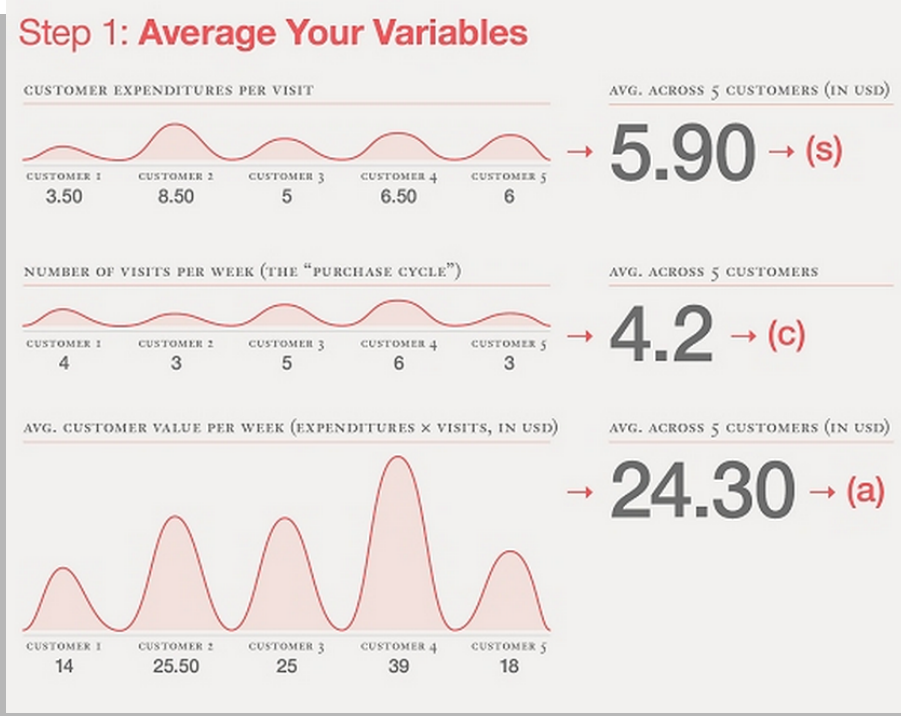
Download the zip file from http://www.kaushik.net/avinash/avinash_ltv.zip

Do the class exercise based on numbers given by instructor

Give a brief supporting statement on analysis

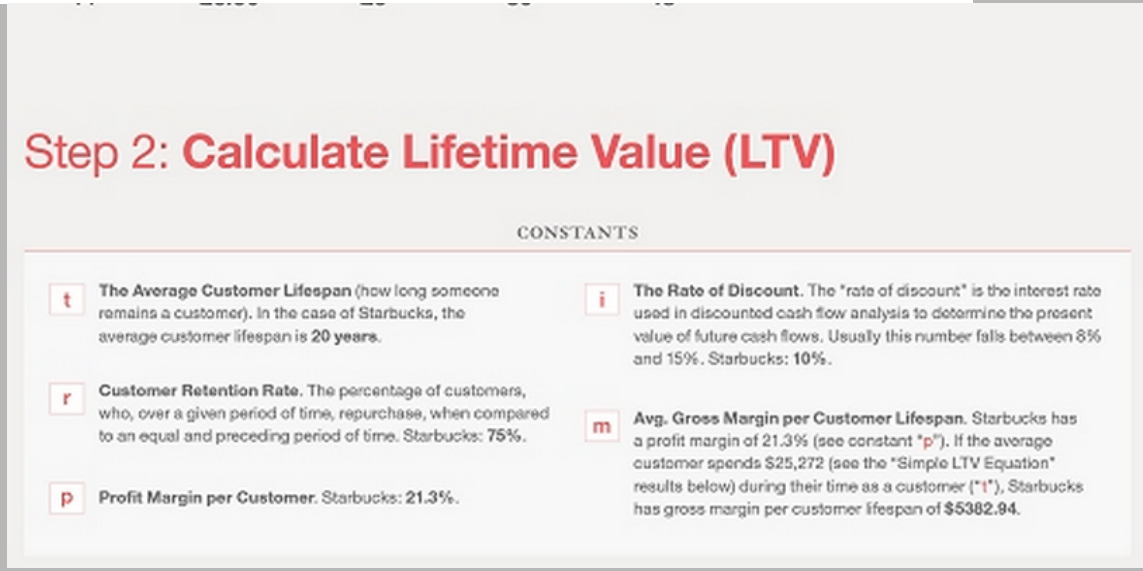
LTV Analytics :Another Approach

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



LTV Analytics

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



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%**.

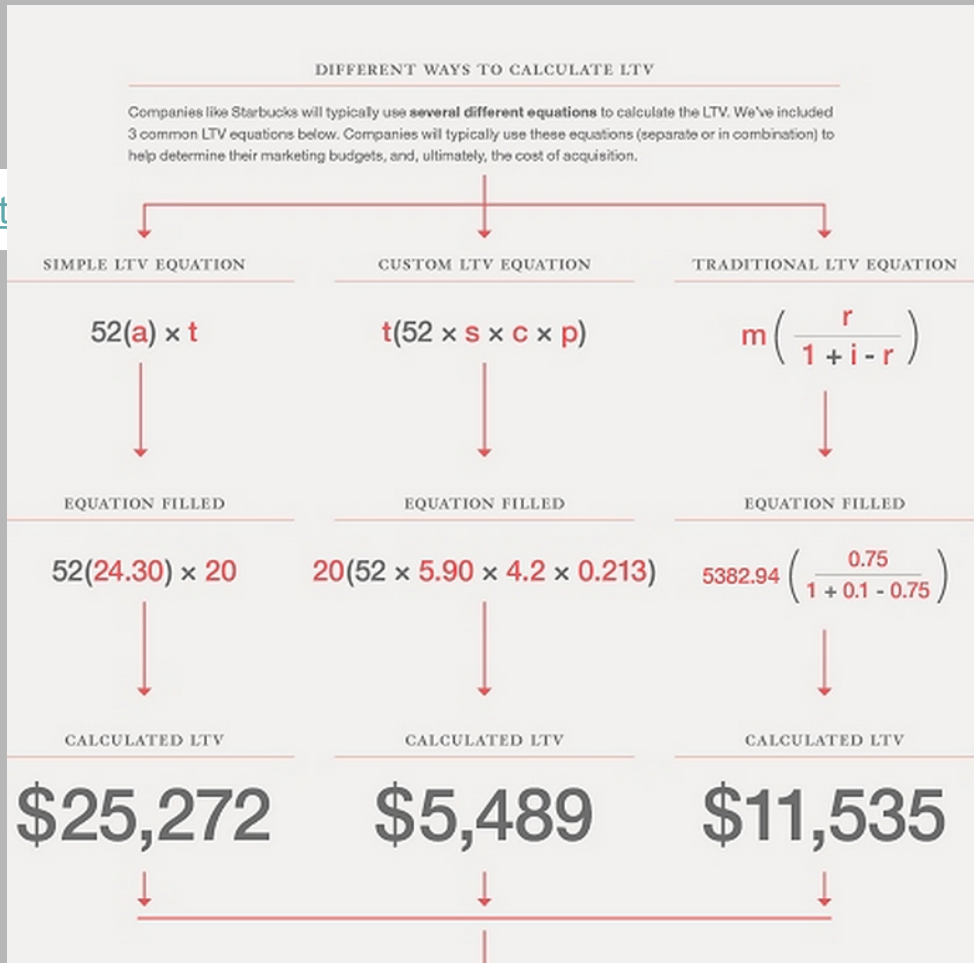
p **Profit Margin per Customer.** Starbucks: **21.3%**.

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**.

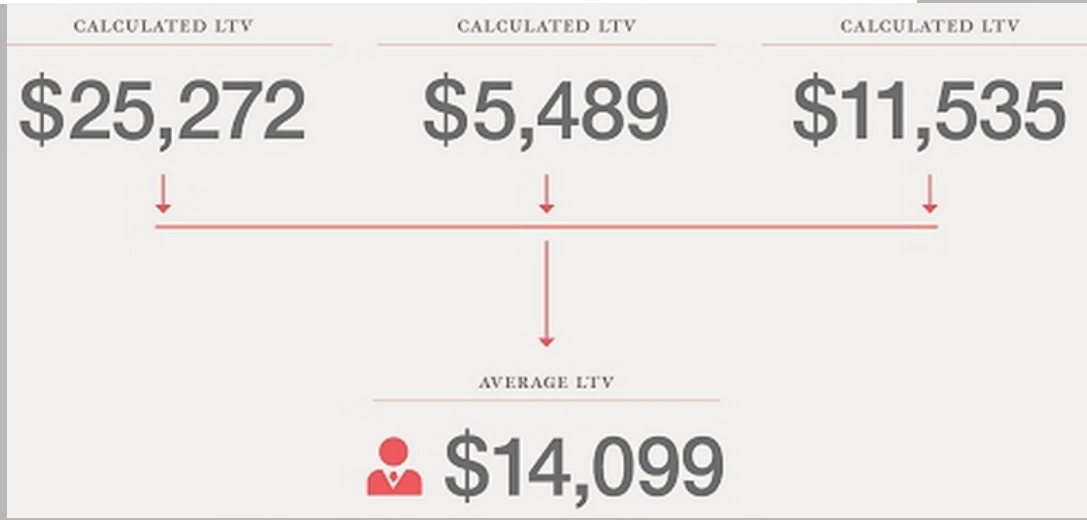
LTV Analytics

https://blog.kissmetrics.com/how-to-calculate-lifet



LTV Analytics

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



Pareto principle

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.

