БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям
_____ О.Г. Прохоренко
12 июня 2023 г.

Регистрационный № УД-1378/м.

Visualization methods in data analysis by R

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

7-06-0533-05 Applied Mathematics and Computer Science

Profiling:
Computer Data Analysis

2023 г.

Учебная программа составлена на основе ОСВО 7-06-0533-05-2023, учебного плана № М53а-5.3-115/уч. от 11.04.2023

СОСТАВИТЕЛИ:
В.В. Мушко, доцент кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета, кандидат физико-математических наук

РЕЦЕНЗЕНТЫ:
Б.А. Залесский, заведующий лабораторией обработки и распознавания изображений ГНУ «Объединённый институт проблем информатики Национальной академии наук Беларуси», доктор физико-математических наук
М.С. Абрамович, заведующий НИЛ статистического анализа и моделирования Учреждения БГУ «Научно-исследовательский институт прикладных проблем математики и информатики» кандидат физико-математических наук, доцент

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:
Кафедрой теории вероятностей и математической статистики БГУ
(протокол № 12 от 23 мая 2023 года);

Научно-методическим советом БГУ
(протокол № 8 от 31 мая 2023 года)

Заведующий кафедрой
теории вероятностей и математической статистики
_____ А.Ю. Харин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная дисциплина «Visualization methods in data analysis by R» преподается на английском языке для студентов углубленной формы высшего образования (магистрантов) специальности *7-06-0533-05 Applied Mathematics and Computer Science*.

Цели и задачи учебной дисциплины

Цель учебной дисциплины – ознакомление студентов магистратуры с основами визуализации данных, а также основными возможностями пакета *ggplot2*, расширяющего базовые графические возможности системы статистических вычислений языка *R*.

Задачи учебной дисциплины:

1. Изучение основных подходов и методов графического анализа данных и способов визуализации данных;

2. Формирование практических умений и навыков работы с пакетом *ggplot2* в рамках *R* и *RStudio*.

Место учебной дисциплины в системе подготовки специалиста с углубленным высшим образованием (магистра).

Учебная дисциплина относится к государственному компоненту и входит **в модуль** «Special software».

Дисциплина «Visualization methods in data analysis by R» согласуется с дисциплиной государственного компонента «Data analysis software» и способствует успешному изучению дисциплин «Multivariate Statistical Analysis» модуля «Methods and computer tools of applied mathematics», «Methods for statistical analysis of complex data» модуля «Specific methods of analysis», «Visual analytics of time-oriented data» и «Infographic and data analysis results presentation» модуля «Visual analytics and infographic», прохождению практики и написанию магистерских диссертаций.

Требования к компетенциям

Освоение учебной дисциплины «Visualization methods in data analysis by R» должно обеспечить формирование следующих компетенций:

универсальные компетенции (UC):

UC-2. To solve scientific and innovative problems on the base of information and communication technologies basis.

UC-5. To develop the innovative acceptability and the ability for innovative activity.

углубленные профессиональные компетенции (UPC):

UPC-6. To apply abilities to use fundamental data visualization methods with *R*.

В результате освоения учебной дисциплины магистрант должен

знать:

- основные целевые функции процесса визуализации данных;
- основные подходы к визуализации данных;
- основные виды графиков, области их применения и типичные ограничения для каждого графического инструмента;
- особенности анализа и визуализации конкретных типов данных;
- типовые способы визуально эффективного представления результатов исследования
- потенциальные ошибки, возможные при визуализации данных;
- основные графические возможности пакета *ggplot2* языка статистических вычислений *R*;

уметь:

- использовать различные методы визуализации данных для подготовки отчетов о результатах анализа;
- выбирать адекватные, оптимальные инструменты визуализации данных для эффективной поддержки принятия решений;
- эффективно использовать инструментарий программного обеспечения наук о данных — пакет *ggplot2* среды *R* — для генерации визуальных представлений данных;

владеть:

- теоретическими знаниями базовых концепций и типовых практических инструментов, необходимых для анализа и визуализации больших данных;
- техническими навыками выбора адекватных инструментов эффективного графического анализа данных различных типов;
- умениями пользоваться инструментальной базой на практике.

Структура учебной дисциплины

Дисциплина изучается в первом семестре. Всего на изучение учебной дисциплины «Visualization methods in data analysis by R» отведено:

– для очной формы получения углубленного высшего образования – 106 часов, в том числе 50 аудиторных часов, из них: лекции – 20 часов, семинарские занятия – 10 часов, лабораторные занятия – 20 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации – экзамен.

