STUDIJŲ DALYKO (MODULIO) APRAŠAS

Dalyko (modulio) pavadinimas	Kodas
Duomenų tvarkyba ir tranformavimas R aplinkoje	

Dėstytojas (-ai)	Padalinys (-iai)
Koordinuojantis: Algimantas Birbilas **Kitas (-i):**	Statistinės analizės katedra

Studijų pakopa	Dalyko lygmuo	Dalyko (modulio) tipas
Pirmoji	Pradedančiųjų	Pasirenkamas

Įgyvendinimo forma	Vykdymo laikotarpis	Vykdymo kalba (-os)
Auditorinė	Septintas (rudens) semestras	Lietuvių

Reikalavimai studijuojančiajam	
Išankstiniai reikalavimai: darbo R aplinkoje pagrindai, anglų kalba B1 (arba aukštesniu) lygiu pagal įprastą Europoje galiojančią lygių sistemą.	**Gretutiniai reikalavimai (jei yra):**

Dalyko (modulio) apimtis kreditais	Visas studento darbo krūvis	Kontaktinio darbo valandos	Savarankiško darbo valandos
5	125	48	77

Dalyko (modulio) tikslas: studijų programos ugdomos kompetencijos

Studentai turėtų išugdyti šias kompetencijas

- gebės surasti reikiamą literatūrą, įsisavinti naujas žinias ir metodus bei taikyti juos praktiškai (1.1);
- gebės rinkti duomenis iš įvairių duomenų šaltinių, įvertinti duomenų patikimumą, klasifikuoti duomenis šaltinio, apimties, dažnumo ir srauto aspektu, sutvarkyti bei paruošti duomenis analizei (5.2);
- Gebėjimas rinktis tinkamą analizės metodologiją bei jai reikalingus įrankius (6);
- Gebėjimas interpretuoti ir reprezentuoti analizės rezultatus (7).

Dalyko (modulio) studijų siekiniai (išklausę modulį studentai turėtų):	Studijų metodai	Vertinimo metodai
• pagilinti žinias apie pagrindinius R duomenų tipus bei duomenų struktūras; • patobulinti programavimo R aplinkoje įgūdžius; • detaliai susipažinti su R paketu tidyverse; • gebėti efektyviai atrinkti bei paruošti duomenis tolimesnei duomenų analizei; • gebėti skaityti literatūrą, skirtą duomenų tvarkybai.	Paskaitos, pratybos naudojant programinę įrangą, savarankiškas namų darbams skirtų užduočių sprendimas	Kontroliniai darbai

Temos	Kontaktinio darbo valandos	Savarankiškų studijų laikas ir užduotys

	Paskaitos	Konsultacijos	Seminarai	Pratybos	Laboratoriniai darbai	Praktika	**Visas kontaktinis darbas**	**Savarankiškas darbas**	**Užduotys**
1. Bazinis R. RStudio. Pagrindiniai duomenų tipai (char, logical, numeric, string, factor, date-time) bei struktūros (vector, list, dataframe). Atributai. Poaibių išrinkimas. Funkcijos.	2				6		**8**	**8**	Išspręsti dėstytojo nurodytus [2] šaltinio 2, 4, 5, 6 poskyrių pratimus
2. Manipuliavimas. Tibble duomenų struktūra**.** Pipe operatorius. Duomenų išrinkimas, filtravimas bei kintamųjų kūrimas pasitelkiant paketą dplyr.	2				6		**8**	**8**	Išspręsti dėstytojo nurodytus [1] šaltinio 5 ir 10 poskyrių pratimus
3. Tekstiniai duomenys. Reguliarios išraiškos: sudarymo principai, panaudojimas tekstinių duomenų filtravimui, paketas stringr.	2				6		**8**	**8**	Išspręsti dėstytojo nurodytus [1] šaltinio 14 poskyrio pratimus
4. Kategoriniai duomenys. Kategorinių kintamųjų kūrimas. Kategorijų rikiavimas, transformavimas bei agregavimas.	1				3		**4**	**6**	Išspręsti dėstytojo nurodytus [1] šaltinio 15 poskyrio pratimus
5. Data ir laikas. Datos tipo kintamųjų kūrimas. Laiko juostos, pateikimo formatai. Manipuliavimas: aritmetiniai veiksmai, datos ir/arba laiko dalies išskyrimas.	2				6		**8**	**8**	Išspręsti dėstytojo nurodytus [1] šaltinio 16 poskyrio pratimus
6. Duomenų importas. Tekstinių dumenų importas pasitelkiant paketą readr. Pradinės žinios apie JSON bei XML duomenų formatus.	1				3		**4**	**4**	Išspręsti dėstytojo nurodytus [1] šaltinio 11 poskyrio pratimus
7. Struktūrizuoti duomenys. Manipuliavimas reliaciniais duomenimis dplyr paketo aplinkoje. Duomenų struktūros formavimas panaudojant tidyr paketą.	2				6		**8**	**8**	Išspręsti dėstytojo nurodytus [1] šaltinio 12 ir 13 poskyrių pratimus
8. Pasiruošimas atsiskaitymams.								**25**	Pasiruošti atsiskaitymams.
Iš viso	**12**				**36**		**48**	**77**	