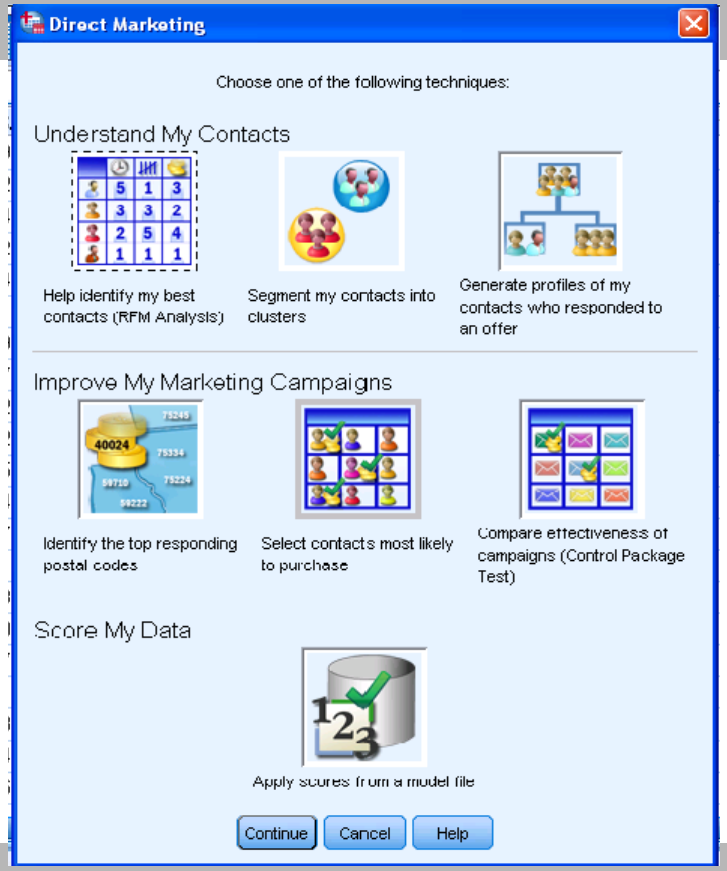
RFM Analysis

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?*

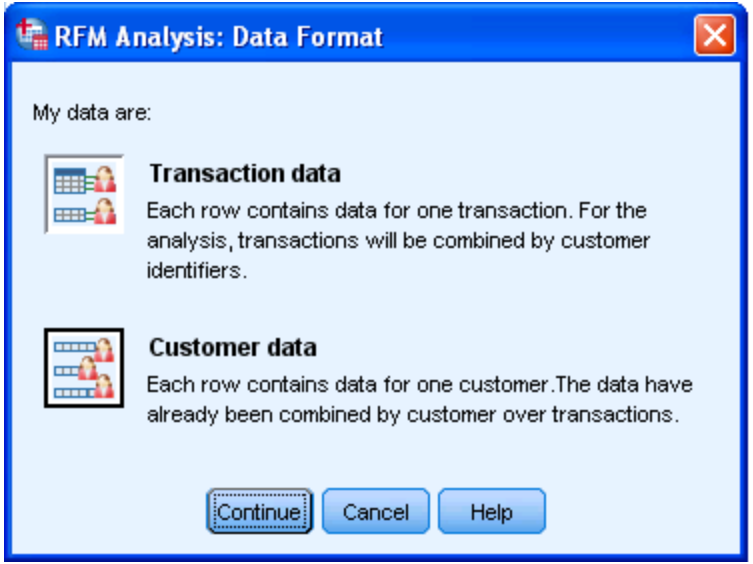
RFM Analysis

Using SPSS 19 - example



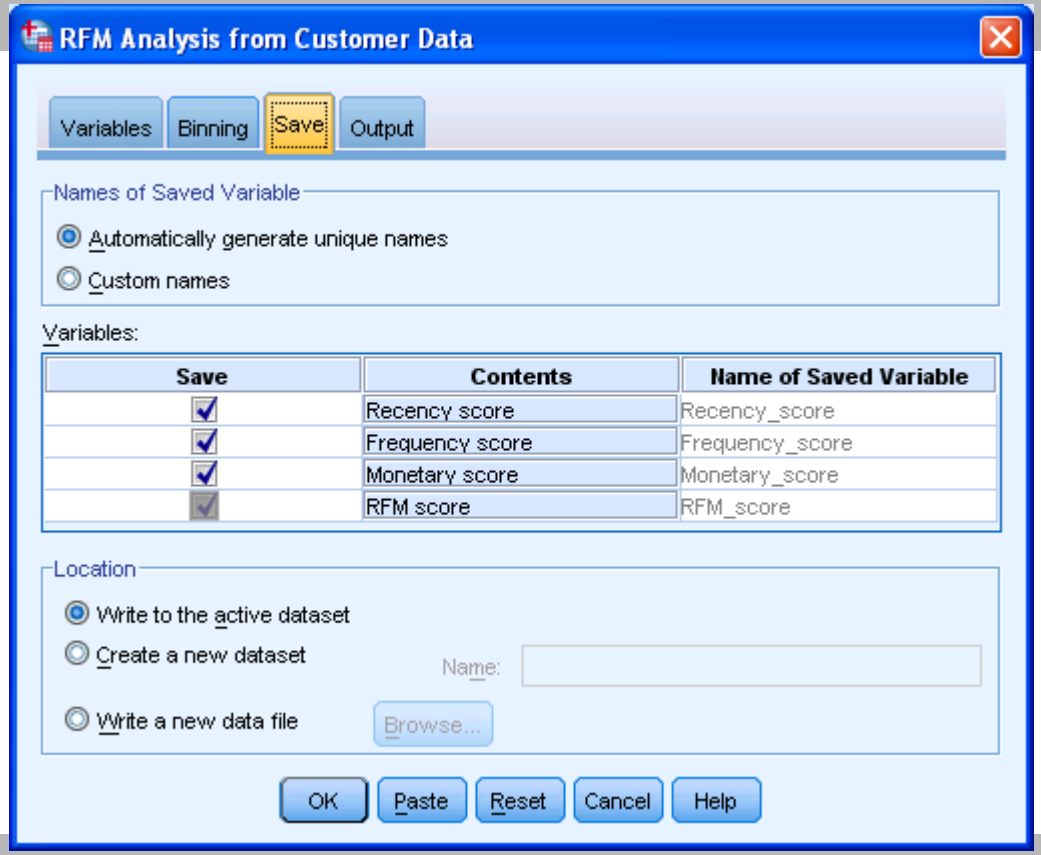
RFM Analysis

Using SPSS 19 - example



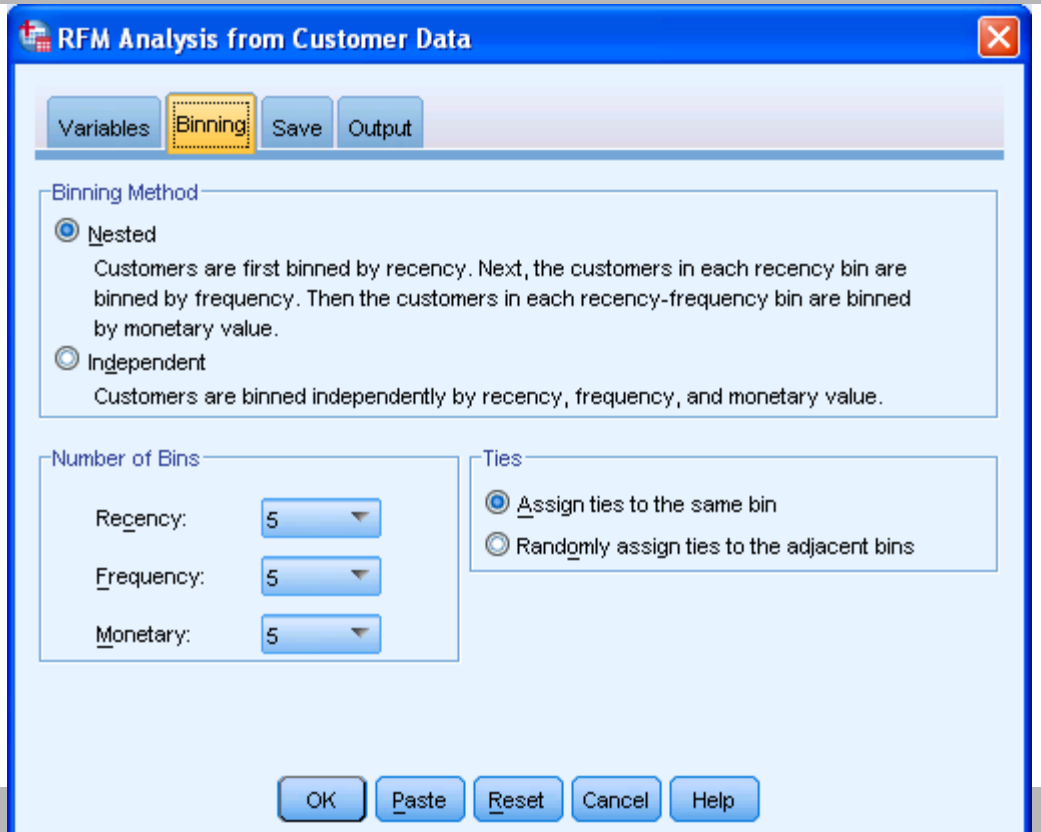
RFM Analysis

Using SPSS 19 - example



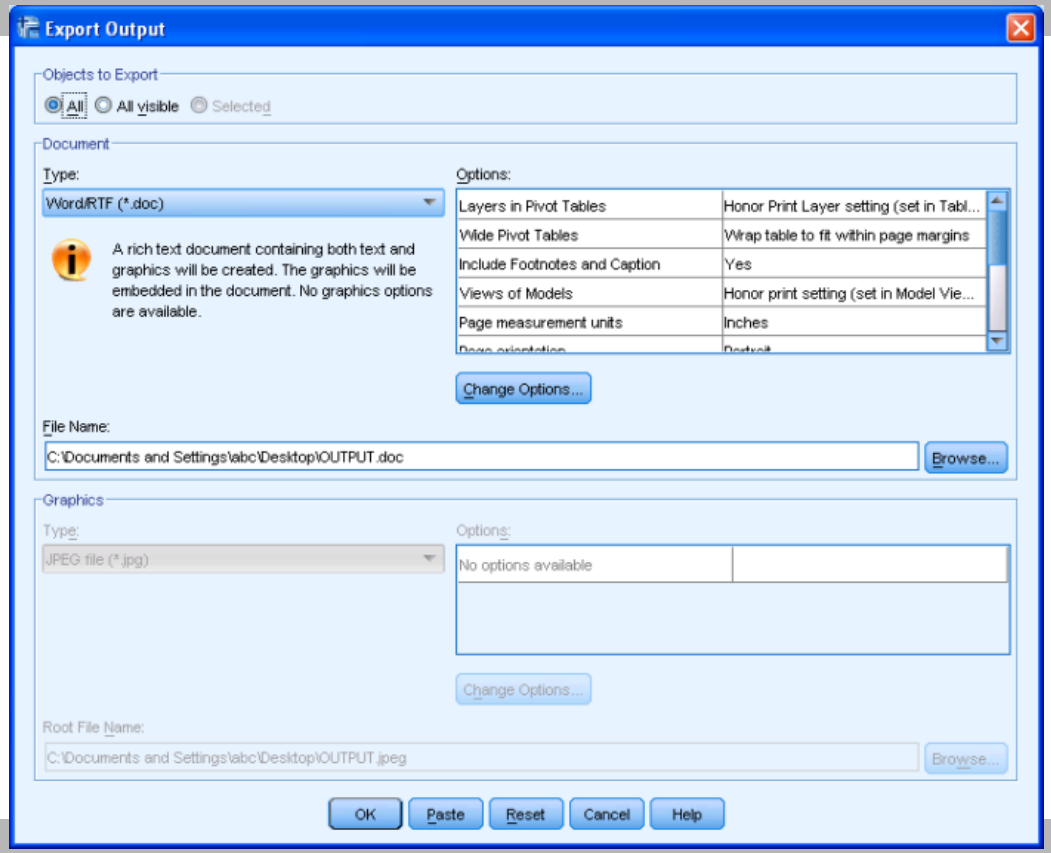
RFM Analysis

Using SPSS 19 - example



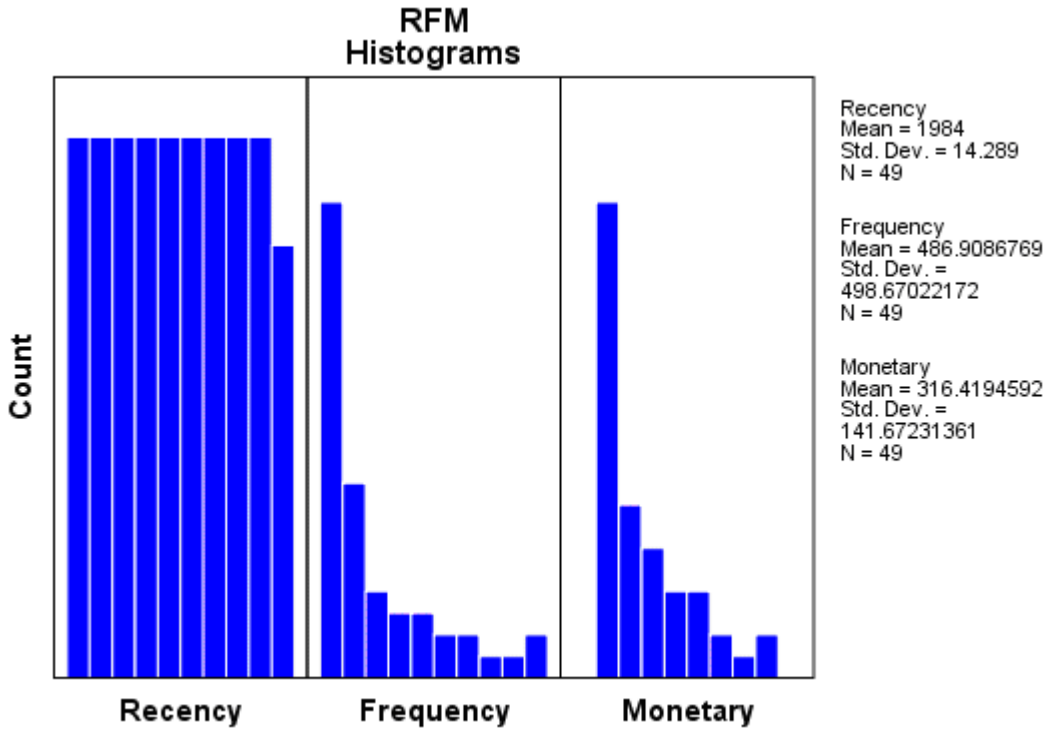
RFM Analysis

Using SPSS 19 - example



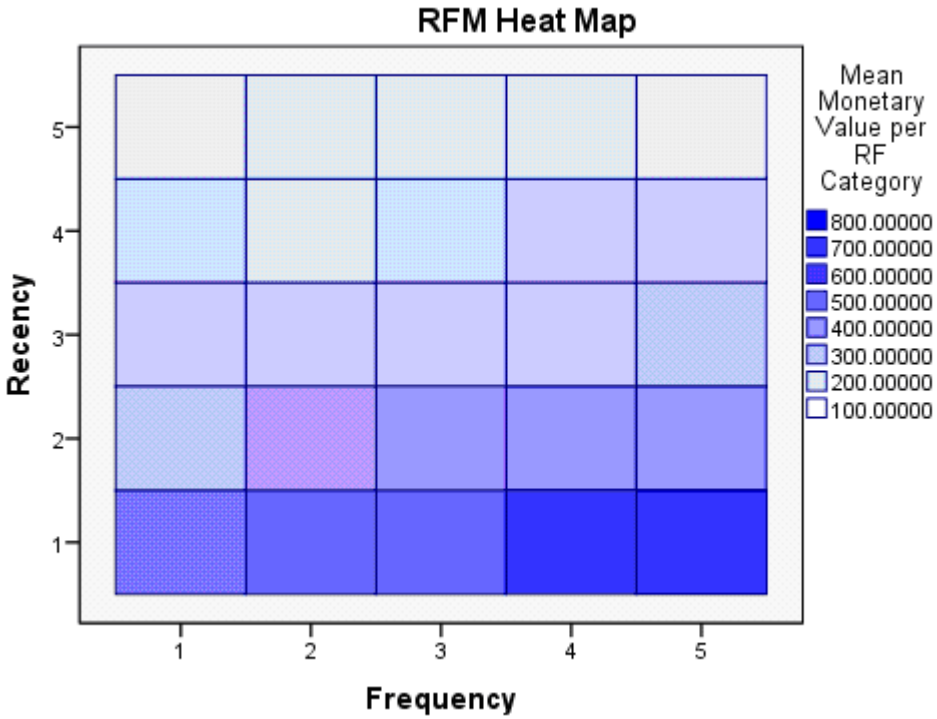
RFM Analysis

Using SPSS 19 - example



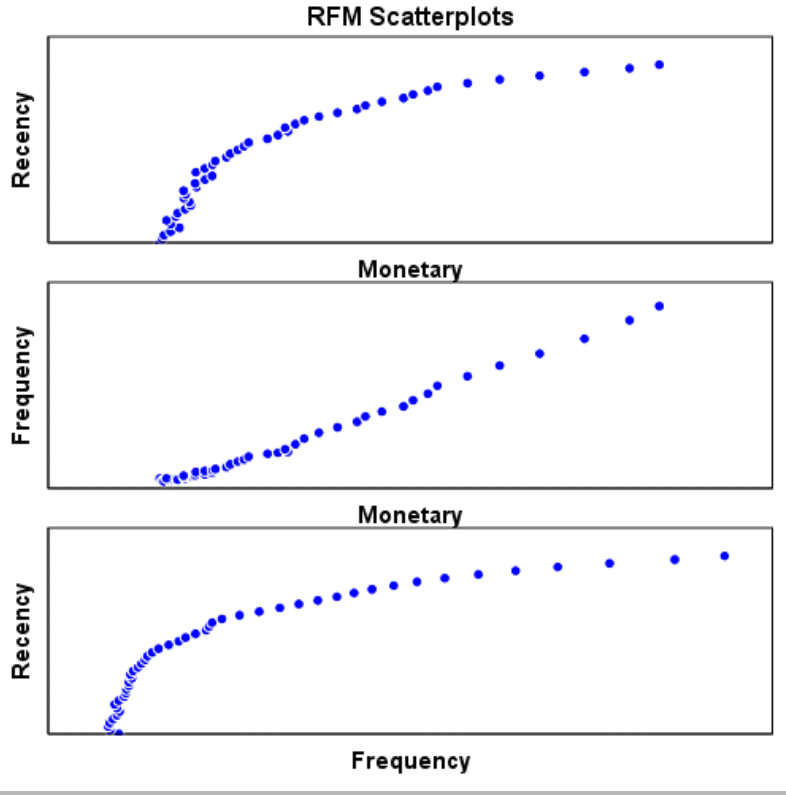
RFM Analysis

Using SPSS 19 - example



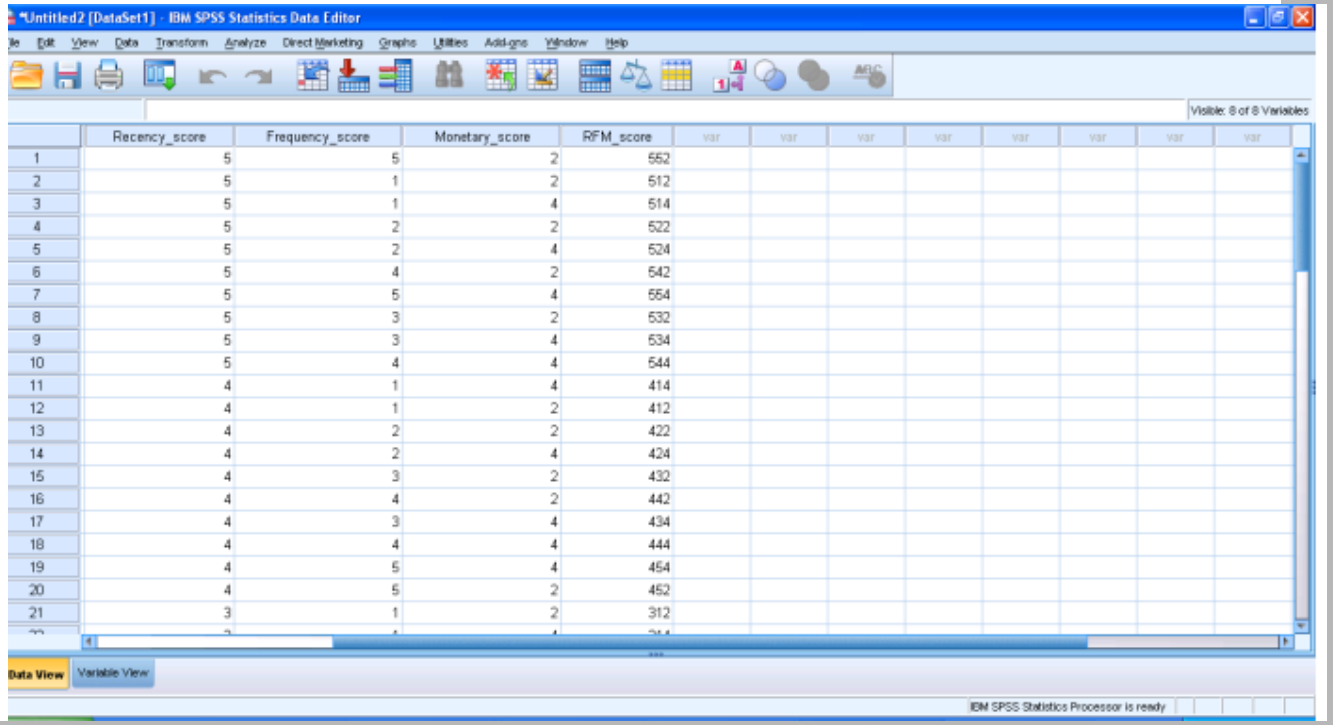
RFM Analysis

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

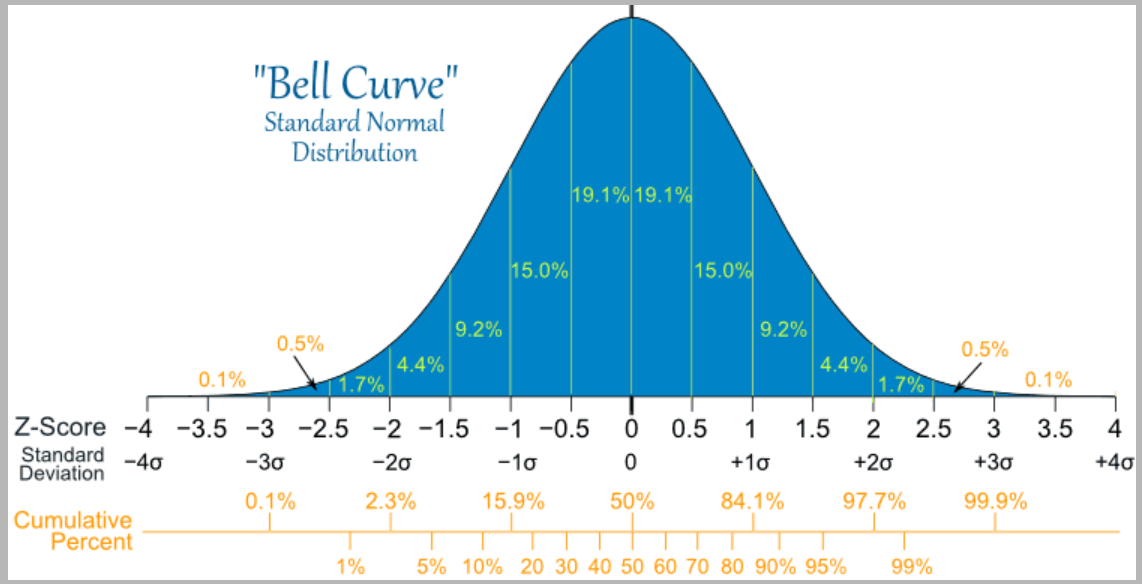
p $q = 1 - p$

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

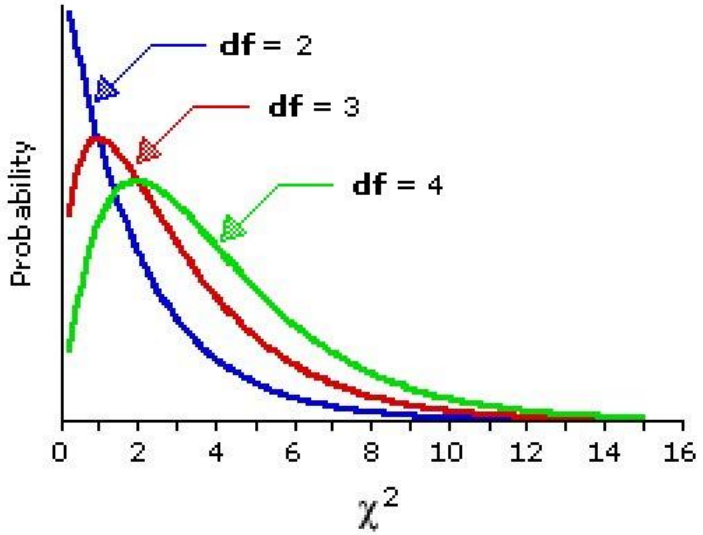
