

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ



УТВЕРЖДАЮ

Проректор по учебной работе
и образовательным инновациям

О.Г. Прохоренко

5 июля 2023 г.

Регистрационный № УД- 1259/м.

Методы визуализации в анализе данных средствами языка R

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:**

7-06-0533-05 Прикладная математика и информатика

Профилизации: Компьютерный анализ данных
Интеллектуальные системы

2023 г.

Учебная программа составлена на основе ОСВО 7-06-0533-05-2023, примерного учебного плана, регистрационный № 7-06-05-016/пр. от 18.01.2023, учебных планов БГУ: № М53-5.3-43/уч. от 15.02.2023, № М53-5.3-79/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 (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

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 «Специальные программные средства».

Дисциплина «Методы визуализации в анализе данных средствами языка R» согласуется с дисциплиной государственного компонента «Программные средства анализа данных» и способствует успешному изучению дисциплин «Многомерный статистический анализ» модуля «Методы и компьютерные средства прикладной математики», «Методы статистического анализа сложных данных» модуля «Специальные методы анализа», «Визуальная аналитика данных с временной структурой» и «Инфографика и представление результатов анализа данных» модуля «Визуальная аналитика и инфографика», прохождению практики и написанию магистерских диссертаций.

Требования к компетенциям

Освоение учебной дисциплины «Методы визуализации в анализе данных средствами языка R» должно обеспечить формирование следующих компетенций:

***универсальные* компетенции:**

УК-2. Решать научно-исследовательские и инновационные задачи на основе применения информационно-коммуникационных технологий.

УК-5. Развивать инновационную восприимчивость и способность к инновационной деятельности.

***углубленные профессиональные* компетенции:**

УПК-6. Применять навыки использования основных методов визуализации больших данных средствами языка R.

В результате освоения учебной дисциплины обучающийся 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 «Методы визуализации в анализе данных средствами языка 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.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Введение

Тема 1.1. Коллекция tidyverse пакетов R для науки о данных

Язык R для статистических вычислений и графики. Интегрированная среда разработки RStudio. Коллекция tidyverse пакетов R для науки о данных.

Тема 1.2. Руководство по стилю программирования с tidyverse

Руководство по стилю программирования с tidyverse. Автоматическое форматирование кода при помощи пакета styler. Автоматическая проверка кода на соответствие руководству по стилю при помощи пакета lintr.

Раздел 2. Пакет ggplot2. Основы графика ggplot2

Тема 2.1. Пакет ggplot2

Пакет ggplot2. Обзор. Установка. Жизненный цикл. Экосистема расширений. Ресурсы для изучения ggplot2. Пакет plotly для создания интерактивных графиков.

Тема 2.2. Основы графика ggplot2

Функции ggplot(), aes(), `+`(<gg>), ` %+% `, ggsave() и их аргументы. Слои. Геометрические объекты. Статистические преобразования. Позиционные настройки. Аннотации. Визуальные атрибуты. Шкалы. Оси и легенды. Панели. Системы координат. Темы.

Тема 2.3. Визуализация и восстановление пропущенных значений

Пакеты panier, VIM для визуализации и восстановления пропущенных значений.

Тема 2.4. Цветовые палитры. Симуляторы цветовой слепоты

Коллекция цветовых палитр пакета paletteer. Симуляторы цветовой слепоты пакетов colorBlindness, colorblindr.

Раздел 3. Автоматизация отчетности

Тема 3.1. Автоматизированный графический исследовательский анализ данных

Пакеты dlookr, brinton для автоматизированного графического исследовательского анализа данных.

Тема 3.2. Система Quarto научно-технических публикаций

Система Quarto научно-технических публикаций с открытым исходным кодом.

Раздел 4. Графическое представление распределения переменной

Тема 4.1. Графическое представление распределения непрерывной переменной

Непрерывные переменные. Особенности. Основные способы графического представления. Альтернативные способы графического представления. Параметры графических представлений.

Тема 4.2. Графическое представление распределения категориальной переменной

Категориальные переменные. Номинальные, порядковые, дискретные переменные. Особенности. Основные способы графического представления. Альтернативные способы графического представления. Параметры графических представлений.

Раздел 5. Графическое представление многомерных данных

Тема 5.1. Графическое представление многомерных непрерывных данных

Многомерные непрерывные данные. Особенности. Основные способы графического представления. Альтернативные способы графического представления. Параметры графических представлений.

Тема 5.2. Графическое представление многомерных категориальных данных

Многомерные категориальные данные. Особенности. Основные способы графического представления. Альтернативные способы графического представления. Параметры графических представлений.

Раздел 6. Графическое представление временных рядов

Тема 6.1. Графическое представление временных рядов

Временные ряды. Особенности. Основные способы графического представления. Альтернативные способы графического представления. Параметры графических представлений.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

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

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
1	Введение	2		2	2			
1.1	Коллекция tidyverse пакетов R для науки о данных	1		1	1			опрос письменный отчет по аудиторным практическим упражнениям
1.2	Руководство по стилю программирования с tidyverse	1		1	1			опрос письменный отчет по аудиторным практическим упражнениям
2	Пакет ggplot2. Основы графика ggplot2	6		6	6			
2.1	Пакет ggplot2	2		2	2			опрос контрольная работа № 1 контрольная работа № 2