Vertinimo strategija	Svoris proc.	Atsiskaitymo laikas	Vertinimo kriterijai
1 kontrolinis darbas	20	trečia studijų savaitė	Kontrolinį sudaro kelios užduotys, skirtos patikrinti žinių lygiui. Bendra užduočių vertė – 2 balai. Kiekvienos užduoties vertė svyruoja nuo 0,1 iki 1 balo. Užduotys atliekamos raštu arba prie kompiuterio. Atskiros užduoties vertinimo principai: a) išskiriamos dalys, už kurias skiriama dalis visos užduoties taškų; b) atlikus atitinkamą dalį be klaidų už ją skiriamas maksimalus taškų skaičius, priešingu atveju taškų skaičius mažinamas atsižvelgiant į padarytas klaidas; c) klaidingas kažkurios dalies atlikimas neturi įtakos kitų dalių vertinimui.
2 kontrolinis darbas	20	šešta studijų savaitė	
3 kontrolinis darbas	20	devinta studijų savaitė	
4 kontrolinis darbas	20	dvylikta studijų savaitė	
5 kontrolinis darbas	20	penkiolikta studijų savaitė	
Kontrolinių perrašymas	≤40	Šešiolikta studijų savaitė, egzaminas	Kiekvienas studentas turi teisę perrašyti bet kuriuos du iš penkių rašytų kontrolinių. Perrašymui skiriama šešiolikta studijų savaitė bei egzamino laikas. Perrašius prasčiau, paliekamas geresnis balas.
Eksterno tvarka			Eksternu dalyko studijuoti negalima.

Autorius	Leidimo metai	Pavadinimas	Periodinio leidinio Nr. ar leidinio tomas	Leidimo vieta ir leidykla ar internetinė nuoroda
Privaloma literatūra				
1. Garrett Grolemund, Hadley Wickham	2016	R for Data Science		O'Reilly (prieiga internete https://r4ds.had.co.nz/)
2. Hadley Wickham	2014	Advanced R		Chapman & Hall (prieiga internete https://adv-r.hadley.nz/)
3. Jeffrey B. Arnold	2019	R for Data Science Solutions Manual		O'Reilly (prieiga internete https://jrnold.github.io/r4ds-exercise-solutions/)
4. W. N. Venables, D. M. Smith and the R Core Team	2018	An introduction to R (Notes on R: A Programming Environment for Data Analysis and Graphics)		https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
Papildoma literatūra				

COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Data preprocessing and transformation using R	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: associate professor V. Skorniakov **Other(s):**	Department of Statistical Analysis

Study cycle	Level of course	Type of the course unit (module)
First	Beginners	Compulsory

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face	Seventh (autumn) semester	English

Requirements for students	
Prerequisites: basics of R; ability to understand English at the level of independent user (B1 according to CEFR classification).	**Additional requirements (if any):**