TEACHING MATERIAL CONTENTS

Section 1. Introduction

Topic 1.1. The *tidyverse* collection of *R* packages for data science

The *R* language for statistical computing and graphics. The *RStudio* integrated development environment. The *tidyverse* collection of *R* packages for data science.

Topic 1.2. The *tidyverse* style guide

The *tidyverse* style guide. Automatic code formatting using the *styler* package. Automatic code checking for style guide compliance using the *lintr* package.

Section 2. The *ggplot2* package. Plot fundamentals in *ggplot2*

Topic 2.1. The *ggplot2* package

The *ggplot2* package. Overview. Installation. Lifecycle. Ecosystem of extensions. Learning *ggplot2*. The *plotly* package for creating interactive graphics.

Topic 2.2. Plot fundamentals in *ggplot2*

The *ggplot()*, *aes()*, *`+`(<gg>)*, *`%+%`*, *ggsave()* functions and their arguments. Layers. Geometric objects. Statistical transformations. Position adjustments. Annotations. Aesthetics. Scales. Axes and legends. Facets. Coordinate systems. Themes.

Topic 2.3. Visualization and recovery (imputation) of missing values

The *naniar*, *VIM* packages for visualization and recovery (imputation) of missing values.

Topic 2.4. Color palettes. Color blindness simulators

Collection of color palettes of *paletteer* package. Color blindness simulators of *colorBlindness*, *colorblindr* packages.

Section 3. Automation of reporting

Topic 3.1. Automated graphical exploratory data analysis

The *dlookr*, *brinton* packages for automated graphical exploratory data analysis.

Topic 3.2. *Quarto* system for scientific and technical publishing

An open-source *Quarto* system for scientific and technical publishing.

Section 4. Graphing of a variable (probability) distribution

Topic 4.1. Graphing of a continuous variable distribution

Continuous variable. Features. Basic methods for graphing of a continuous variable (probability) distribution. Alternative methods for graphing of a continuous variable distribution. Plot options.

Topic 4.2. Graphing of a categorical variable distribution

Categorical variable. Nominal variable, ordinal variable, discrete variable. Features. Basic methods for graphing of a categorical variable (probability) distribution. Alternative methods for graphing of a categorical variable distribution. Plot options.

Section 5. Graphing of multivariate data

Topic 5.1. Graphing of multivariate continuous data

Multivariate continuous data. Features. Basic methods for graphing of multivariate continuous data. Alternative methods for graphing of multivariate continuous data. Plot options.

Topic 5.2. Graphing multivariate categorical data

Multivariate categorical data. Features. Basic methods for graphing multivariate categorical data. Alternative methods for graphing multivariate categorical data. Plot options.

Section 6. Time series graphing

Topic 6.1. Time series graphing

Time series. Features. Basic methods for time series graphing. Alternative methods for time series graphing. Plot options.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Очная форма получения углубленного высшего образования с применением дистанционных образовательных технологий (ДОТ)

Numbering	Section, topic	In-class hours					Supervised self-study, hours	Knowledge test form
		Lectures	Practical classes	Seminars	Labs	Other		
1	2	3	4	5	6	7	8	9
1	**Introduction**	**2**		**2**	**2**			
1.1	The *tidyverse* collection of *R* packages for data science	1		1	1			oral test written report on in-class practical exercises
1.2	The *tidyverse* style guide	1		1	1			oral test written report on in-class practical exercises
2	**The *ggplot2* package. Plot fundamentals in *ggplot2***	**6**		**6**	**6**			
2.1	The *ggplot2* package	2		2	2			oral test test № 1 test № 2
2.2	Plot fundamentals in *ggplot2*	2		2	2			oral test test № 3 test № 4

2.3	Visualization and recovery (imputation) of missing values	1		1	1			written report on in-class practical exercises
2.4	Color palettes. Color blindness simulator	1		1	1			written report on in-class practical exercises
3	**Automation of reporting**	**2**		**2**	**2**			
3.1	Automated graphical exploratory data analysis	1		1	1			colloquium
3.2	*Quarto* system for scientific and technical publishing	1		1	1			written report on in-class practical exercises
4	**Graphing of a variable (probability) distribution**	**4**			**4**			
4.1	Graphing of a continuous variable distribution	2			2			written report on home practical exercises with their oral defense
4.2	Graphing of a categorical variable distribution	2			2			written report on home practical exercises with their oral defense
5	**Graphing of multivariate data**	**4**			**4**			
5.1	Graphing of multivariate continuous data	2			2			written report on home practical exercises with their oral defense
5.2	Graphing multivariate categorical data	2			2			written report on home practical exercises with their oral defense
6	**Time series graphing**	**2**			**2**			
6.1	Time series graphing	2			2			written report on home practical exercises with their oral defense

