The Course of STATISTICAL DATA ANALYSIS - cfu 12
Recap and Project Work for the Exam

Roberta Siciliano

Stad Lab

Statistics, tèchnè-loghìa, analysis of data

University of Napoli Federico II – http://www.stad.unina.it

DATA SCIENCE UNINA

A thougth

Father of Indian Statistics: Prof. Prasanta Chandra Mahalanobis

“Statistics must have a clearly defined purpose, one aspect of which is scientific advancement and the other human welfare and national development.”

https://artsandculture.google.com/story/father-of-indian-statistics-prof-prasanta-chandra-mahalanobis-indian-statistical-institute/wAURK23-669lLA?hl=en

Contents

Statistics, tèchnè-loghìa, anàlysis data

Stad - Statistics, tèchnè-loghìa, anàlysis dàto

- ***Statistics, tèchnè-loghìa, anàlysis data*** are key words for reading the reality as it is manifested in **qualitative** expressions (categories, attributes, classes, groups, labels, etc.) and **quantitative** ones (numbers, measurements, values, etc.) in various domain contexts or application fields.
- All starts with a real-world case study and specific questions that require **statistical data analysis.**

Stad – The etymological derivation

- ***Statistics*** from the Latin ***Status*** = ***"science dealing with data about the condition of a state or community"*** coined by the German political scientist Gottfried Aschenwall (1719-1772) in his "Vorbereitung zur Staatswissenschaft" (1748).
- ***tèchnè-loghìa*** combines two Greek words:
  - ***tèchnè*** = *art, ability of doing, namely the human brain to 'remember', 'think', 'come to mind', 'cross the mind', 'invent something new',*
  - ***loghìa*** = *discourse, explanation,*
  - ***tèchnè-loghìa*** = ***"the systematic treatise on an art or how to do"***
- ***anàlysis*** from the Greek = *solution of a problem,* ***"the process of breaking a complex topic into smaller parts in order to gain a better understanding of it"*** *(Aristotele, 384-322 B.C.),*
- ***dàta*** from the Latin = ***known quantities****.*

Statistics, tèchnè-loghìa, anàlysis dàto

- ***Statistics*** has two fundamental missions:
  - ✓ **Exploratory Data Analysis** *(from data to information)* uses a *"deducting approach"* to discover significant facts, patterns, groups, anomalies, associations, correlation, similarities, typologies, clusters, and so on;
  - ✓ **Confirmatory Data Analysis** *(from information to knowledge)* uses an *"inductive approach"* to justify theories and hypotheses, to build up models for decision-making and prediction.
- ***tèchnè-loghìa*** is the rationalization process that starting from the real-world questions build up a "project" to move from theory to practice. The concrete ability pass through two attitudes:
  - ✓ ***Heuristic Experience*** through trial and error to seek solutions for a new problem,
  - ✓ ***Algorithmic Experience*** by applying known solutions for a problem that is very similar to others previously addressed.
- ***analysis of data*** is to find solutions from known quantities, the data.

Statistical Data Analysis: the mission

The statistician or data scientist, to understand the facts (exploration) and to confirm the theories (confirmation), makes use of *reasoning and concrete skills* (statistical thinking and methodology) to make the most of the available data with the related domain context information as a result of observing reality.

The final goal of **Statistical Data Analysis** is *to learn from data* and *find solutions* to a real-world problem.

All starts with real-world case study in a domain context.

The statistician or data scientist needs to transform ***raw data (input)*** into ***useful information and knowledge (output)*** (tables, plots and graphs, info-graphics, statistical indexes, models, prediction) to be correctly interpreted and analyzed such to provide answers to the real-world questions as well as some ***actions and consequences (outcome) to add value*** (utility, revenue, satisfaction, innovation plan, etc.) for some stakeholders.