The simplest case of a normal distribution is known as the *standard normal distribution*. This is a special case where μ=0 and σ=1,



Distributions

Chi Square

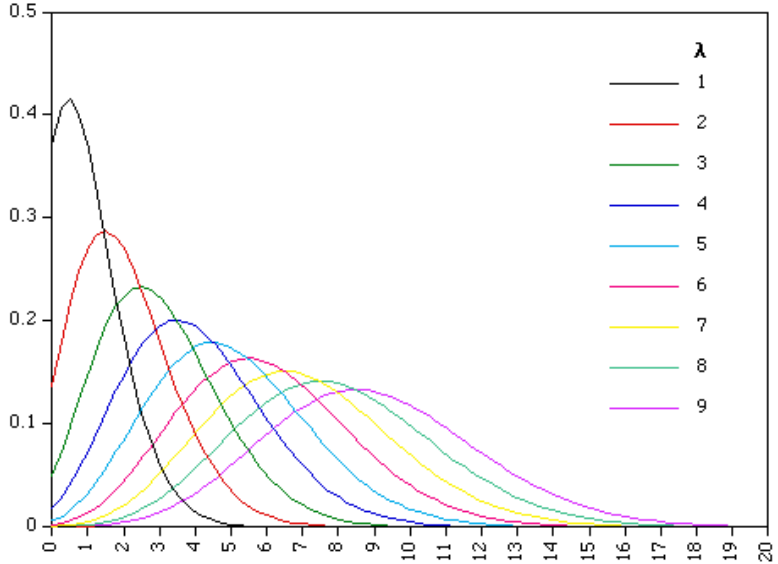
the distribution of a sum of the squares of *k* independent standard normal random variables.



Distributions

Poisson

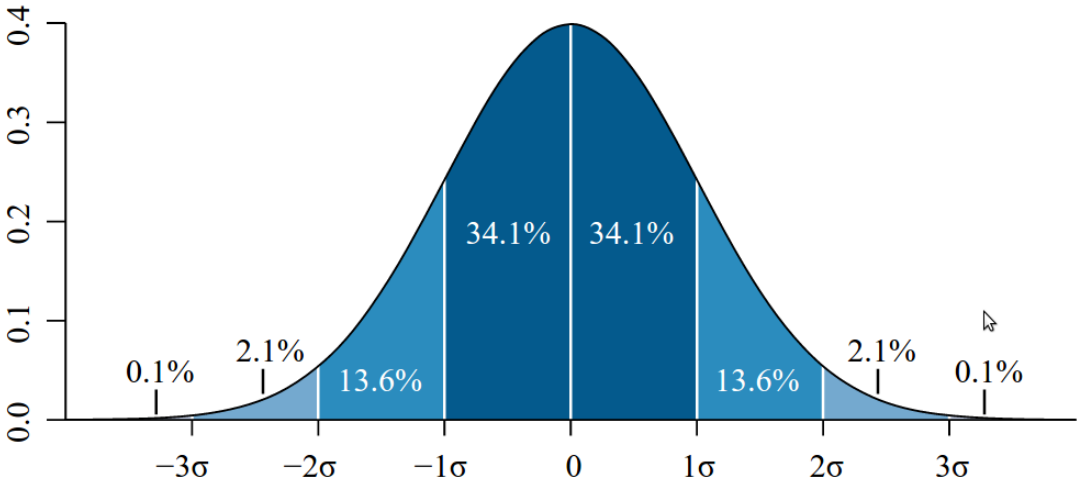
a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event



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.

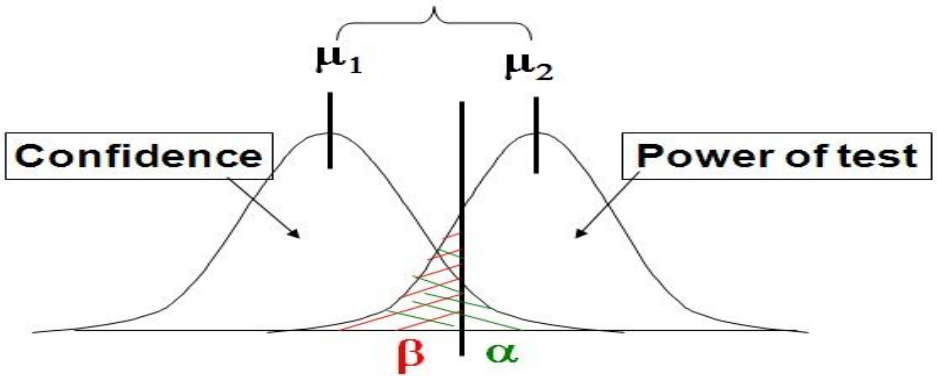
3. Compute the P-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the -value, the stronger the evidence against the null hypothesis.

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

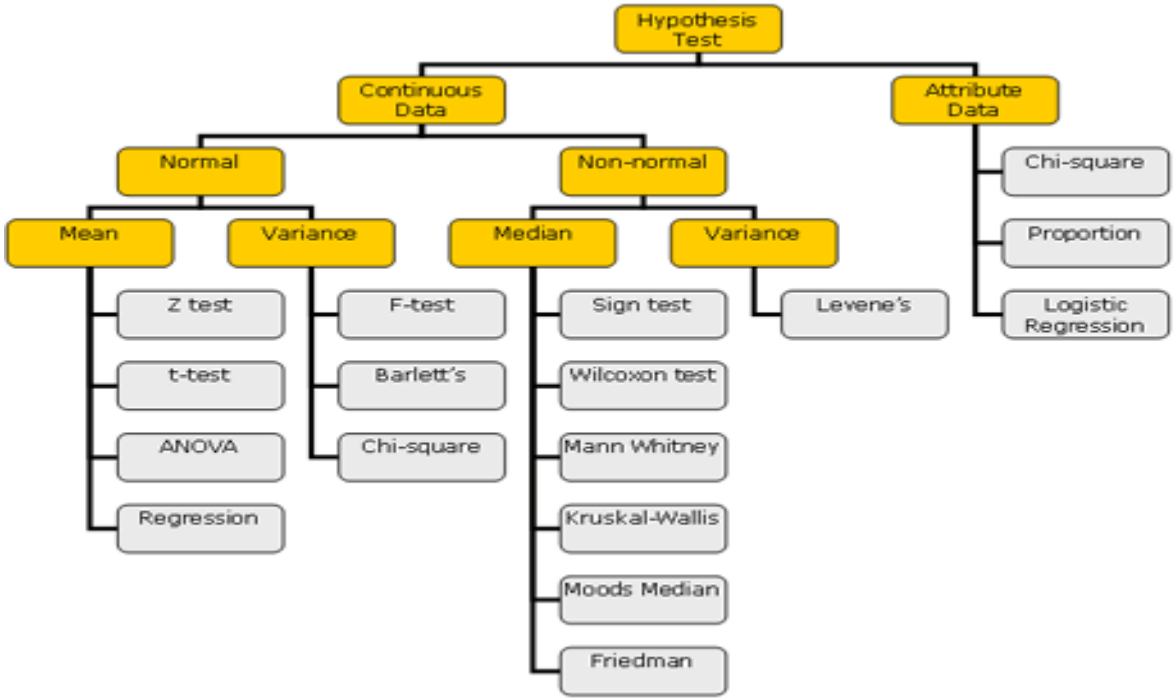
Hypothesis testing

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 Researcher's Decision	The Truth (unknown to the researcher): The Null Hypothesis is True	The Truth (unknown to the researcher): 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

