РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ ЯЗЫКОВ R И PYTHON

1. Место дисциплины в структуре ОПОП

Дисциплина «Разведочный анализ данных с помощью языков R и Python» является дисциплиной профиля по выбору студента вариативной части для ОПОП «Общая биология и экология» мс_антропология. Изучается в 7 семестре студентами кафедры антропологии отделения «Общая биология и экология».

Дисциплина «Разведочный анализ данных с помощью языков R и Python» разработана для ознакомления обучающихся с основами синтаксиса, статистической обработки и визуализации данных с помощью языка R и Python, акцент сделан на основные средства разведочного анализа применительно к антропологическим данным. Первая часть курса посвящена основам синтаксиса языков, работе с файлами и базами данных, особенностям работы с антропологическими данными (работа с пропущенными значениями, малыми выборками, балловыми признаками и т.д), основам машинного обучения. Вторая часть курса посвящена визуализации полученных результатов. Рассказывается про алгоритм подбора графического метода и их дальнейшая реализация с помощью разных пакетов.

Освоение данной дисциплины необходимо как предшествующее для курса «Прикладная антропология», а также для выполнения выпускной квалификационной работы.

Цели освоения дисциплины

Ознакомление обучающихся с синтаксисом, основными пакетами и функциями языков программирования R и Python.

Задачи курса:

- получение базовых практических навыков работы с наиболее популярными языками программирования, применяющихся для анализа данных;
- умение использовать их при анализе реальных данных;
- умение выбирать различные методы визуализации в зависимости от типа поставленной задачи;
- умение максимально понятно представлять свои данные пользуясь настройками графиков.

2. Входные требования

Перед началом освоения дисциплины «Разведочный анализ данных с помощью языков R и Python» студент должен изучить курс «Статистический анализ антропологических данных».

3. Планируемые результаты изучения дисциплины, соотнесенные с требуемыми компетенциями выпускников

— ***Компетенции выпускников (коды)***:

СПК-6. Способность использовать современные компьютерные средства и специализированное программное обеспечение для статистической обработки антропометрических и других экспериментальных и обсервационных данных; владение навыками глубокого анализа и биологической интерпретации результатов статистической обработки антропологических данных; способность компьютерного моделирования антропологических и геногеографических интерполяционных карт.

— ***Планируемые результаты обучения по модулю, сопряженные с компетенциями***:

Освоение базовых навыков работы со стандартными библиотеками языков R и Python, формирование устойчивого навыка работы с прикладными пакетами статистической обработки данных, умения выбирать адекватные методы для извлечения информации из массивов данных, формирование навыка интерпретации результатов статистического анализа и умения использовать средства визуализации цифровой информации.

— ***Индикаторы (показатели) достижения компетенций***:

Знает:

- основы синтаксиса языков R и Python, основные библиотеки и их возможности.

Умеет:

- применять адекватные методы получения информации с помощью функций различных библиотек R и Python;
- визуализировать полученные данные с помощью графических возможностей.

Владеет навыками:

- анализа и визуализации данных на языках R и Python.

Демонстрирует готовность:

- применять полученные навыки работы в научной и профессиональной деятельности антрополога.

4. Объем дисциплины «Разведочный анализ данных с помощью языков R и Python»
у обучающихся на ОПОП «Общая биология и экология» по подплану мс_антропология:

- Общая трудоемкость дисциплины – 2 з.е. (72 ч).
- Аудиторная нагрузка – 56 ч. (4 ч. в неделю), из них семинары – 56 ч.
- Самостоятельная работа – 16 ч.
- Форма промежуточной аттестации – экзамен (7 семестр).

5. Форма обучения – очная

6. Содержание и структура дисциплины

№ п/п	Раздел дисциплины	Семинары (ч)	Самостоятельная работа (часы)
1	Тема 1. Введение в программирование, сравнение двух языков	4	0
2	Тема 2. Введение в статистическую обработку данных на языках программирования	4	
3	Тема 3. Введение в синтаксис языка R.	4	0
4	Тема 4. Введение в синтаксис языка Python. Работа с файлами. Работа с Data.frame	4	1
5	Тема 5. Особенности работы с разными типами антропологических данных	4	1
6	Тема 6. Функции для расчета описательной статистики	4	1
7	Тема 7. Многомерные методы анализа данных	4	1
8	Тема 8. Базовая графика на языке R	4	1
9	Тема 9. Графики с использованием пакета ggplot2	4	1
10	Тема 10. Визуализация многомерных данных с использованием пакета ggplot2	4	1
11	Тема 11. Продвинутая графика в контексте антропологических данных	4	1

12	Тема 12. Графические предоставления корреляционных матриц	4	2
13	Тема 13. Многомерные анализы и их визуализация с помощью языка Python	4	1
14	Тема 14. Методы машинного обучения на языке Python	4	2
15	Промежуточная аттестация – Экзамен		2
	Итого	56	16

6.1. Программа дисциплины «Разведочный анализ данных с помощью языков R и Python»

Тема 1. Введение в программирование. Сравнение функционала двух языков. Использование основных языковых конструкций: условия, циклы, функции и другие.

Тема 2. Введение в статистическую обработку данных на языках программирования.

Тема 3. Введение в синтаксис языка R. Базовые объекты языка R (вектора, списки, матрицы и многомерные матрицы). Простейшие математические операции. Работа с Data.frame. Загрузка и сохранение данных.