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	125	48	77

Purpose of the course unit (module): programme competences to be developed (the number in the brackets coincides with that given in the official description of the programme)		
Students should develop the following competencies: • be able to find the necessary literature, acquire new knowledge and methods, and apply them in practice (1.1); • be able to collect data from various data sources, assess data reliability, classify data in terms of source, volume, frequency, and flow, organize and prepare data for analysis (5.2); • ability to choose the appropriate analysis methodology and the tools required for it (6); • • Ability to interpret and represent the results of the analysis (7).		
Learning outcomes of the course unit (module); after completing the course, students should:	**Teaching and learning methods**	**Assessment methods**
• extend their knowledge regarding base R data types and data structures; • enhance R programming skills; • get familiar with R package tidyverse in detail; • be able to pre-process data for further analysis effectively; • be able to read specialized literature devoted to data pre-processing.	Lectures, class works by making use of software, homework	Tests

Content: breakdown of the topics	Contact hours	Self-study work: time and assignments

	Lectures	Tutorials	Seminars	Exercises	Laboratory work	Internship/work placement	**Contact hours**	**Self-study hours**	**Assignments**
1. Base R. RStudio. Main data types (char, logical, numeric, string, factor, date-time) and data structures (vector, list, dataframe). Attributes. Subsetting. Functions.	2				6		**8**	**8**	Solve assigned exercises from reference [2], ch. 2, 4, 5 and 6
2. Wrangling. Tibbles**.** Piping. Data selection, filtering and creation of new variables by using dplyr.	2				6		**8**	**8**	Solve assigned exercises from reference [1], ch. 5 and 10
3. Text data. Regular expressions: rules, filtering, package stringr.	2				6		**8**	**8**	Solve assigned exercises from reference [1], ch. 14
4. Factors. Creation, sorting, aggregation and transformation.	1				3		**4**	**6**	Solve assigned exercises from reference [1], ch. 15
5. Date and time. Variable **c**reation. Time zones and formats. Wrangling: arithmetic, part extraction.	2				6		**8**	**8**	Solve assigned exercises from reference [1], ch. 16
6. Data import. Import of CSV files by making use of package readr. Introduction to JSON and XML data formats.	1				3		**4**	**4**	Solve assigned exercises from reference [1], ch. 11
7. Structured data. Working with relational data by making use of package dplyr. Data transformations with the help of tidyr package.	2				6		**8**	**8**	Solve assigned exercises from reference [1], ch. 12 and 13
8. Preparation for assessements.								**25**	Prepare for tests.
Total	**12**				**36**		**48**	**77**	

Assessment strategy	**Weight, %**	**Deadline**	**Assessment criteria**
Test 1	20	3rd study week	The test consists of several practical tasks intended to check the level of knowledge obtained. The total weight of these tasks equals to 2 points. The weight of each task ranges from 0.1 to 1 point. Tasks are designed to be solved in a written form or by using a computer and appropriate software. Each task is evaluated as follows: a) the task is divided into parts, and each part is assigned an appropriate number of points; b) if the student accomplishes the part without mistakes, the whole amount of that part is attained; otherwise, the amount is reduced considering the mistakes made; c) the parts are evaluated independently.
Test 2	20	6th study week	
Test 3	20	9th study week	
Test 4	20	12th study week	
Test 5	20	15th study week	
Rewriting of tests	≤40	16th study week, exam	Each student is allowed to rewrite two tests optionally. Rewriting takes place during the 16th study week and time devoted to the exam. If the rewritten test's outcome is worse than that of the initial one, the grade remains unchanged.
External order			The subject cannot be studied externally.

Author	Year of publication	Title	Issue of a periodical or volume of a publication	Publishing place and house or weblink
Compulsory reading				
1. Garrett Grolemund, Hadley Wickham	2016	R for Data Science		O'Reilly (available online https://r4ds.had.co.nz/)
2. Hadley Wickham	2014	Advanced R		Chapman & Hall (available online https://adv-r.hadley.nz/)
3. Jeffrey B. Arnold	2019	R for Data Science Solutions Manual		O'Reilly (available online https://jrnold.github.io/r4ds-exercise-solutions/)
4. W. N. Venables, D. M. Smith and the R Core Team	2018	An introduction to R (Notes on R: A Programming Environment for Data Analysis and Graphics)		https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
Optional reading				