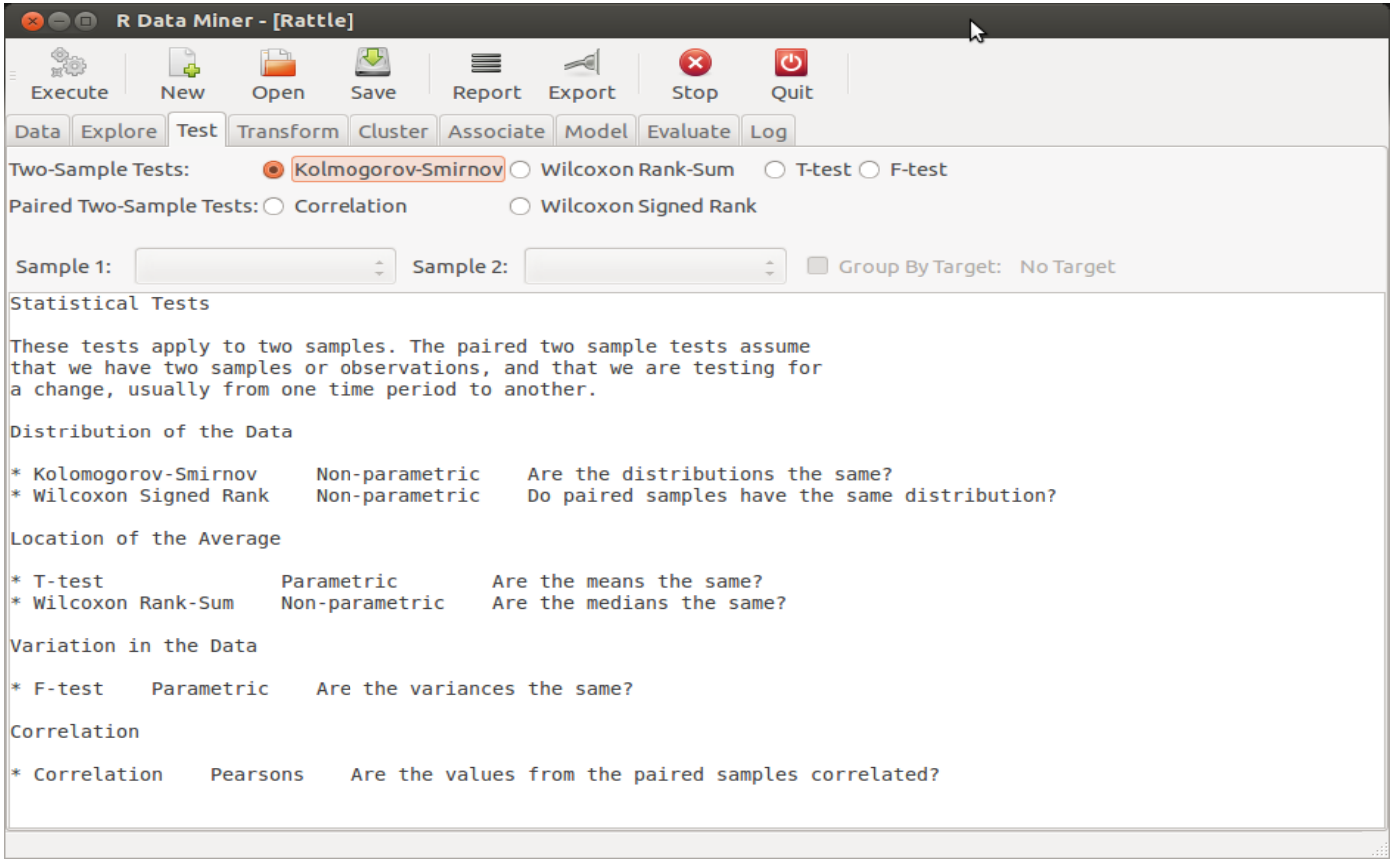
Hypothesis testing



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.	$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
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.	$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$
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	$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
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	$t = \frac{\overline{d}\sqrt{n}}{s_d}$
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 Ma

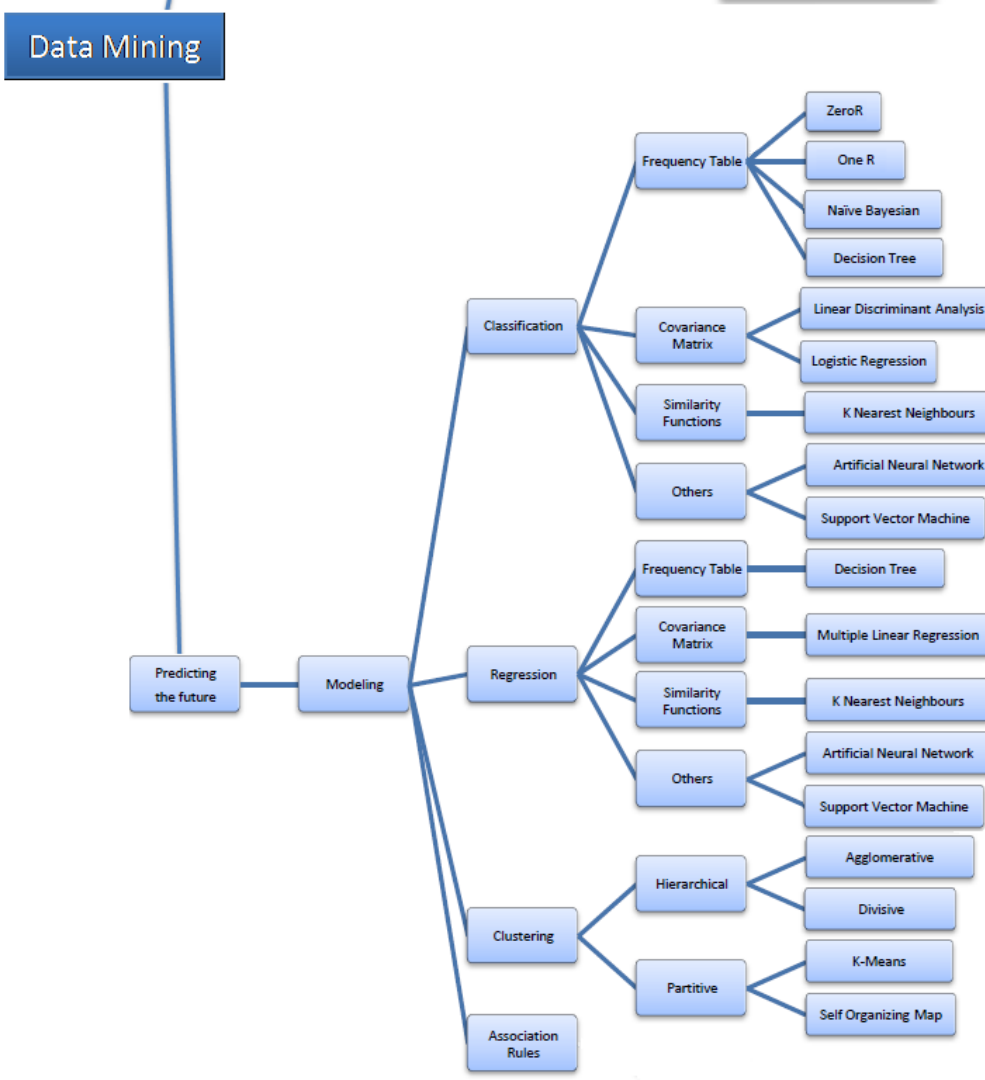
Source-

http://www.saedsayad.com/

Data Mining Map

Source-

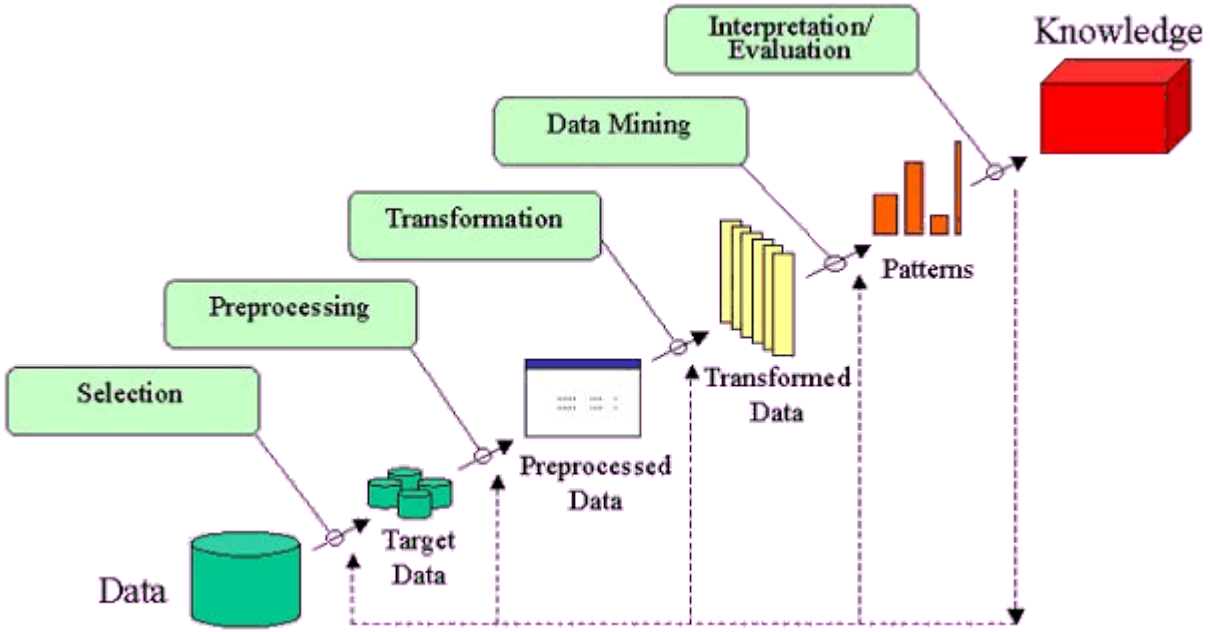
http://www.saedsayad.com/



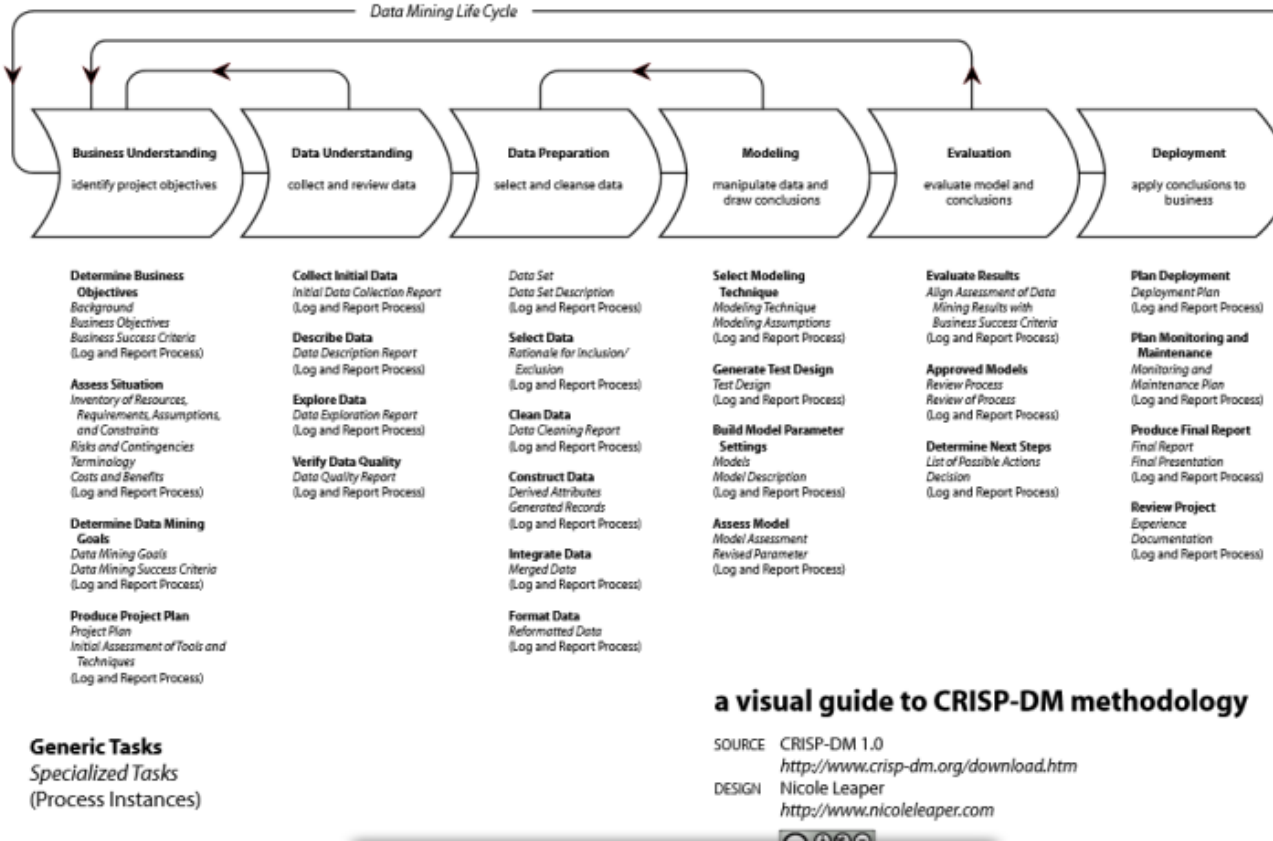
Examples of Data Mining

- 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)

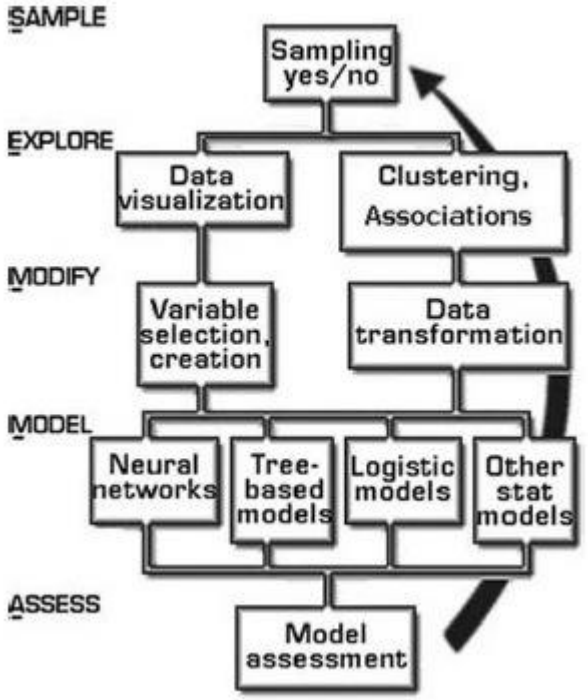
KDD


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

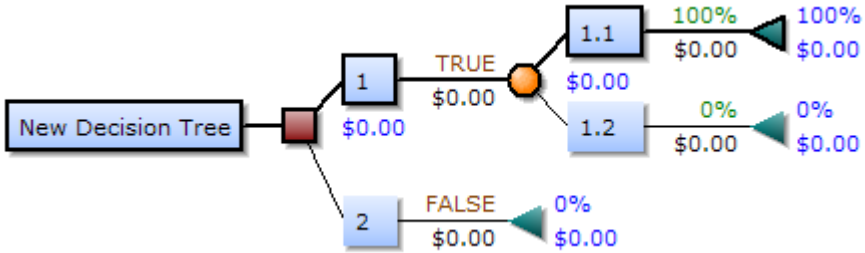
Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational **learning** theory in artificial intelligence.