Statistical Data Analysis Course: the Syllabus

Statistical Data Analysis - Module A: Fundaments of Statistics
Introduction to Statistics, *Technè-Loghìa*, Analysis of Data
Exploratory Data Analysis (Data Pre-processing, Descriptive Statistics, Data Visualization, Data Cross-Classification Analysis)
Probability and Probability Models (Axiomatic Theory of Probability, Discrete and Continuous Models)
Confirmatory Data Analysis (Classical and Bayesian Inference, Estimation and Testing Theory)
Linear Regression Models (ANOVA, Linear Regression, Use of Dummy Variables in Regression, Regression Diagnostics)
Non-Parametric Tests
Multiple Testing Procedures
Linear Models for Time Series Analysis
Multidimensional Data Analysis (Principal Component Analysis, Hierarchical Clustering, K-Means Clustering)
Statistical Data Analysis - Module B: Statistical Learning
Statistical Learning in theory and practice
Linear Models for Regression
Linear Models for Supervised Classification
Computational Inference
Regularization and Shrinkage Methods
Model Selection
Methods for Non-Linearity in Regression and Classification
Classification and Regression Trees
Ensemble Methods
Support Vector Machines
Projection Pursuit Regression
Introduction to Deep Learning
Survival Analysis and Censored Data

Readings/Bibliography

The Basic Practice of Statistics, (the newest is the) 9th Edition (2021), David S. Moore, William I. Notz, Michael A. Fligner, W.H. Freeman Publishers

Introduction to Statistical Learning, with applications in R. James Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani (2009).

https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

The elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer (2009).

https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Slide and teaching material by the docent are uploaded on the Team of The Course.

Lectures Recording in the Team of The Course.

Lectures Scheduling

SDA - LM DATA SCIENCE			Exploratory Data Analysis			Theory	Laboratory
Statistical Data Analysis			Module A Fundaments of Statistics Inference				
			Module B Statistical Learning Modeling and Prediction				
			Project Work				
#	**Date**	**Day**	**Lecture**	**C**	**#H**	**# T**	**# L**
1	04/10/22	Tuesday	Introduction to Statistics, Technè-Loghìa, Analysis of Data	Intro	3	3	
2	05/10/22	Wednesday	Data Pre-Processing and Descriptive Statistics	EDA	2	2	
3	06/10/22	Thursday	Bivariate Data Analysis and Visualization / Laboratory in R	EDA	3	2	1
4	11/10/22	Tuesday	Probability Models / Laboratory in R	I	3	2	1
5	12/10/22	Wednesday	Probability Models	I	2	2	
6	13/10/22	Thursday	Inference / Laboratory in R	I	3	2	1
7	18/10/22	Tuesday	Laboratory in R	I	3		3
8	19/10/22	Wednesday	Bayesian Inference/Laboratory in R	I	2		2
9	20/10/22	Thursday	Linear Models for Experimental Data: ANOVA Modeling	M	3	3	
10	25/10/22	Tuesday	Linear Regression Models / Statistical Thinking	M	3	2	1
11	26/10/22	Wednesday	Laboratory in R	M	2	2	
12	27/10/22	Thursday	FATER CHALLENGE	PW	3	2	1
13	02/11/22	Wednesday	Multiple Linear Regression	M	2	2	
14	03/11/22	Thursday	Regression Diagnostics / Laboratory in R	M	3	2	1
15	08/11/22	Tuesday	Linear Models for Time Series	M	3	3	
16	09/11/22	Wednesday	Non-Parametric Tests	I	2	2	
17	10/11/22	Thursday	Non-Parametric Tests	EDA	3	3	
18	15/11/22	Tuesday	Multidimensional Data Analysis: PCA, Cluster	EDA	3		3
					48	34	14
19	16/11/22	Wednesday	Statistical Learning Theory and Machine Learning	Intro	2	2	
20	17/11/22	Thursday	Linear Models for Regression / Laboratory in R	M	3	2	1
21	22/11/22	Tuesday	Linear Models for Classification / Laboratory in R	M	3	2	1
22	23/11/22	Wednesday	Resampling Methods	I	2	2	
23	24/11/22	Thursday	Exploratory and Decision Trees / Laboratory in R	M	3	2	1
24	29/12/22	Tuesday	Exploratory and Decision Trees / Laboratory in R	M	3	2	1
25	30/11/22	Wednesday	Ensemble Methods	M	2	2	
26	01/12/22	Thursday	Laboratory in R	M	3		3
27	06/12/22	Tuesday	Linear Model Selection and Regularization / Laboratory in R	M	3	3	
28	07/12/22	Wednesday	Methods for Non-Linearity	M	2	2	
29	13/12/22	Tuesday	Methods for Non Linearity/ Laboratory in R	M	3	2	1
30	14/12/22	Wednesday	Support Vector Machines	M	2	2	
31	15/12/22	Thursday	Laboratory in R	M	3		3
32	20/12/22	Tuesday	Project Work for the Exam / Laboratory in R	*PW*	3	2	1
33	21/01/22	Wednesday	Laboratory in R	*PW*	2		2
34	11/01/23	Tuesday	Laboratory in R	*PW*	3		3
35	12/01/23	Wednesday	Learning by doing	*M*	2	2	
36	13/01/23	Thursday	Learning by doing	*PW*	4	2	2
					48	29	19
					96	63	33