Тема 4. Введение в синтаксис языка Python. Основные языковые конструкции, отличия синтаксиса от R, работа с файлами, работа с Data.frame. загрузка и сохранение данных.

Тема 5. Особенности работы с разными типами антропологических данных: балловые признаки, частоты, признаки с ненормальным распределением и т.д.). Работа с малыми выборками и пропущенными значениями. Особенности работы на Python.

Тема 6. Функции для расчета описательной статистики. Достоверности различий.

критерии проверки на нормальность, для выборок разной численности с использованием R и Python.

Тема 7. Многомерные методы в языке R: анализ главных компонент, расчет матриц расстояний, кластерный анализ.

Тема 8. Базовая графика на языке R: основные пакеты и аргументы.

Тема 9. Графики с использованием пакета ggplot2: ящики с усами, гистограммы, столбцовые диаграммы, линейные диаграммы, диаграммы рассеяния, географические данные.

Тема 10. Визуализация многомерных данных с использованием пакета ggplot2: диаграммы рассеяния, пиктограммы.

Тема 11. Продвинутая графика в контексте антропологических данных: Оценка таксономической значимости признаков. Классификация групп, по разным наборам признаков. Определение положения неизвестной выборки в более крупной общности, на основе разных методов машинного обучения.

Тема 12. Графические предоставления корреляционных матриц: температурная карта, эллипсы, графы.

Тема 13. Многомерные анализы и их визуализация с помощью языка Python: главные компоненты; кластерный анализ многомерное шкалирование.

Тема 14. Методы машинного обучения на языке Python: дерево классификаций, Random Forest, дискриминантный анализ

7. Фонд оценочных средств для оценивания результатов обучения по дисциплине:

7.1. Типовые задания и иные материалы, необходимые для оценки результатов обучения

Примерный список заданий для проведения текущей аттестации (для подготовки к коллоквиумам, контрольным, опросам)

По данной дисциплине в качестве текущего контроля успеваемости предусмотрено выполнение практических заданий по изученной теме и их обсуждение в рамках семинарского занятия.

Примерный список вопросов для промежуточной аттестации (экзамен)

1. Чем характеризуются языки программирования R и Python, общее и различия
2. Как загружать данные в R?
3. Как осуществлять работу с пакетами и директориями?
4. Типы данных в R и работа с ними.
5. Типы данных в Python и работа с ними
6. Работа с пропущенными значениями в разных пакетах языков программирования.
7. Расчет параметров описательной статистики, статистические критерии и их реализация в R.
8. Расчет параметров описательной статистики, статистические критерии и их реализация в Python.
9. Внутригрупповой анализ в R (на примере анализа главных компонент).
10. Линейный регрессионный анализ в R и Python.

11. Корреляционный анализ, виды коэффициентов корреляции их расчет в R.
12. Факторный анализ на языке R и Python.
13. Расчет матрицы различий и матрицы расстояний и реализация кластерного анализа в R и Python.
14. Канонический дискриминантный анализ в R.
15. Основные возможности и функции пакета ggplot2, основные настраиваемые параметры.
16. Построение boxplot и гистограмм, интерпретация данных.
17. Построение столбчатых и линейных диаграмм, интерпретация данных.
18. Построение диаграмм рассеяния на языке R и Python.
19. Обработка данных связных с географией и картами.
20. Визуализация результатов корреляционного анализа температурная карта, графы.
21. Классификационный анализ, построение деревьев классификации.
22. Построение деревьев решений и проведение анализа Random Forest с помощью языка Python

7.2. Описание критериев и шкал оценивания

Описание критериев оценивания выполнения задания

Показатель	Баллы
Студент выполняет менее 50% задания	0-20
Задание студент выполняет все или большей частью, есть отдельные неточности, способен при направляющих вопросах исправить допущенные неточности	21-32
Задание выполнено студентом правильно, самостоятельно в полном объеме	33-40

Шкала оценивания сформированности компетенций

Уровень сформированности компетенции	Баллы	Оценка в 5-балльной шкале
Недостаточный	Менее 20	неудовлетворительно
Базовый	20-26	удовлетворительно
Высокий (повышенный)	27-32	хорошо
Продвинутый (повышенный)	33-40	отлично

8. Ресурсное обеспечение:

8.1. Перечень основной и дополнительной учебной литературы

Основная литература

1. R in Action. Data analysis and graphics with R. ROBERT I. KABACOFF
2. Python для сложных задач: наука о данных и машинное обучение. Вандер Плас Д.O'Reilly. 2023.
3. https://r-graphics.org/
4. https://www.r-project.org/
5. https://r-graph-gallery.com/

Дополнительная литература

1. https://tsamsonov.github.io *(*Самсонов Т.Е. Пространственная статистика и моделирование на языке R. М.: Географический факультет МГУ, 2021.)
2. https://pozdniakov.github.io (Поздняков И. Анализ данных и статистика в R., 2021)

8.2. Перечень лицензионного программного обеспечения

1. R base,
2. R studio,
3. Python,
4. Jupyter Notebook.

8.3. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

Электронная библиотека механико-математического факультета МГУ http://lib.mexmat.ru/books

9. Язык преподавания

Русский

10. Преподаватель

Кузнецова Ольга Алексеевна – кандидат биологических наук, научный сотрудник кафедры антропологии биологического факультета МГУ

11. Автор программы

Кузнецова Ольга Алексеевна – кандидат биологических наук, научный сотрудник кафедры антропологии биологического факультета МГУ