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

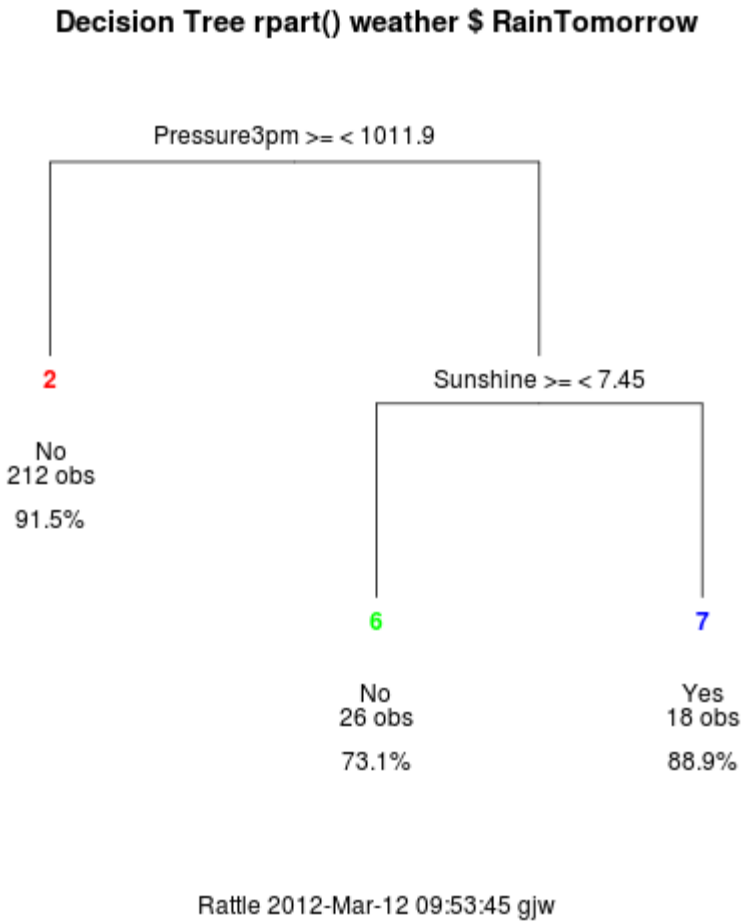
- Supervised learning. The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- Unsupervised learning, no labels are given to the learning algorithm, leaving it on its own to find structure in 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



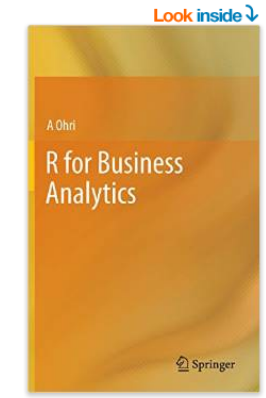
Association Analysis

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



R for Business Analytics Hardcover – Import, 11 Sep 2012

by A Ohri (Author)

1 customer review

See all 2 formats and editions

Kindle Edition ₹ 1,673.49 — Read with our free app

Hardcover ₹ 3,865.16 — 10 New from ₹ 3,285.86

EMI Available. Options

Delivery to pincode 110001 - Delhi within **1 - 2 weeks**. Details

R for Business Analytics looks at some of the most common tasks performed by business analysts and helps the user navigate the wealth of information in R and its 4000 packages. With this information the reader can select the packages that can help process the analytical tasks with minimum effort and maximum usefulness. The use of Graphical User Interfaces (GUI) is emphasized in this book to further cut down and bend the famous learning curve in learning R. This book is aimed to help you kick-start with analytics including chapters on data visualization, code examples on web analytics and social media analytics, clustering, regression models, text mining, data mining models and forecasting. The book tries to expose the reader to a breadth of business analytics topics without burying the user in needless

Read more

See all 2 images

Share

100% Purchase Protection
Genuine Products | Secure Payments
Easy Returns

₹ 3,865.16 + FREE Delivery
Inclusive of all taxes

Sold and fulfilled by B2A UK (4.4 out of 5 | 1,934 ratings).

Add to Cart

Buy Now

Add to Wish List

Other Sellers on Amazon

₹ 4,250.00 Add to Cart
+ ₹ 300.00 Delivery charge
Sold by: A1webstores

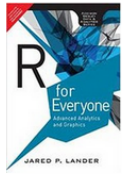
₹ 4,875.56 Add to Cart
+ FREE Delivery
Sold by: B2A US

₹ 4,937.00 Add to Cart
+ ₹ 300.00 Delivery charge
Sold by: uRead-shop

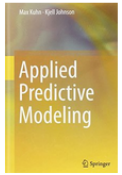
10 New from ₹ 3,285.86

Customers Who Bought This Item Also Bought

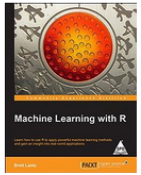
Page 1 of 4



R for Everyone: Advanced Analytics and Graphics

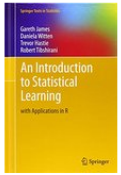


Applied Predictive Modeling



Machine Learning With R: Learn How to Use R to…

Data Smart: Using Data Science to Transform…



An Introduction to Statistical Learning:…

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

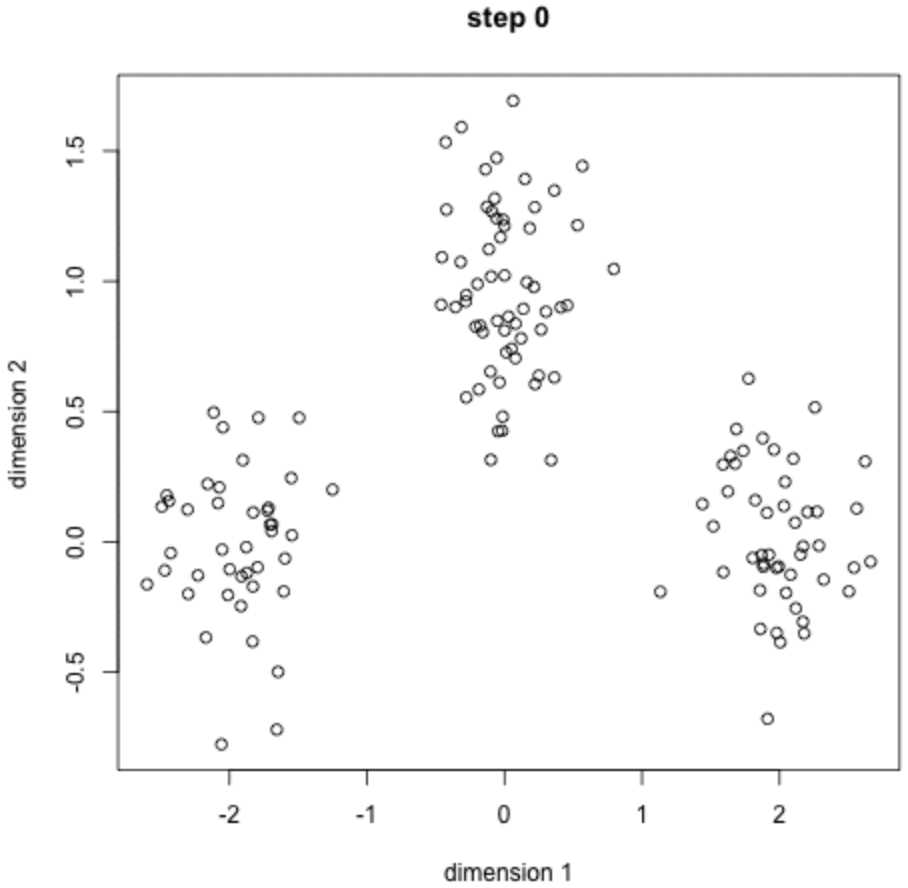
Clustering

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

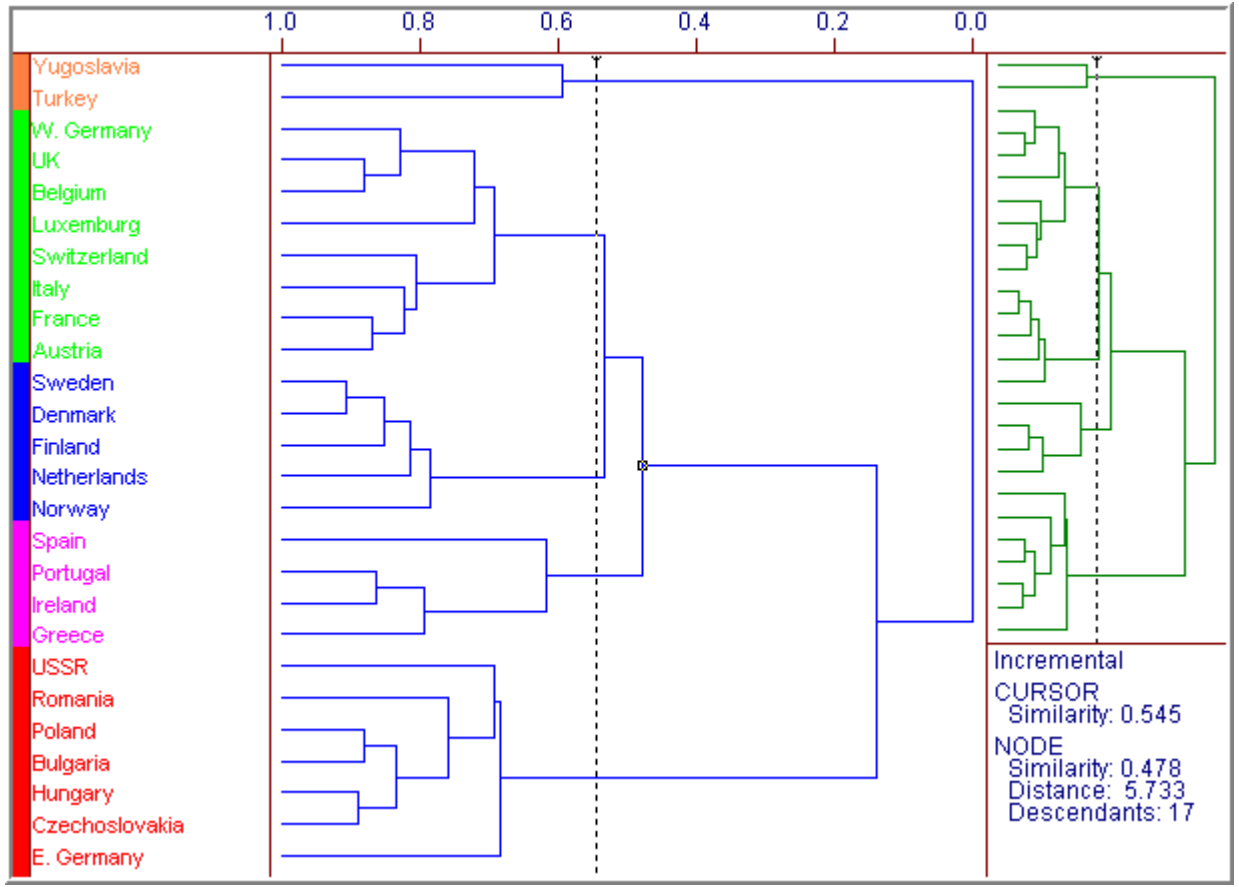


Clustering

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.

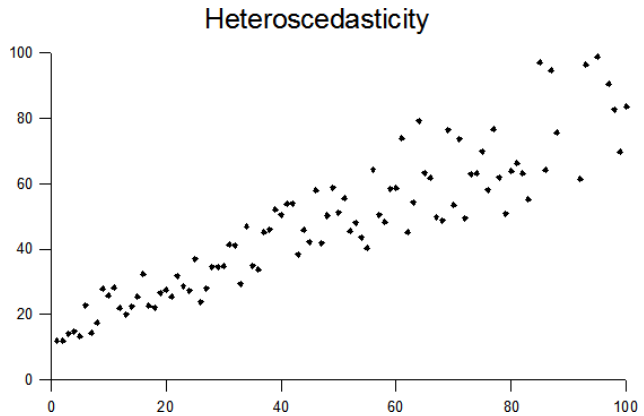
$y = a + bx$

$y = a + bx + cy$

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

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

Text mining, also referred to as **text** data **mining**, roughly equivalent to **text** analytics, refers to the process of deriving high-quality information from **text**.