Lectures Road Map

1. Statistics, *Technè-Logia*, Analysis of Data
2. Data Cleansing and Data Pre-Processing
3. Bivariate Data Analysis
4. Probability Models and Inference
5. Linear Models: ANOVA and Simple Regression
6. Linear Models for Time Series Analysis
7. Multidimensional Data Analysis: PCA, Cluster Analysis
8. Statistical Learning in practice
9. Linear Methods for Regression: Multiple Linear Regression
10. Linear Methods for Classification: Discriminant Analysis and Logistic Regression
11. Resampling Methods and Computational Inference
12. Tree-based Methods and Ensembles
13. Linear Model Selection and Regularization
14. Methods to deal with Non-Linearity
15. Support Vector Machines
16. Statistical Learning Theory and Recap of Learning Machines
17. Project Work
18. Exam

Examination/Evaluation Criteria

Exam

- The exam consists in developing a Project Work on a Real-World Case Study, submitting a technical report (one week earlier the exam session), presenting the quantitative story telling of the results at the exam session.
- Guidelines to develop and prepare the Technical Report are given in this presentation.
- Exam sessions will be communicated on the Team of the Course as well as on the Lecturer Web site. https://www.docenti.unina.it/.

Evaluation pattern

- The evaluation will consider the methodological knowledge, the computational aspects (code), the presentation and communication of the results (technical report and oral presentation with final discussion).

Evaluation grid

- The attribution of the vote takes place according to the criteria shown in the Table.

