Data Science with R

Day 1 - Introduction to Data Science

TACT 360° Training

LEARN. ACHIEVE. STANDOUT

www. tactlearn.com

Today’s Agenda

✓ **Data Science Introduction**

- What is Data Science?
- Objectives of a Data Science Project

✓ **Data Science Toolkit**

- What you need to learn?

✓ **Job outlook**

- What are the opportunities?

✓ **Prerequisites**

- What you should know before getting started with Data Science?

✓ **Your Specific Questions**

Hemant Rathore

*"Data Scientist & Freelance Corporate Trainer with vast industry experience in the fields of **Business Intelligence, Data Science, Statistics, Machine Learning, Data warehouse, Data Modeling,** and **Analytics.** A Data Science enthusiast and an adept Data Analyst with rich experience over wide range of BI tools and technologies."*

- **Data Modeling/DWH designing** Concepts using **ER Studio**
- **Data Science, Predictive & Descriptive Analysis using R**
- **Visual Analytics** using **Tableau & SAP Business Objects**
- **Complete Statistics** using **R**
- **Statistical and Inferential Analysis** using R
- **Machine Learning** using **R, MS Azure Machine Learning Studio**
- **ETL/Data Integration** concepts using **Informatica**
- **Data profiling** and **data quality** analysis using **Informatica IDQ**
- **Data Virtualisation** Concepts using **Cisco Composite (CIS)**
- **Data Replication** using **HVR**
- **PL/SQL Concepts**

Data Science Introduction

What is Data Science?

“Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data present in various forms, either structured or unstructured.”

Data science employs **techniques** and **theories** drawn from a wide range of disciplines like **Mathematics, Statistics, Information Science,** and **Computer Science**, in particular from the subdomains of **Machine learning, Classification, Association, Cluster analysis, Data mining, forecasting** and **Visualisation** in order to understand and analyse actual phenomena with data.

Data Science : Data + Science = Knowledge

Data Science Introduction

Data Science

Some Common Objectives/ Types of a Data Science Project

- **Regression Analysis –** Finding the relationship between a dependent variable and one or more independent variables, **Predicting Diamond price based on Carat, Cut & Clarity**
- **Classification Analysis –** Dividing objects into 2 or more known classes, **Distinguishing cancer and normal cells**
- **Anomalies Detection(Outliers Analysis)** - Finding unusual, **Credit card transactions**
- **Association Analysis** – Finding links, **Shopping cart analysis**
- **Cluster Analysis** (Segmentation) – Grouping similar objects together (unknown classes/groups), **Grouping customers into different clusters based on their previous shopping data/transactions.**
- **Time Series Analysis** – Time dependent Data, **Stock prediction**

Hemant Rathore

Data Science Toolkit

What you need to learn?

Data Science

- **Data Engineering** – Getting and Processing the data, **R, Python, PL/SQL, SAS, ETL, Big Data**
- **Data Analysis** – Exploratory Data Analysis, Inferential Statistics, Predictive and Descriptive Analysis **R, Python, SAS**
- **Statistics & Probabilities** – Basic and advance concepts of Statistics and probability to understand data, **R, Python, MS Excel**
- **Reporting or Analytics** – Effective visual Presentation of data and findings, **R, Python, SAS, Tableau or other visualization tools**
- **Machine Learning** – To Apply the algorithm on data, **R, Python, SAS, Microsoft Azure ML Studio, IBM SPSS, SAP Predictive Analytics**

++ Business or Domain Knowledge ++

Job outlook

What are the opportunities? – “Data Science is not a Technology but a whole New world”

- **Data Engineer**
- **Dashboard/Analytics Expert**
- **Data Analyst**
- **Machine Learning Expert**
- **Data Scientist – One for ALL!!**

Prerequisites & Target Audience

Prerequisite or who is suitable to get into Data Science?

Data Science

Professional already working as Data/Business Analyst, Analytics Expert, BI Developers, ETL/Big data Engineer.

- Basic understanding of programming concepts **PL/SQL, R, SAS, Python, ETL, Big Data, or any other programming languages like C, C++, Java**
- Basic knowledge of **Mathematics** and **Statistics** Concepts
- Basic Knowledge of Reporting or Visualization, **Tableau, Spotfire, SAP Business Objects or other reporting tools**
- Great Determinations! Dedications! Consistency! and Confidence !!

Get Set Ready!!!