Corpus - **text corpus** is a large and structured set of **texts**

A **document-term matrix** or **term-document matrix** is a mathematical **matrix** that describes the frequency of **terms** that occur in a collection of **documents**.

- D1 = "I like databases"
- D2 = "I hate databases",

then the document-term matrix would be:

	I	**like**	**hate**	**databases**
D1	1	1	0	1
D2	1	0	1	1

Text Mining

http://www.rdatamining.com/examples/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**

Sentiment Analysis

Sentiment analysis (also known as **opinion mining**) refers to the use of natural language processing, text analysis andcomputational 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.

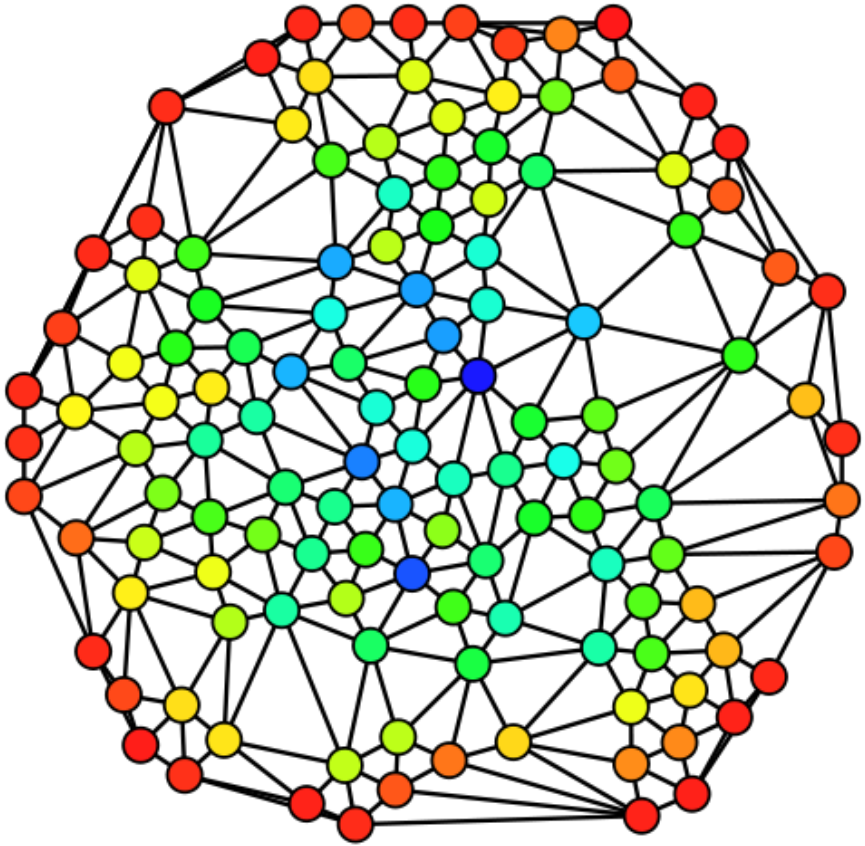
Density: The proportion of direct ties in a network relative to the total number possible.[25][26]

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

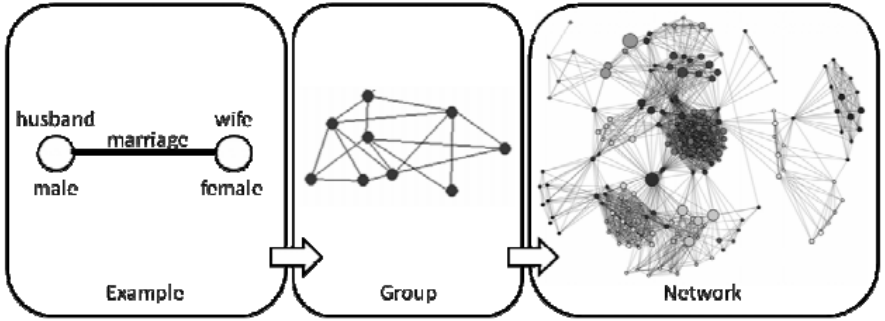


Figure 1. Social Network Analysis

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.

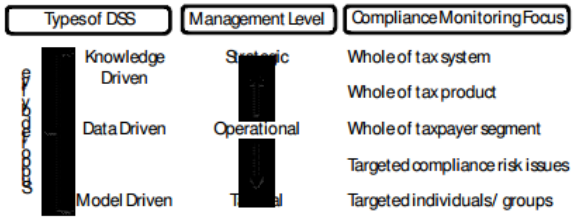
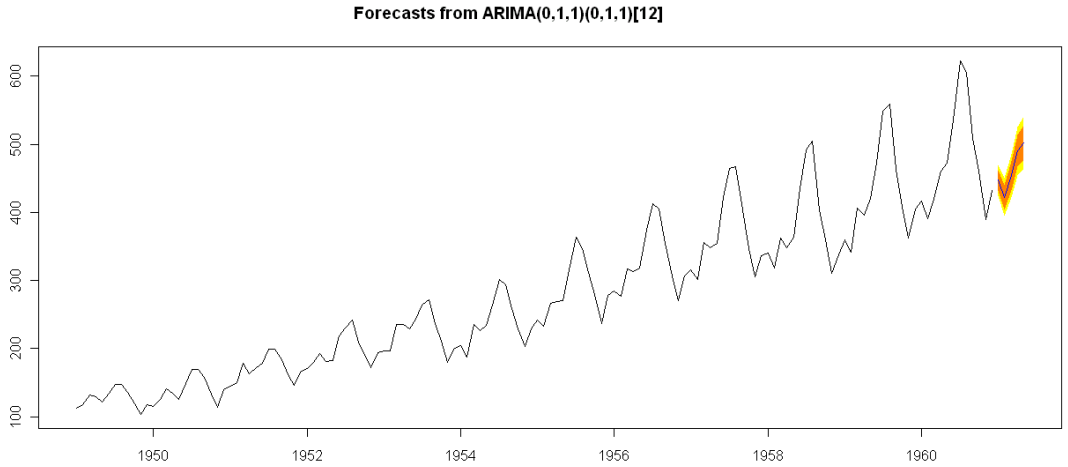


Figure 2. Compliance Management and Decision Support Systems (Derived from OECD, 2008:10; Power, 2002:13-16)

RESEARCH OBJECTIVE

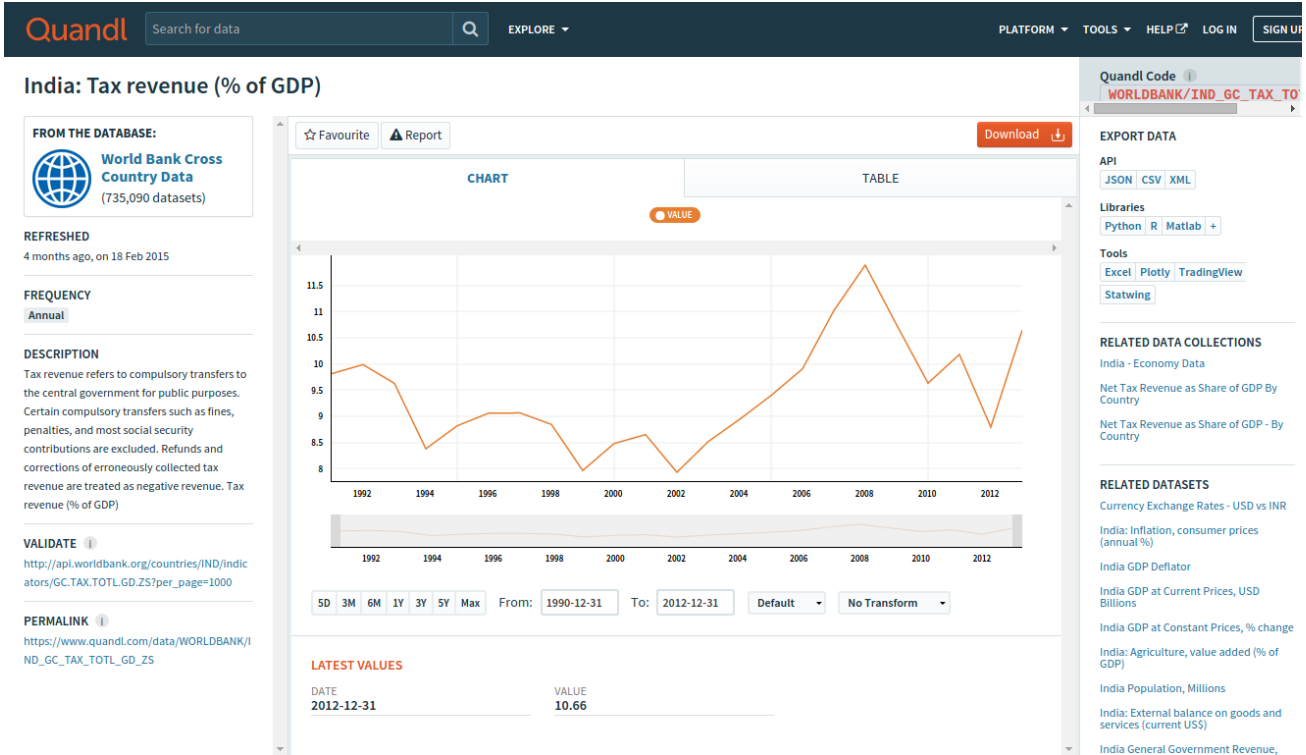
Time Series Forecasting

A **time series** is a sequence of data points, typically consisting of successive measurements made over a time interval. **Time series *analysis*** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series *forecasting*** is the use of a model to predict future values based on previously observed values.



Web Data for Time Series

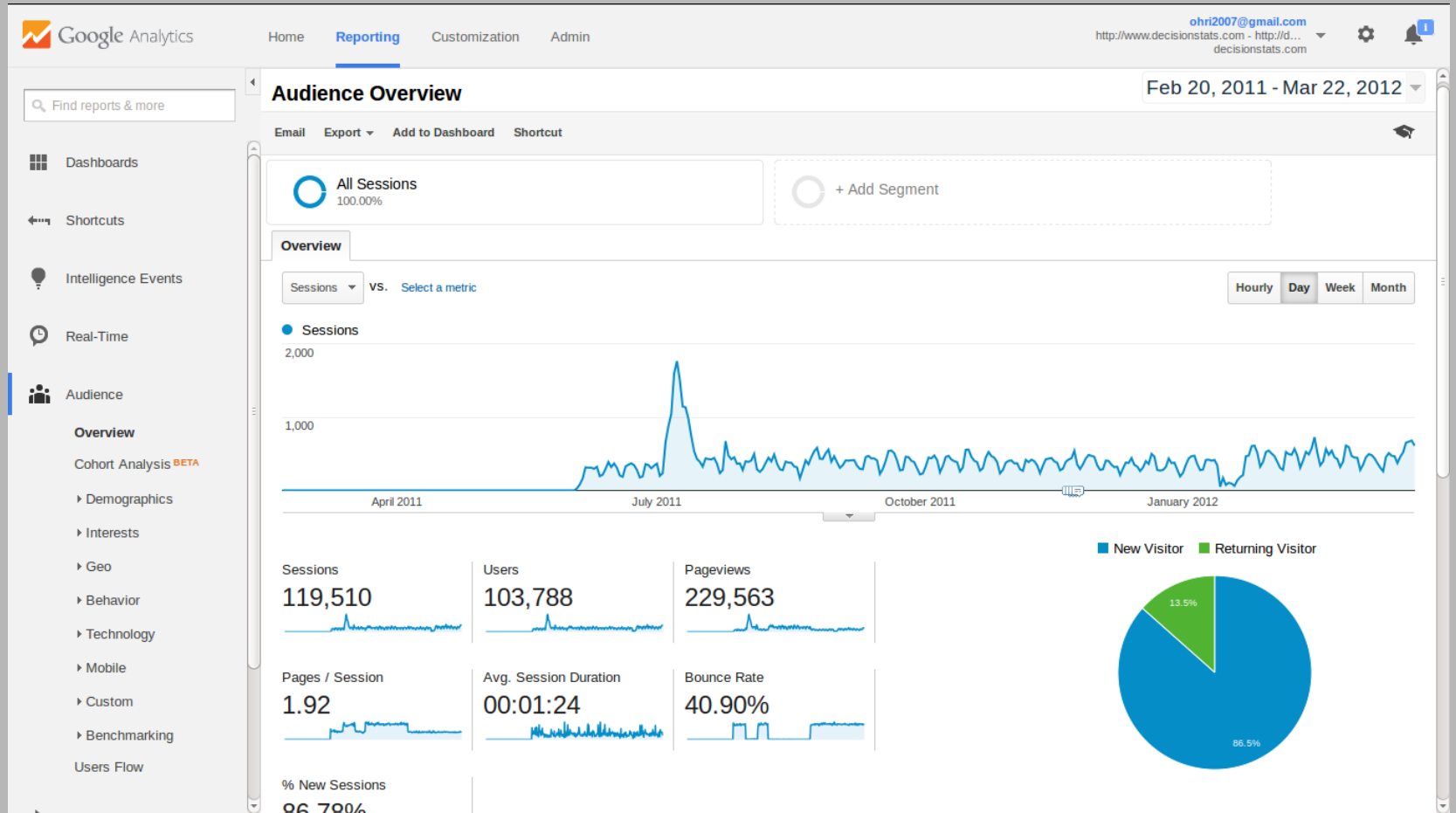
https://www.quandl.com/data/WORLDBANK/IND_GC_TAX_TOTL_GD_ZS-India-Tax-revenue-of-GDP