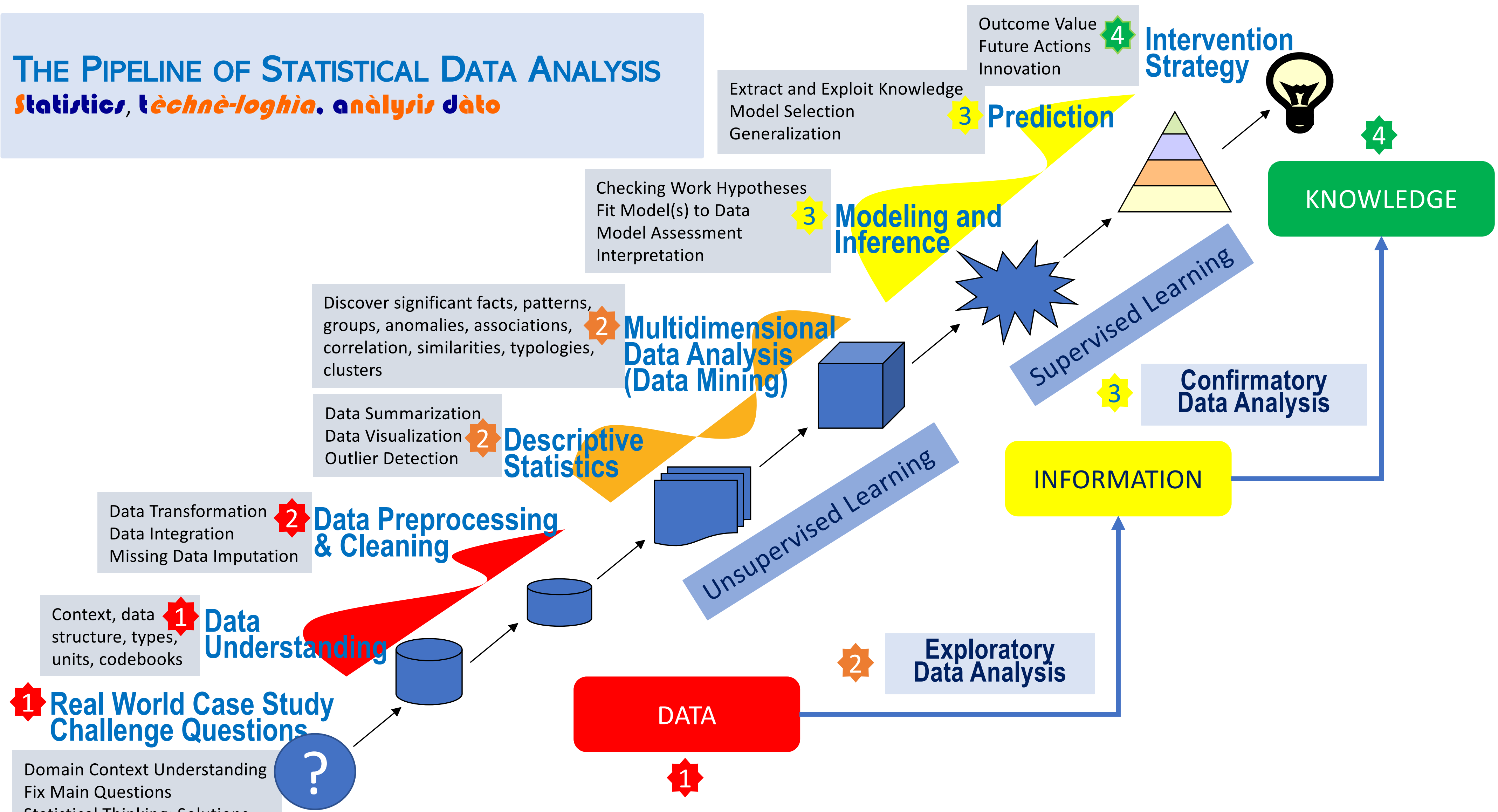
<18 **Not sufficient**	**Fragmented and superficial knowledge of the contents, errors in applying the concepts, insufficient written test and insufficient exposure**
18-20	Sufficient but general knowledge of the contents, simple exposition, uncertainties in the application of theoretical concepts
21-23	Appropriate but not in-depth knowledge of contents, ability to apply theoretical concepts, ability to present contents in a simple way
24-25	Appropriate and broad knowledge of contents, fair ability to apply knowledge, ability to present contents in an articulated way
26-27	Precise and complete knowledge of contents, good ability to apply knowledge, analytical skills, clear and correct presentation
28-29	Broad, complete, and in-depth knowledge of contents, good application of contents, good ability to analyze and synthesize, safe and correct exposure
30 **30 et lauda**	Very broad, complete, and in-depth knowledge of contents, well-established ability to apply contents, excellent ability to analyze, synthesize and interdisciplinary connections, mastery of exposure

How to pass the Exam?