2.2	Основы графика ggplot2	2		2	2			опрос контрольная работа № 3 контрольная работа № 4
2.3	Визуализация и восстановление пропущенных значений	1		1	1			письменный отчет по аудиторным практическим упражнениям
2.4	Цветовые палитры. Симуляторы цветовой слепоты	1		1	1			письменный отчет по аудиторным практическим упражнениям
3	Автоматизация отчетности	2		2	2			
3.1	Автоматизированный графический исследовательский анализ данных	1		1	1			коллоквиум
3.2	Система Quarto научно-технических публикаций	1		1	1			письменный отчет по аудиторным практическим упражнениям
4	Графическое представление распределения переменной	4			4			
4.1	Графическое представление распределения непрерывной переменной	2			2			письменный отчет по домашним практическим упражнениям с их устной защитой
4.2	Графическое представление распределения категориальной переменной	2			2			письменный отчет по домашним практическим упражнениям с их устной защитой

5	Графическое представление многомерных данных	4			4			
5.1	Графическое представление многомерных непрерывных данных	2			2			письменный отчет по домашним практическим упражнениям с их устной защитой
5.2	Графическое представление многомерных категориальных данных	2			2			письменный отчет по домашним практическим упражнениям с их устной защитой
6	Графическое представление временных рядов	2			2			
6.1	Графическое представление временных рядов	2			2			письменный отчет по домашним практическим упражнениям с их устной защитой портфолио работ
ИТОГО		20		10	20			

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Брюс, П. Практическая статистика для специалистов Data Science: 50+ важнейших понятий с использованием R и Python / П. Брюс, Э. Броус, П. Гедек; [пер. с англ. А. Логунова]. - 2-е изд., перераб. и доп. - Санкт-Петербург: БХВ-Петербург, 2021. - 346 с.
2. О'Нил, К. Data Science. Инсайдерская информация для новичков. Включая язык R = Doing Data Science / Кэти О'Нил, Рэйчел Шатт; [пер. с англ. И. Пальти, К. Сеницы, С. Черникова]. - Санкт-Петербург [и др.]: Питер, 2019. - 364 с.
3. Мэтлофф, Н. Искусство программирования на R. Погружение в большие данные = The Art of R Programming. A Tour of Statistical Software Design / Норман Мэтлофф; [пер. с англ. Е. Матвеев]. - Санкт-Петербург [и др.]: Питер, 2019. - 413 с. - URL: <https://ibooks.ru/reading.php?short=1&productid=365272>.

Перечень дополнительной литературы

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 «Методы визуализации в анализе данных средствами языка 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 %.

Примерная тематика семинарских занятий

Семинар № 1. Коллекция tidyverse пакетов R для науки о данных. Руководство по стилю программирования с tidyverse.

Семинар № 2. Пакет ggplot2.

Семинар № 3. Основы графика ggplot2.

Семинар № 4. Визуализация и восстановление пропущенных значений. Цветовые палитры. Симуляторы цветовой слепоты.

Семинар № 5. Автоматизированный графический исследовательский анализ данных. Система Quarto научно-технических публикаций.

Примерная тематика лабораторных занятий

Занятие № 1. Коллекция tidyverse пакетов R для науки о данных. Руководство по стилю программирования с tidyverse.

Занятие № 2. Пакет ggplot2.

Занятие № 3. Основы графика ggplot2.

Занятие № 4. Визуализация и восстановление пропущенных значений. Цветовые палитры. Симуляторы цветовой слепоты.

Занятие № 5. Автоматизированный графический исследовательский анализ данных. Система Quarto научно-технических публикаций.

Занятие № 6. Графическое представление распределения непрерывной переменной.

Занятие № 7. Графическое представление распределения категориальной переменной.

Занятие № 8. Графическое представление многомерных непрерывных данных.

Занятие № 9. Графическое представление многомерных категориальных данных.

Занятие № 10. Графическое представление временных рядов.

Рекомендуемая тематика контрольных работ

Контрольная работа № 1. Геометрические объекты. Статистические преобразования.

Контрольная работа № 2. Позиционные настройки. Аннотации. Визуальные атрибуты.

Контрольная работа № 3. Шкалы. Оси и легенды.

Контрольная работа № 4. Панели. Системы координат. Темы.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

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

- освоение содержания 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.

Примерный перечень вопросов к экзамену

1. Коллекция tidyverse пакетов R для науки о данных
2. Руководство по стилю программирования с tidyverse
3. Пакет ggplot2
4. Основы графика ggplot2
5. Визуализация и восстановление пропущенных значений
6. Цветовые палитры. Симуляторы цветовой слепоты

7. Автоматизированный графический исследовательский анализ данных
8. Система Quarto научно-технических публикаций
9. Основные способы графического представления распределения непрерывной переменной
10. Альтернативные способы графического представления распределения непрерывной переменной
11. Основные способы графического представления распределения категориальной переменной
12. Альтернативные способы графического представления распределения категориальной переменной
13. Основные способы графического представления многомерных непрерывных данных
14. Альтернативные способы графического представления многомерных непрерывных данных
15. Основные способы графического представления многомерных категориальных данных
16. Альтернативные способы графического представления многомерных категориальных данных
17. Основные способы графического представления временных рядов
18. Альтернативные способы графического представления временных рядов.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

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)
Многомерный статистический анализ	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Программные средства анализа данных	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Методы статистического анализа сложных данных	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Визуальная аналитика данных с временной структурой	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения (протокол № 12 от 23 мая 2023 года)
Инфографика и представление результатов анализа данных	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)