DECISION STATS

BETTER DECISIONS === FASTER STATS

Introduction to Web Analytics

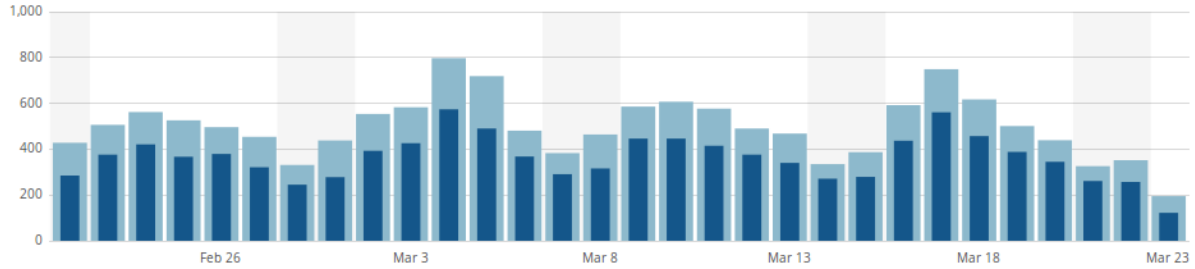


March 23, 2015, 7:50 am

DECISION STATS (WP.com)

Have you tried the upgraded stats page? Show Me

Days Weeks Months — Views Visitors — Summaries →



Today		Best ever	All time	
122	195	2,993	565,032	817
Visitors	Views	views	views	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

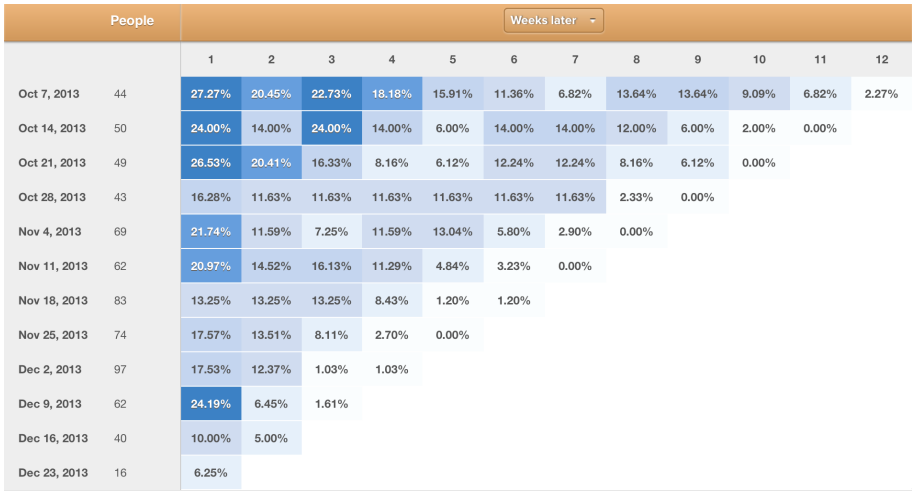
Today Yesterday — Summaries →

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
Running R on Amazon EC2	6

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%											

DECISION STATS

BETTER DECISIONS === FASTER STATS

Introduction to Blogging

Why Blogging?

Because this is the age of bloggers

You can build your own identity and reputation globally at no cost

Improves your communication

Helps job chances

Even PM of India recommends social media

Blog

how - content (topics, tags, categories)
navigation (themes, custom, widgets)
sharing (social, email , links)