								portfolio
TOTAL		**20**		**10**	**20**			

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Principal textbooks

1. Брюс, П. Практическая статистика для специалистов Data Science: 50+ важнейших понятий с использованием R и Python / П. Брюс, Э. Броюс, П. Гедек; [пер. с англ. А. Логунова]. - 2-е изд., перераб. и доп. - Санкт-Петербург: БХВ-Петербург, 2021. - 346 с. - URL: https://ibooks.ru/reading.php?short=1&productid=380029.
2. Лонг, Дж. Д. R. Книга рецептов: проверенные рецепты для статистики, анализа и визуализации / Дж. Д. Лонг, Пол Титор ; [пер. с англ. Д. А. Беликова]. - Москва : ДМК Пресс, 2020. - 508 с. - URL: https://ibooks.ru/reading.php?short=1&productid=387334.
3. Ланц, Б. Машинное обучение на R = Machine Learning with R : экспертные техники для прогностического анализа / Б. Ланц ; [пер. с англ. Е. Сандицкой]. - Санкт-Петербург [и др.] : Питер, 2020. - 462 с. - URL: https://ibooks.ru/bookshelf/367984.
4. Ын, А. Теоретический минимум по Big Data. Всё, что нужно знать о больших данных / Анналин Ын, Кеннет Су ; [пер. с англ. А. Тимохина]. - Санкт-Петербург [и др.] : Питер, 2021. - 205 с. - URL: https://ibooks.ru/reading.php?short=1&productid=359225.

Optional textbooks

1. Bertin J. Semiology of Graphics: Diagrams, Networks, Maps. — Esri Press, 2010. — 456 p.

2. Chen C, Hardle W., Unwin A. Handbook of Data Visualization. Springer Handbooks of Computational Statistics. — Springer-Verlag Berlin Heidelberg, 2008. — 936 p.

3. Gerbing D. R Visualizations. Derive Meaning from Data. — Chapman & Hall/CRC, 2020. — 249 p.

4. Kabacoff R. Modern Data Visualization with R. — Chapman & Hall/CRC The R Series, 2024. — 256 p.

5. Unwin A. Graphical Data Analysis with R. The R Series. — Chapman & Hall/CRC, 2018. — 310 p.

6. Unwin A., Theus M., Hofmann H. Graphics of Large Datasets: Visualizing a Million. — Springer-Verlag New York, 2006. — 275 p.

7. Wickham H. Advanced R. — Chapman & Hall/CRC The R Series, 2019. — 588 p.

8. Wickham H., Çetinkaya-Rundel M., Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, 2nd Edition. — O'Reilly Media, 2023. — 576 p.

9. Wickham H. ggplot2: Elegant Graphics for Data Analysis, 3rd Edition. — Springer International Publishing, 2024. — 260 p.

10. Wilkinson L. The Grammar of Graphics. Statistics and Computing, 2nd Edition. — Springer-Verlag New York, 2005. — 691 p.

Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

Объектом диагностики компетенций магистрантов являются знания, умения, полученные ими в результате изучения учебной дисциплины. Выявление учебных достижений магистрантов осуществляется с помощью мероприятий текущего контроля и промежуточной аттестации.

Для диагностики компетенций используются следующие формы:

1. Устная форма: опрос;

2. Письменная форма: письменные отчеты по аудиторным практическим упражнениям, коллоквиум, контрольные работы, портфолио работ;

3. Устно-письменная форма: письменные отчеты по домашним практическим упражнениям с их устной защитой.

Формой промежуточной аттестации по дисциплине «Visualization methods in data analysis by R» учебным планом предусмотрен **экзамен.**

При формировании итоговой отметки используется рейтинговая система оценки знаний магистранта, дающая возможность проследить и оценить динамику процесса достижения целей обучения.

Рейтинговая система предусматривает использование весовых коэффициентов в ходе проведения контрольных мероприятий текущей аттестации.

Примерные весовые коэффициенты, определяющие вклад текущей аттестации в отметку при прохождении промежуточной аттестации:

Формирование отметки за текущую аттестацию:

- коллоквиум – 15 %;
- выполнение контрольных работ – 35 %;
- подготовка письменных отчетов по домашним практическим упражнениям – 35 %;
- подготовка портфолио работ – 15 %.