- Build up a Team for the Project
- Select a Real-World Case Study Challenge
- Develop the Project Work following the Pipeline of Statistical Data Analysis
- Prepare the Technical Report using Markdown and submit it one week before the exam session
- Prepare a presentation of the results and do the exam, taking into account the examination/evaluation grid

Case studies

- Kaggle Repository
  https://www.kaggle.com/datasets
- UCI Machine Learning Repository
  https://archive.ics.uci.edu/ml/index.php
- A general entry point:
  https://datasetsearch.research.google.com



THE PIPELINE OF STATISTICAL DATA ANALYSIS
Statistics, tèchnè-loghìa, anàlysis dàto
Outcome Value
Future Actions
Innovation
4
Intervention Strategy
Extract and Exploit Knowledge
Model Selection
Generalization
3
Prediction
4
KNOWLEDGE
Checking Work Hypotheses
Fit Model(s) to Data
Model Assessment
Interpretation
3
Modeling and Inference
Discover significant facts, patterns, groups, anomalies, associations, correlation, similarities, typologies, clusters
2
Multidimensional Data Analysis (Data Mining)
Supervised Learning
3
Confirmatory Data Analysis
Data Summarization
Data Visualization
Outlier Detection
2
Descriptive Statistics
INFORMATION
Unsupervised Learning
Data Transformation
Data Integration
Missing Data Imputation
2
Data Preprocessing & Cleaning
Context, data structure, types, units, codebooks
1
Data Understanding
2
Exploratory Data Analysis
1
Real World Case Study Challenge Questions
DATA
Domain Context Understanding
Fix Main Questions
Statistical Thinking: Solutions
?
1

Technical Report: *sessions and contents structure*

1 Case Study Challenge

Introduce the Domain Context and the Real-World Case Study: **domain understanding** and which **challenging questions to satisfy**.

Refer to **the link of the databases**.

Data Understanding: describe the context, the data structure, the typology of variables, the units, time/space, etc.

Statistical thinking: convert questions into statistical problems, summarize and motivate which statistical methods to apply.

2 Exploratory Data Analysis

Data Pre-Processing and Cleaning: data transformation, data integration, missing data imputation.

Descriptive Statistics: data summarization, data visualization outlier detection.

Multidimensional Data Analysis (basic Data Mining strategy): dimensionality data reduction and clustering.

Statistical thinking: Summarize the relevant information derived from the Exploratory Data Analysis.

3 Confirmatory Data Analysis

Modeling and Inference: Checking Work Hypotheses, Fit Model(s) to Data, Model Assessment, Interpretation.

Prediction: Extract and Exploit Knowledge, Model Selection and Generalization.

Statistical thinking: Summarize the knowledge extraction due to the Confirmatory Data Analysis.

4 Outcome and concluding remarks

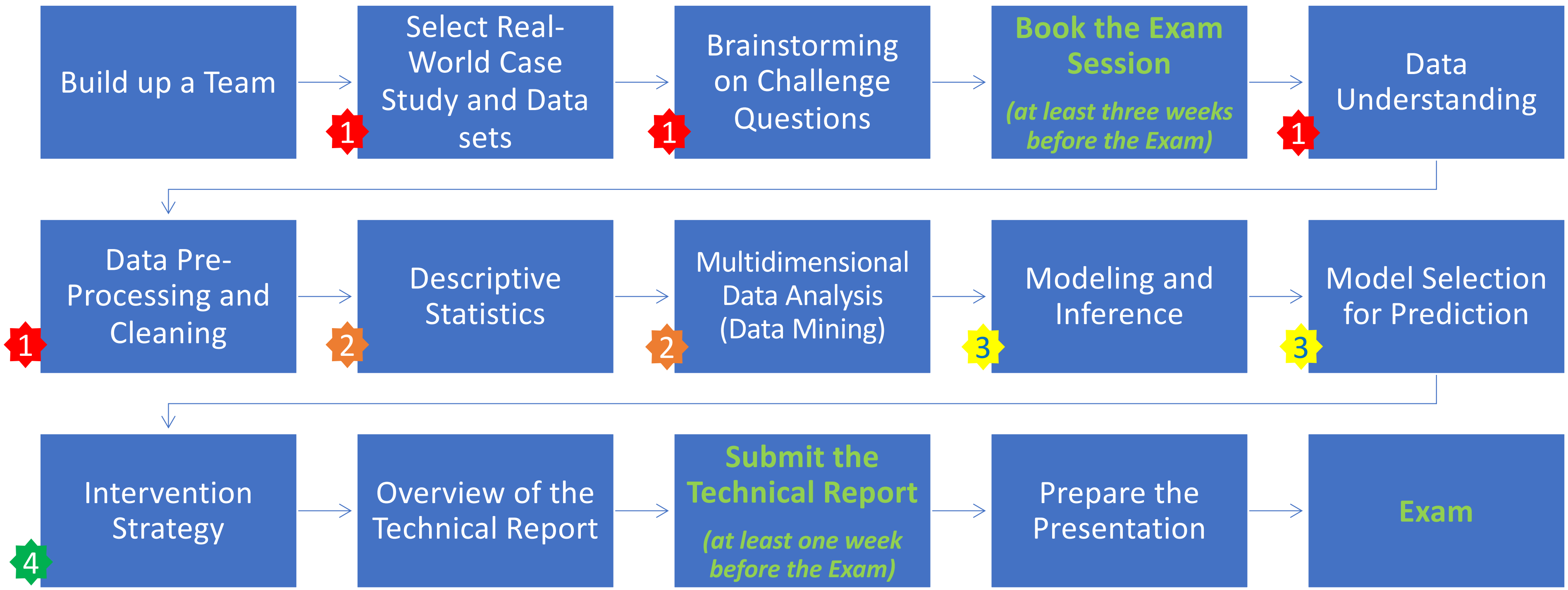
Summarize the key points of the outcome of the Statistical Data Analysis.