Wordpress Basics

Admin Dashboard

Themes

Plugins

General

Blogging Basics

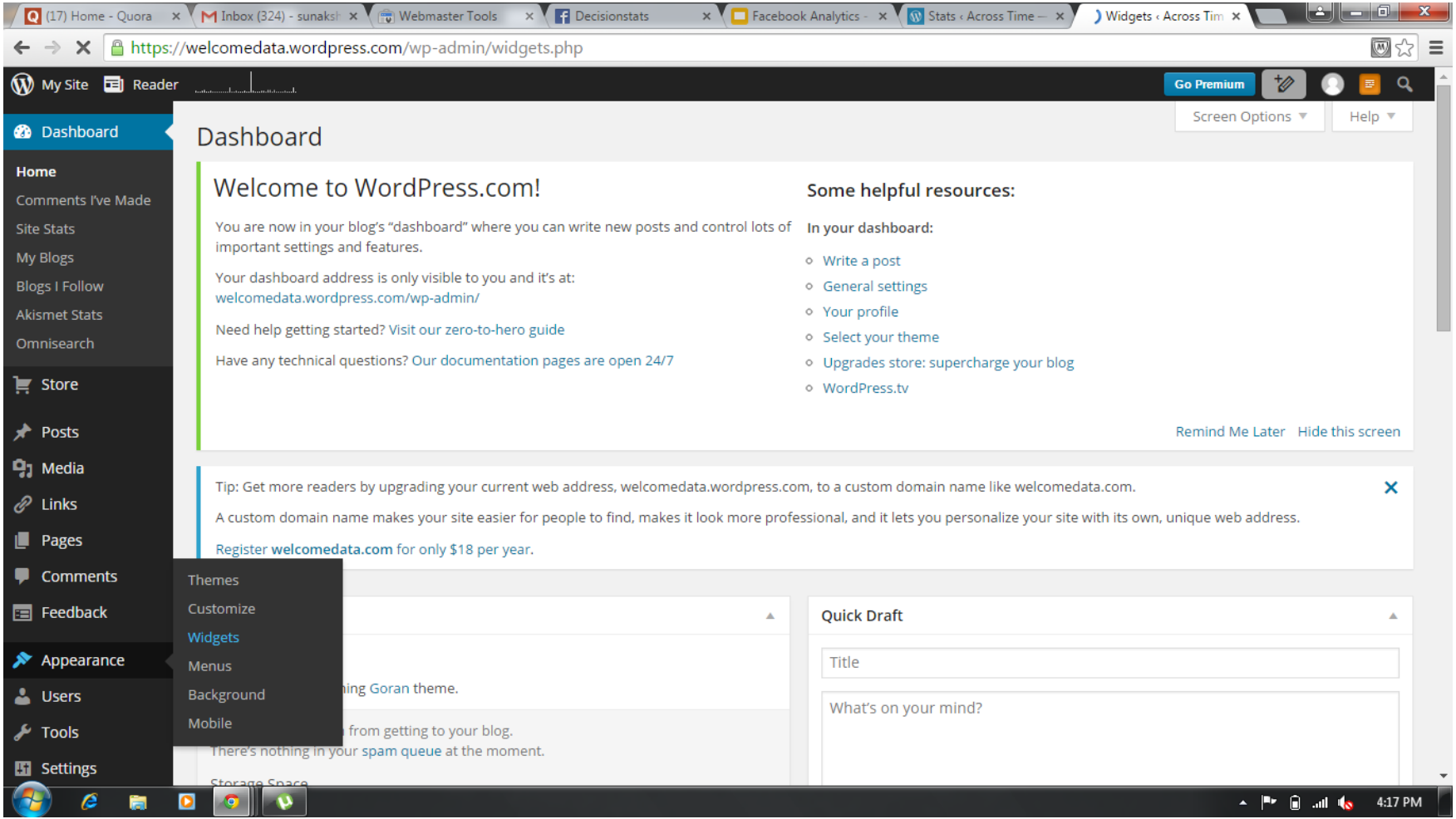
Content

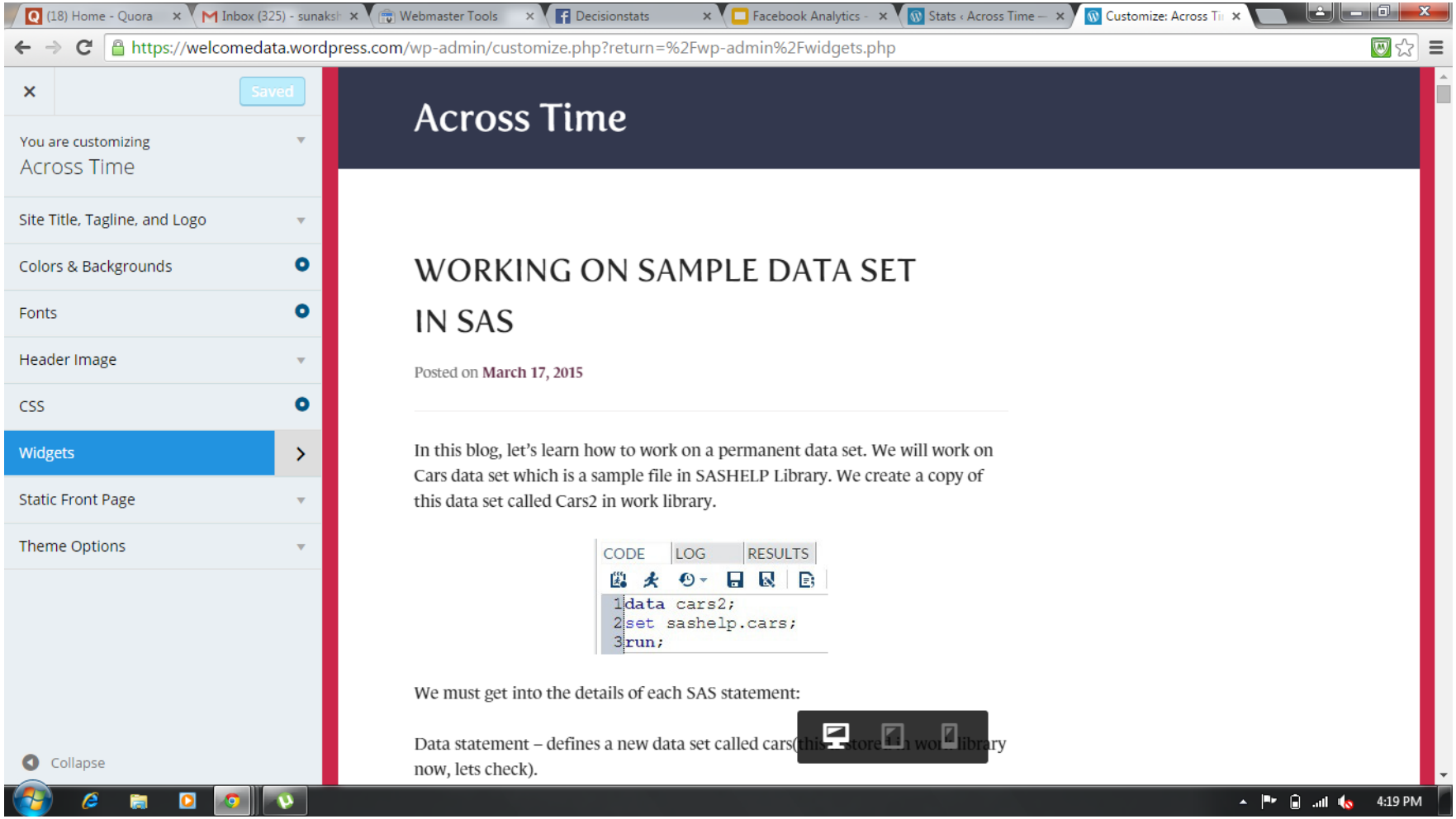
Blog Post Title

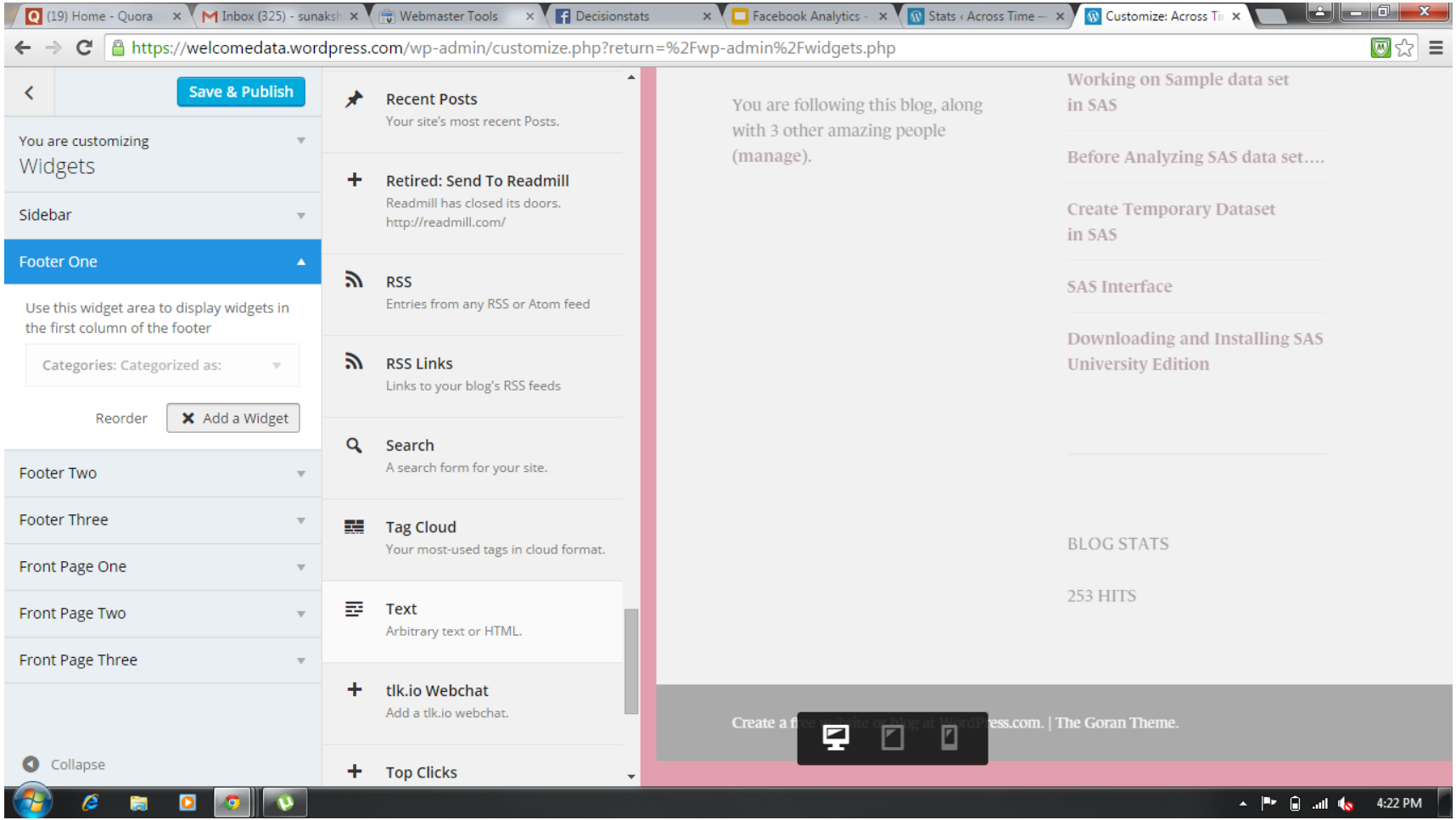
Navigation

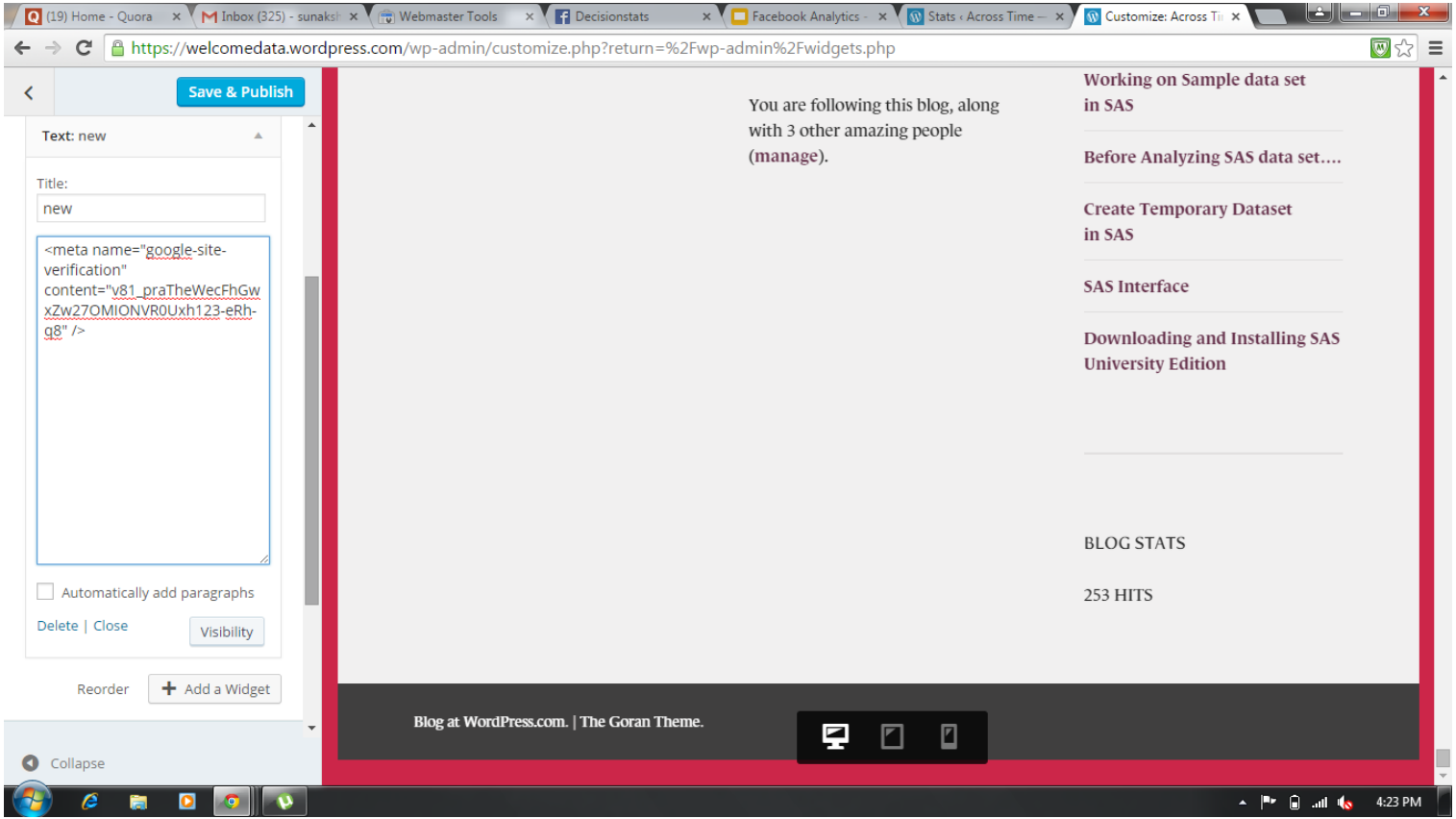
Theme

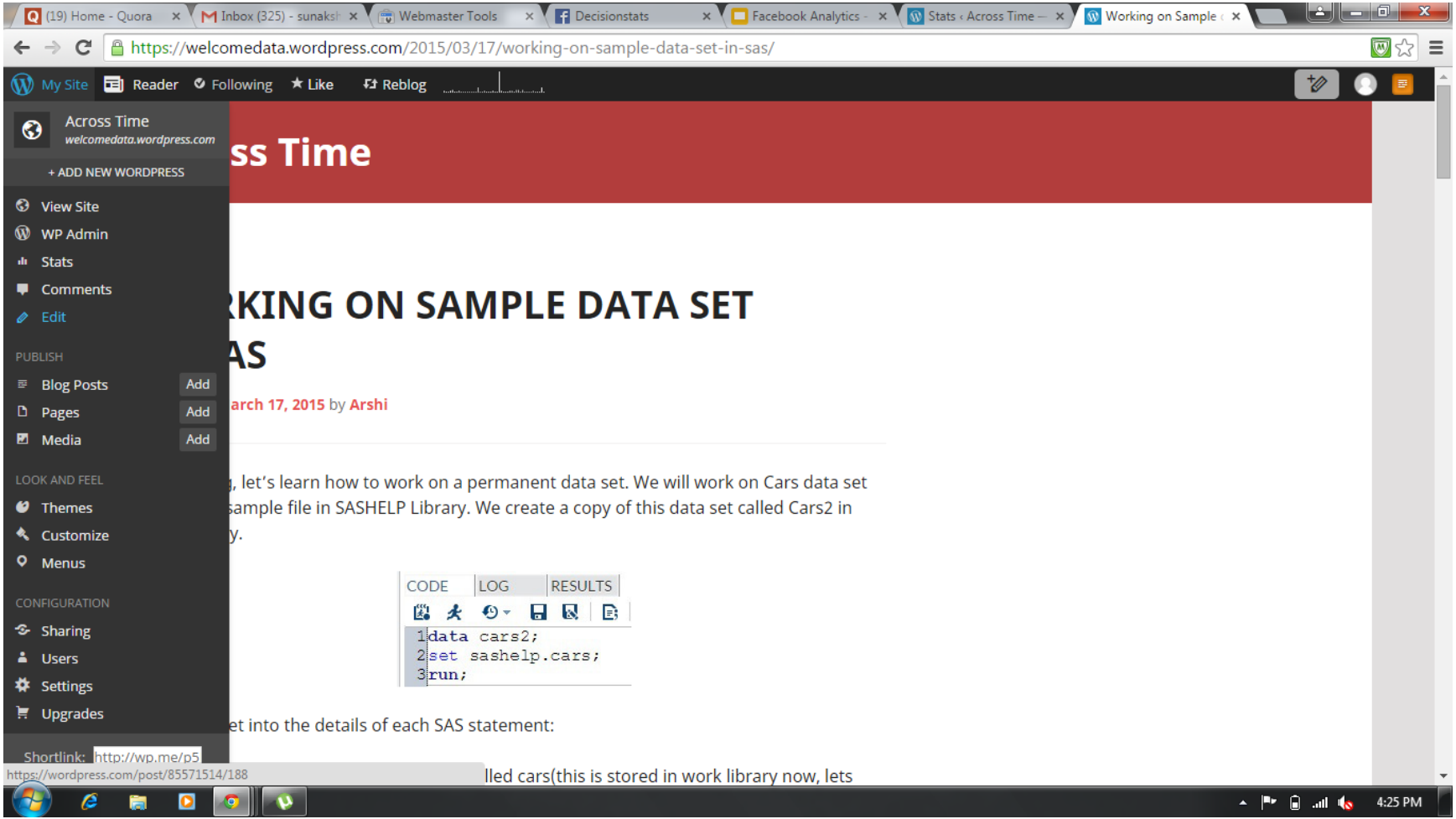
Sharing Content


(17) Home - Quora
Inbox (324) - sunaksh
Webmaster Tools
Decisionstats
Facebook Analytics -
Stats ‹ Across Time –
Widgets ‹ Across Tim
https://welcomedata.wordpress.com/wp-admin/widgets.php
My Site
Reader
Go Premium
Dashboard
Home
Comments I've Made
Site Stats
My Blogs
Blogs I Follow
Akismet Stats
Omnisearch
Store
Posts
Media
Links
Pages
Comments
Feedback
Appearance
Users
Tools
Settings
Themes
Customize
Widgets
Menus
Background
Mobile
Screen Options
Help
Dashboard
Welcome to WordPress.com!
You are now in your blog's "dashboard" where you can write new posts and control lots of important settings and features.
Your dashboard address is only visible to you and it's at:
welcomedata.wordpress.com/wp-admin/
Need help getting started? Visit our zero-to-hero guide
Have any technical questions? Our documentation pages are open 24/7
Some helpful resources:
In your dashboard:
Write a post
General settings
Your profile
Select your theme
Upgrades store: supercharge your blog
WordPress.tv
Remind Me Later
Hide this screen
Tip: Get more readers by upgrading your current web address, welcomedata.wordpress.com, to a custom domain name like welcomedata.com.
A custom domain name makes your site easier for people to find, makes it look more professional, and it lets you personalize your site with its own, unique web address.
Register welcomedata.com for only $18 per year.
Quick Draft
Title
What's on your mind?
ing Goran theme.
from getting to your blog.
There's nothing in your spam queue at the moment.
Storage Space
4:17 PM


(18) Home - Quora
Inbox (325) - sunaksh
Webmaster Tools
Decisionstats
Facebook Analytics -
Stats ‹ Across Time —
Customize: Across Ti
https://welcomedata.wordpress.com/wp-admin/customize.php?return=%2Fwp-admin%2Fwidgets.php
Saved
You are customizing
Across Time
Site Title, Tagline, and Logo
Colors & Backgrounds
Fonts
Header Image
CSS
Widgets
Static Front Page
Theme Options
Collapse
Across Time
WORKING ON SAMPLE DATA SET IN SAS
Posted on March 17, 2015
In this blog, let's learn how to work on a permanent data set. We will work on Cars data set which is a sample file in SASHELP Library. We create a copy of this data set called Cars2 in work library.
CODE
LOG
RESULTS
1 data cars2;
2 set sashelp.cars;
3 run;
We must get into the details of each SAS statement:
Data statement – defines a new data set called cars
now, lets check).
4:19 PM


(19) Home - Quora
Inbox (325) - sunaksh
Webmaster Tools
Decisionstats
Facebook Analytics -
Stats ‹ Across Time –
Customize: Across Ti
https://welcomedata.wordpress.com/wp-admin/customize.php?return=%2Fwp-admin%2Fwidgets.php
Save & Publish
You are customizing
Widgets
Sidebar
Footer One
Use this widget area to display widgets in the first column of the footer
Categories: Categorized as:
Reorder
Add a Widget
Footer Two
Footer Three
Front Page One
Front Page Two
Front Page Three
Collapse
Recent Posts
Your site's most recent Posts.
Retired: Send To Readmill
Readmill has closed its doors. http://readmill.com/
RSS
Entries from any RSS or Atom feed
RSS Links
Links to your blog's RSS feeds
Search
A search form for your site.
Tag Cloud
Your most-used tags in cloud format.
Text
Arbitrary text or HTML.
tlk.io Webchat
Add a tlk.io webchat.
Top Clicks
You are following this blog, along with 3 other amazing people (manage).
Working on Sample data set in SAS
Before Analyzing SAS data set....
Create Temporary Dataset in SAS
SAS Interface
Downloading and Installing SAS University Edition
BLOG STATS
253 HITS
ess.com. | The Goran Theme.
4:22 PM


(19) Home - Quora
Inbox (325) - sunaksh
Webmaster Tools
Decisionstats
Facebook Analytics -
Stats ‹ Across Time —
Customize: Across Ti
https://welcomedata.wordpress.com/wp-admin/customize.php?return=%2Fwp-admin%2Fwidgets.php
Save & Publish
Text: new
Title:
new
<meta name="google-site-verification" content="v81_praTheWecFhGwxZw27OMIONVR0Uxh123-eRh-q8" />
Automatically add paragraphs
Delete | Close
Visibility
Reorder
Add a Widget
Collapse
You are following this blog, along with 3 other amazing people (manage).
Working on Sample data set in SAS
Before Analyzing SAS data set....
Create Temporary Dataset in SAS
SAS Interface
Downloading and Installing SAS University Edition
BLOG STATS
253 HITS
Blog at WordPress.com. | The Goran Theme.
4:23 PM


(19) Home - Quora
Inbox (325) - sunaksh
Webmaster Tools
Decisionstats
Facebook Analytics -
Stats ‹ Across Time —
Working on Sample
https://welcomedata.wordpress.com/2015/03/17/working-on-sample-data-set-in-sas/
My Site
Reader
Following
Like
Reblog
Across Time
welcomedata.wordpress.com
+ ADD NEW WORDPRESS
View Site
WP Admin
Stats
Comments
Edit
PUBLISH
Blog Posts
Add
Pages
Add
Media
Add
LOOK AND FEEL
Themes
Customize
Menus
CONFIGURATION
Sharing
Users
Settings
Upgrades
Shortlink: http://wp.me/p5
ss Time
RKING ON SAMPLE DATA SET
AS
arch 17, 2015 by Arshi
, let's learn how to work on a permanent data set. We will work on Cars data set
sample file in SASHELP Library. We create a copy of this data set called Cars2 in
y.
CODE
LOG
RESULTS
1 data cars2;
2 set sashelp.cars;
3 run;
et into the details of each SAS statement:
lled cars(this is stored in work library now, lets
https://wordpress.com/post/85571514/188
4:25 PM

Quiz Time

LTV

https://docs.google.com/forms/d/1lLbkLTZgZVrM7EovRibhCD10qN38Tk9xrwOcZ6WPYiY/viewform

RFM

https://docs.google.com/forms/d/1_LPANhgPURQi_8zi840TGs9ahS4G6VqexnNWRDCvq5w/viewform

Summer School