Итоговая отметка по дисциплине рассчитывается на основе отметки текущей аттестации – (рейтинговой системы оценки знаний) и экзаменационной отметки с учетом их весовых коэффициентов. Вес отметки по текущей аттестации составляет 40 %, экзаменационной отметки – 60 %.

Sample list of topics for seminars

Seminar № 1. The *tidyverse* collection of *R* packages for data science. The *tidyverse* style guide.

Seminar № 2. The *ggplot2* package.

Seminar № 3. Plot fundamentals in *ggplot2*.

Seminar № 4. Visualization and recovery (imputation) of missing values. Color palettes. Color blindness simulators.

Seminar № 5. Automated graphical exploratory data analysis. *Quarto* system for scientific and technical publishing.

Sample list of topics for labs

Class № 1. The *tidyverse* collection of *R* packages for data science. The *tidyverse* style guide.

Class № 2. The *ggplot2* package.

Class № 3. Plot fundamentals in *ggplot2*.

Class № 4. Visualization and recovery (imputation) of missing values. Color palettes. Color blindness simulators.

Class № 5. Automated graphical exploratory data analysis. *Quarto* system for scientific and technical publishing.

Class № 6. Graphing of a continuous variable distribution.

Class № 7. Graphing of a categorical variable distribution.

Class № 8. Graphing of multivariate continuous data.

Class № 9. Graphing multivariate categorical data.

Class № 10. Time series graphing.

Sample list of test topics

Test № 1. Scales.

Test № 2. Markers and visual channels.

Test № 3. Geometric objects.

Test № 4. Layers.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется ***практико-ориентированный подход,*** который предполагает:

- освоение содержание образования через решения практических задач;

- приобретение навыков эффективного выполнения разных видов профессиональной деятельности;

- ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;

- использованию процедур, способов оценивания, фиксирующих сформированность профессиональных компетенций.

Также при организации образовательного процесса ***используется метод портфолио*****,** который является эффективным средством реализации индивидуальной образовательной программы обучающихся. Все результаты и достижения группируются вокруг основных видов деятельности обучающихся: учебной, научно-исследовательской и иной.

Также при организации образовательного процесса ***используется метод группового обучения,*** который представляет собой форму организации учебно-познавательной деятельности обучающихся, предполагающую функционирование разных типов малых групп, работающих как над общими, так и специфическими учебными заданиями.

Методические рекомендации по организации самостоятельной работы обучающихся

При изучении учебной дисциплины рекомендуется использовать следующие формы самостоятельной работы:

– поиск (подбор) и обзор литературы и электронных источников по индивидуально заданной проблеме дисциплины;

– выполнение домашнего задания;

– работы, предусматривающие решение задач и выполнение упражнений, выдаваемых на лабораторных занятиях;

– изучение материала, вынесенного на самостоятельную проработку;

– подготовка к семинарским занятиям;

– подготовка к экзамену;

– научно-исследовательские работы;

– анализ материалов по заданной теме, проведение расчетов, составление схем и моделей на основе статистических материалов.

Sample list of questions for the examination

1. The *tidyverse* collection of *R* packages for data science.
2. The *tidyverse* style guide.
3. The *ggplot2* package.
4. Plot fundamentals in *ggplot2*.
5. Visualization and recovery (imputation) of missing values.
6. Color palettes. Color blindness simulators.

7. Automated graphical exploratory data analysis.
8. *Quarto* system for scientific and technical publishing.
9. Basic methods for graphing of a continuous variable (probability) distribution.
10. Alternative methods for graphing of a continuous variable distribution.
11. Basic methods for graphing of a categorical variable (probability) distribution.
12. Alternative methods for graphing of a categorical variable distribution.
13. Basic methods for graphing of multivariate continuous data.
14. Alternative methods for graphing of multivariate continuous data.
15. Basic methods for graphing multivariate categorical data.
16. Alternative methods for graphing multivariate categorical data.
17. Basic methods for time series graphing.
18. Alternative methods for time series graphing.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Multivariate Statistical Analysis	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Data analysis software	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Methods for statistical analysis of complex data	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Visual analytics of time-oriented data	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Infographic and data analysis results presentation	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения

			(протокол № 12 от 23 мая 2023 года)

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

_____ _____ _____

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____

(ученая степень, ученое звание) (подпись) (И.О.Фамилия)