Intervention Strategy: Outcome Value, Future Actions, Innovation, how the beneficiary/stakeholders utilize the outcome of statistical data analysis.

How to organize the Project Work and Exam

- **Build up a team of max 5 members** **willing to sit for the same date**, possibly with different bachelor degree. You can also do the project work alone. The team does not necessarily correspond to those for the homework.
- **Select the Real-World Case Study** with one or more data sets from one of the Public Repositories and **define a set of challenging real-world questions.**
- **Team brainstorming** to discuss how to convert the real-world questions into statistical questions and methodologies to apply. You can also discuss with the Professors at the ***Weekly Question Time Planned Meeting*** receiving important suggestions.
- **Book the Exam *three weeks before the exam session* using the form linked in the exam channel of the Team of the Course.**
- **Run your work and share the tasks in the team following the PIPELINE OF STATISTICAL DATA ANALYSIS.**
- **Write the Technical Report of your project using Markdown** and considering the following points: *introduce the case study challenge, describe well the data, justify the selected methods and R packages, comment R code, provide the correct interpretation of all output, derive the outcomes of the statistical data analysis referred to the challenge questions.*
- **Submit your Technical Report *one week before the exam session***.
- **Prepare the Quantitative-Story Telling of your project work results using the ppt presentation:** focalize the attention on how to communicate the results in interesting way for the beneficiary of the statistical data analysis.
- **Present yourself at the Exam Session**: be ready to answer methodological questions and to discuss about the outcome of your statistical data analysis, be convinced about the good results that you have found but also be critical toward yourself, what you could have done better, what are potential future developments.
- **Mind that Your exam will be evaluated according to the published examination/evaluation criteria.**

Road Map for the Project Work and Exam



Calendar of Exam Sessions and **Deadlines**
(January-February-March 2023)

31 Jan: Exam

Deadlines
10 Jan: Book the Exam
24 Jan: Submit the Technical Report

7 Feb: Exam

Deadlines
17 Jan: Book the Exam
31 Jan: Submit the Technical Report

14 Feb: Exam

Deadlines
24 Jan: Book the Exam
7 Feb: Submit the Technical Report

28 Feb: Exam

Deadlines
7 Feb: Book the Exam
21 Feb: Submit the Technical Report

7 Mar: Exam

Deadlines
14 Feb: Book the Exam
28 Feb: Submit the Technical Report

14 Mar: Exam

Deadlines
21 Feb: Book the Exam
7 Mar: Submit the Technical Report

28 Mar: Exam

Deadlines
7 Mar: Book the Exam
21 Mar: Submit the Technical Report