Homework

- Do Some Research on What is Data Science & Machine Learning
- Try to understand CRISP-DM Methodology

https://www.the-modeling-agency.com/crisp-dm.pdf

“Qs & As”

TACT 360^{0} Training

LEARN. ACHIEVE. STANDOUT

For more information or to set up an appointment, please contact us today.

jointact@tactlearn.com

Data Science with R

Day 2 - Data Science Project Life Cycle & Basics of Statistics

TACT 360⁰ Training

LEARN. ACHIEVE. STANDOUT

www. tactlearn.com

Today’s Agenda

✓ **Data Science Project Life Cycle**

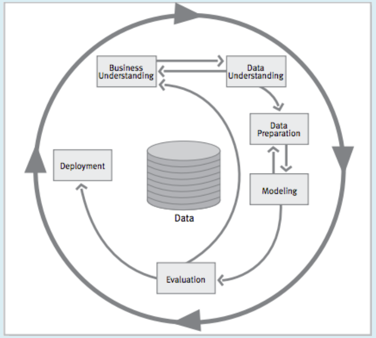
- CRISP - DM Process Model
- CRISP - DM Phases
  - ▪ Business understanding
  - ▪ Data understanding
  - ▪ Data preparation
  - ▪ Modeling
  - ▪ Evaluation
  - ▪ Deployment

✓ **Basics of Statistics**

- Random Variable
- Type of Random Variables
- Central Tendencies- Mean Mode, Median
- Probability, Probability Distribution of Random Variables

Data Science Project Life Cycle

What is CRISP-DM ?



CRISP-DM was conceived in late 1996 by three veterans of the young and immature data mining market.CRISP Stands for “CRoss-Industry Standard Process for Data Mining”

This Process model for data mining provides an overview of the life cycle of a data mining project. It contains all the phases of a project, their respective tasks, and the relationships between these tasks. Relationships could exist between any data mining tasks depending on the goals, the background, and the interest of the user-and most importantly-on the data.

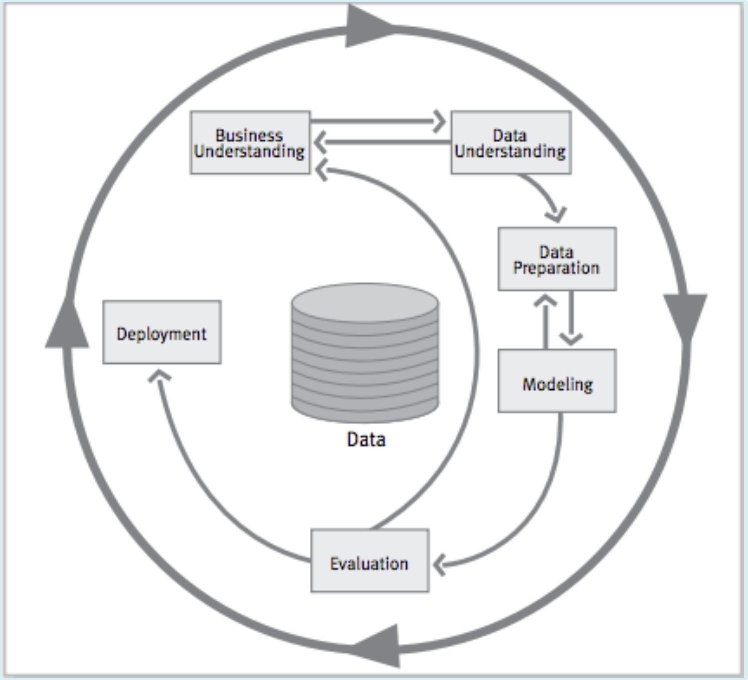
https://www.the-modeling-agency.com/crisp-dm.pdf

Data Science Project Life Cycle

CRISP-DM Phases

The life cycle of a data mining project consists of six phases as shown -

1. **Business understanding** - This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. **Data understanding -**The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data.
3. **Data preparation -** The data preparation phase covers all activities needed to construct the final dataset or data that will be fed into the modeling tool(s) from the initial raw data.Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.



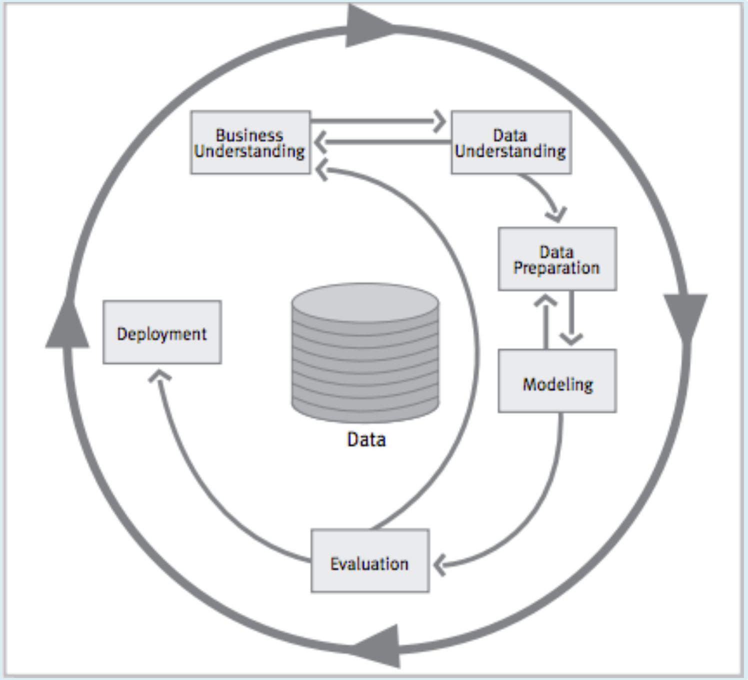
https://www.the-modeling-agency.com/crisp-dm.pdf

Data Science Project Life Cycle

CRISP-DM Phases

4. **Modeling** - In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.

