UNIVERSITÀ DEGLI STUDI DI PALERMO

DEPARTMENT	Scienze Economiche, Aziendali e Statistiche
ACADEMIC YEAR	2021/2022
MASTER'S DEGREE (MSC)	STATISTICS AND DATA SCIENCE
SUBJECT	EXPLORATORY METHODS FOR BIG DATA
TYPE OF EDUCATIONAL ACTIVITY	B
AMBIT	50606-Statistico
CODE	18165
SCIENTIFIC SECTOR(S)	SECS-S/01
HEAD PROFESSOR(S)	PLAIA ANTONELLA Professore Ordinario Univ. di PALERMO
OTHER PROFESSOR(S)	
CREDITS	9
INDIVIDUAL STUDY (Hrs)	162
COURSE ACTIVITY (Hrs)	63
PROPAEDEUTICAL SUBJECTS	
MUTUALIZATION	
YEAR	1
TERM (SEMESTER)	1° semester
ATTENDANCE	Not mandatory
EVALUATION	Out of 30
TEACHER OFFICE HOURS	**PLAIA ANTONELLA** Wednesday 15:00 17:00 La modalita, in studio o su Teams, va concordata col docente

DOCENTE: Prof.ssa ANTONELLA PLAIA

PREREQUISITES	Basic notions of Descriptive Statistics and Probability Theory. R software
LEARNING OUTCOMES	Knowledge and understanding Knowledge of the methods and statistical procedures for exploratory analyzes of "big data" Applying knowledge and understanding The student must 'be able to: 1. formulate correctly a problem 2. choose appropriate statistical and computer analysis solutions, 3. translate the results into operational decisions. Making judgments The student must 'be able to: 1. translate into statistical terms a knowledge requirement arose in different application fields 2. apply the opportune cleaning activities, reorganization, descriptive analysis and interpretation of data, 3. process and communicate consistently the results of its analysis. Communication The student must be able to: 1. communicate clearly, both to specialists and non-specialists in the subject, the concepts and techniques of analysis applied, 2. justify the choice of the instruments used for the analysis, 3. communicate the results with appropriate language. Learning skills The student will have developed learning skills necessary to deepen independently the study of the statistical techniques more commonly used in the analysis of large data sets.
ASSESSMENT METHODS	Presentation of 2 reports, and oral exam. The exam is done through an oral interview, subject to the submission of - for the exams immediately after the class 1. a report assigned during the sixth week to groups of at most three students. 2. a report containing an analysis of a dataset assigned at the end of the course (to be delivered one week before the oral exam) - for the exams during other periods 1. a report assigned 3 weeks before the exam. 2. a report containing an analysis of a dataset assigned 3 weeks before the exam (to be delivered one week before the oral exam) Both the reports (no more than 4 pages, plus tables and plots) must be delivered, together with the corresponding R code. The oral test consists in presenting (power point or otheìher sw) the analyses and results shown in the final report in order to evaluate better knowledge, skills and abilities held by the student as well as his/her ability to provide them with a suitable statistical language. FINAL EVALUATION The final exam assessment will take into account three aspects: i) the mastery of the subjects; ii) the ability to application of knowledge and iii) the correct use of language, assessed in the intermediate and final report and the oral examination. The evaluation will be sufficient if the candidate has chosen the correct methods of cleaning and tidy of the data provided for both reports and identified the appropriate analysis tools, even if the analysis will not be efficient. Furthermore, for a sufficient evaluation, the candidate must demonstrate a sufficient argumentative and expository capacity during the oral exam. The more the candidate will give evidence, in the writing of the reports and in the oral test, of his argumentative and expositive abilities, as well as of properties of statistical language, and of efficient use of the programming language R for data analysis, the more the evaluation will be positive. The Selection Committee is formed by the chair of the class and at least one other teacher, professor or assistant professor, or an expert on the subject.
EDUCATIONAL OBJECTIVES	The course introduces statistical methods for the analisys of high dimensional data.The main objective is the analysis of large databases in order to find patterns, associations, changes, faults and structures of particular interest.At the

	end of the course the learner will be able to identify the most appropriate tools for the analysis to be played and apply them, summarizing the results effectively.
TEACHING METHODS	Lectures, laboratory
SUGGESTED BIBLIOGRAPHY	Dispense rese disponibili dal docente sul portale di Ateneo. Risorse on-line indicate dal docente durante il corso. Breiman, L. Friedman, J. H. Olshen, R. A. Stone, C. J. (1984) Classification and regression trees, Chapman & Hall. Capp. 1-5, 8 G. James, D. Witten, T. Hastie, R. Tibshirani . (2013) An Introduction to Statistical Learning, with applications in R. Springer. Cap. 8 Stef van Buuren, (2012) Flexible Imputation of Missing Data, Chapman & Hall, capp 1-4, 7.2

SYLLABUS

Hrs	Frontal teaching
4	Exploratory data analysis
4	Data integration
4	Data visualization
2	Data transformation
4	Web scraping
6	Missing data handling
12	Classification and clustering techniques

Hrs	Practice
4	Data integration
4	Data visualization
2	Web scraping
5	Missing data handling
10	Classification and clustering techniques
2	How to write a report and its conversion to a power point presentation