5. **Evaluation** - At this stage you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered.

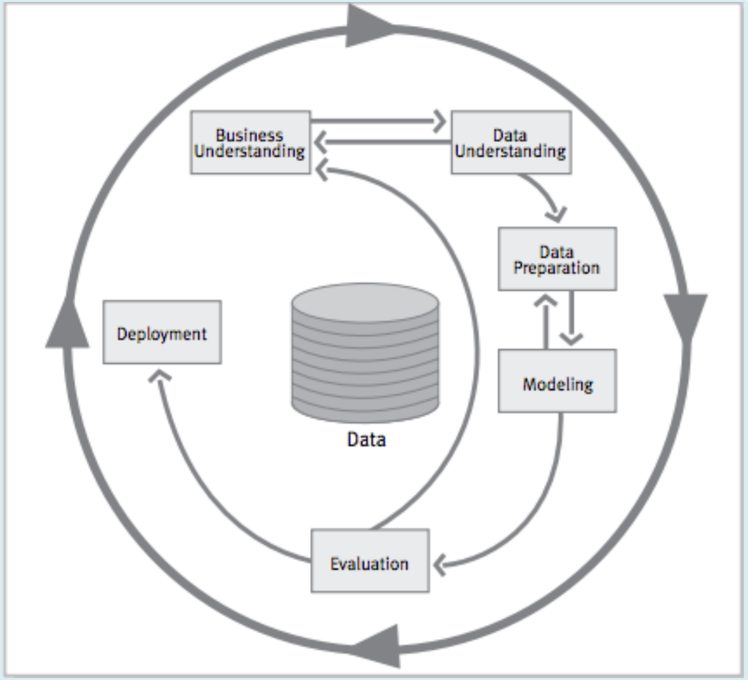


https://www.the-modeling-agency.com/crisp-dm.pdf

Data Science Project Life Cycle

CRISP-DM Phases

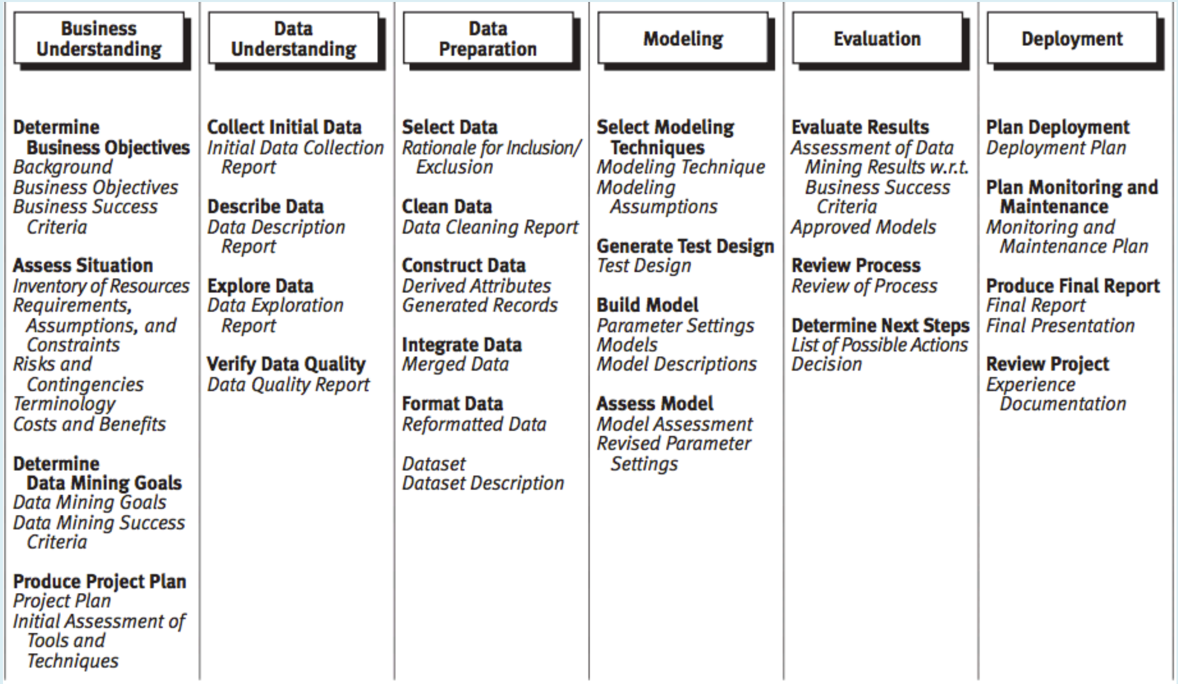
6. **Deployment-** Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organised and presented in a way that the customer can use it. It often involves applying “live” models within an organisation’s decision making processes.Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.



https://www.the-modeling-agency.com/crisp-dm.pdf

Data Science Project Life Cycle

CRISP-DM Phases



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

https://www.the-modeling-agency.com/crisp-dm.pdf

Basics of Statistics

What is Statistics?

Statistics is a branch of Mathematics dealing with the collection, analysis, interpretation, presentation, and organisation of data.

Random variable

A Variable which is used to store value corresponding to each outcome of a Random Experiment/Event/Activity. Ex. Coin Flip, RV= {H,T}

Type of Random variable

Based on the nature of outcome RV can be Discrete or Continuos.
Discrete - Finite measurement, no decimals, Ex. Number of people
Continues - Infinite measurements between 2 consecutive values, Ex. Weight, Age

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

Hemant Rathore

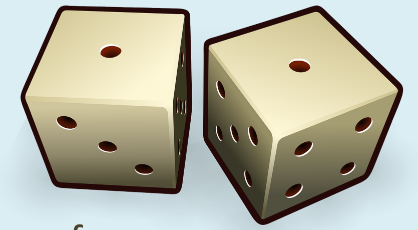
Basics of Statistics

Central Tendencies- Mean Mode, Median

Type	Description	Example	Result
Mean	Sum of values of a data set divided by number of values:	(1+2+2+3+4+7+9) / 7	4
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, 3, 4, 7, 9	3
Mode	Most frequent value in a data set	1, 2, 2, 3, 4, 7, 9	2

https://en.wikipedia.org/wiki/Mode_(statistics)

Basics of Statistics



Basic Probability

Probability is the measure of the likelihood that an event will occur. In case of Random variable we are interested in knowing the Probabilities of getting different values.

Probability - $P(X=Xi) = F(Xi) / F_{Total}$

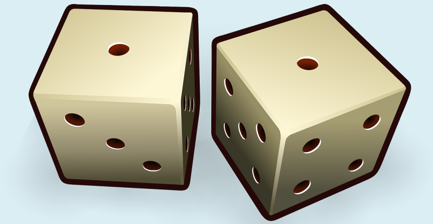
Ex. Rolling Dices, $P(Output=1) = 1/6$

Hemant Rathore

Basics of Statistics

Probability Distribution of RV

Table/ Chart/Formula to show relationship between Values and Corresponding Probabilities or shows the distribution of probabilities by values.



Type of Probability Distribution

Based on type of RV, Probability Distribution can also be either Discrete or Continuous

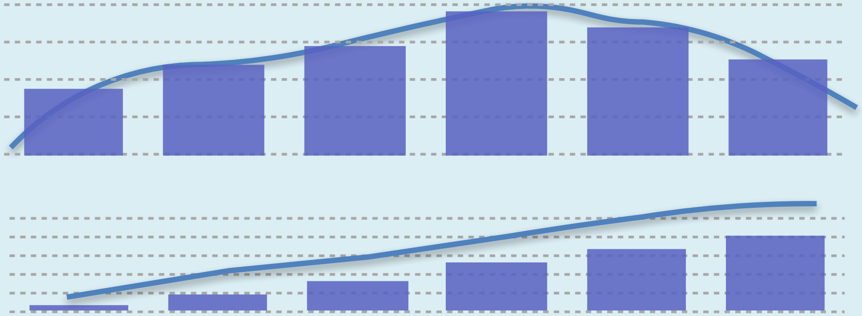
1. Discrete Probability Distribution – Probability Mass Function (PMF)

Discrete Probability Distribution (PMF)						
Values(X)	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

Basics of Statistics

Type of Probability Distribution

2. Continuos Probability Distribution – Probability Density Function (PDF)



AGE GROUP	#PERSONS(X) (IN K)	P(X)	CP(X)
0-10	11	0.105	0.105
10-20	15	0.143	0.248
20-30	18	0.171	0.419
30-40	24	0.229	0.648
40-50	21	0.200	0.848
50-60	16	0.152	**1**
Total	**105**	**1**	

Hemant Rathore

“Qs & As”

Data Science with R

Day 3 - Statistics for Data Science – Basics and Advanced

LEARN. ACHIEVE. STANDOUT

www. tactlearn.com