Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ:
Директор
_____А. В. Замятин
«_16_» ____мая_____ 20_22_ г.

Рабочая программа дисциплины

Представление знаний и визуализация данных

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки :
Интеллектуальный анализ больших данных

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: Б1.В.ДВ.01.02.02

СОГЛАСОВАНО:
Руководитель ОП
_____А.В. Замятин

Председатель УМК
_____С.П. Сущенко

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

– ПК-1 – способность разрабатывать и применять математические методы, алгоритмы, программное обеспечение для решения задач научно-исследовательской и проектной деятельности.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-1.3 Разрабатывает новые методы, модели, алгоритмы и программное обеспечение для решения задач в области профессиональной деятельности.

2. Задачи освоения дисциплины

– Формирование у студентов теоретических знаний и практических навыков в области методов, средств, подходов и принципов визуального представления результатов научно-исследовательской деятельности, основанных на основных положениях интеллектуального анализа данных, машинного обучения и реализующихся в выборе инструментов и технологий, к которым можно отнести современные скриптовые языки Python и R..

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Дисциплина входит в модуль «Биоинформатика и биомедициная».

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Второй семестр, зачет

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Введение в интеллектуальный анализ данных», «Математические методы и модели для компьютерных наук». .

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
-лекции: 16 ч.
-лабораторные: 16 ч.
в том числе практическая подготовка: 0 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Раздел 1. Визуализация многомерных данных. Задачи визуализации. Способы визуализации.
Цель визуализации, задачи визуализации многомерных данных, классификация по визуализируемым объектам. Выбор правильного типа визуализации. Первичный анализ данных с использованием методов визуализации. Библиотеки Python (или R) для решений задач визуализации. Сравнение полученных визуальных образов.
Раздел 2. Python-библиотеки для визуализации данных в Data Science .

Возможности библиотек Matplotlib, Seaborn, Missingno, Altair, Plotly, Bokeh, Pygal, Networkx. Примеры использования перечисленных библиотек. Визуализация многомерных данных с использованием диаграмм Эндрюса.

Раздел 3. Визуализация данных средствами дашбордов.

Что из себя представляет дашборд, его свойства, отличие от отчета. Модули дашборда. Виды дашбордов. Основные инструменты: Google Sheets, Яндекс.Метрика, Google Analytics, Google Data Studio, Qlik, Power Bi, Owox Bi.

Раздел 4. Методы визуализации для решения прикладных задач.

Методы визуализации для задач классификации, кластеризации. Визуализация решения задачи временных рядов, демонстрация примеров.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля лабораторных работ.

Лабораторная работа № 1

Исходные данные: изучите открытые данные по выборам депутатов Государственной Думы Федерального Собрания Российской Федерации седьмого созыва: http://www.vybory.izbirkom.ru/region/region/izbirkom?action=show&root=1&tvd=100100067795854&vrn=100100067795849&region=0&global=1&sub_region=0&prver=0&pronetvd=0&vibid=100100067795854&type=233. Выберите для дальнейшего анализа данные по одному из округов.

Задание: 1) продумайте и подберите тип визуализации для ответа на следующие вопросы:

А) какие партии являются лидерами (аутсайдерами) по количеству голосов;

Б) какие избирательные участки являются лидерами (аутсайдерами) по количеству пришедших на выборы;

В) есть какие-то закономерности (связи) между количеством голосов за лидирующие партии и конкретными избирательными участками;

Г) можно ли определить какие-то другие связи?

Д) сделайте выводы.

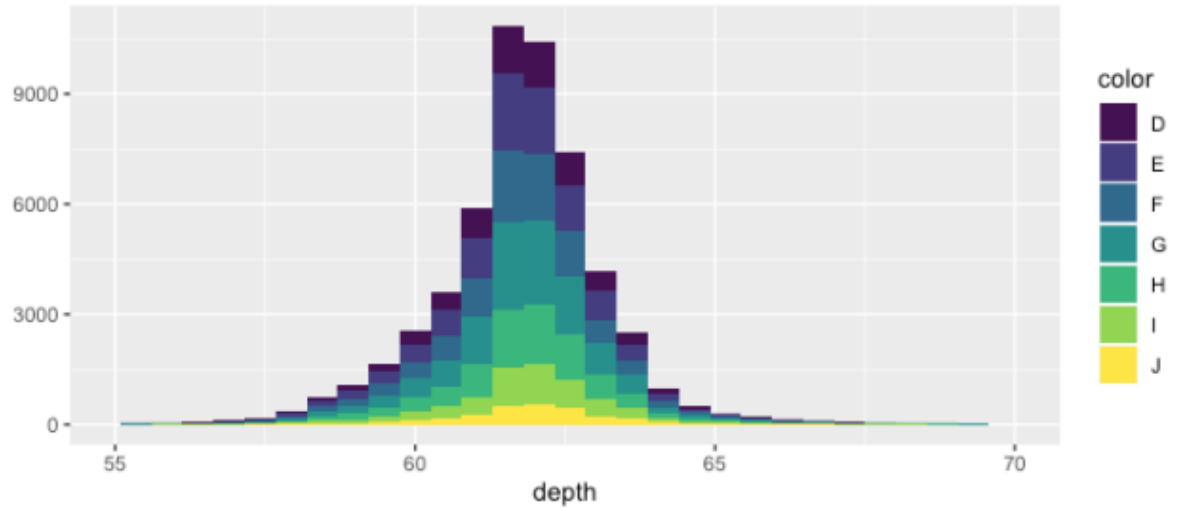
Лабораторная работа № 2

Цель работы: познакомиться с основными элементами графической грамматики основных библиотек визуализации.

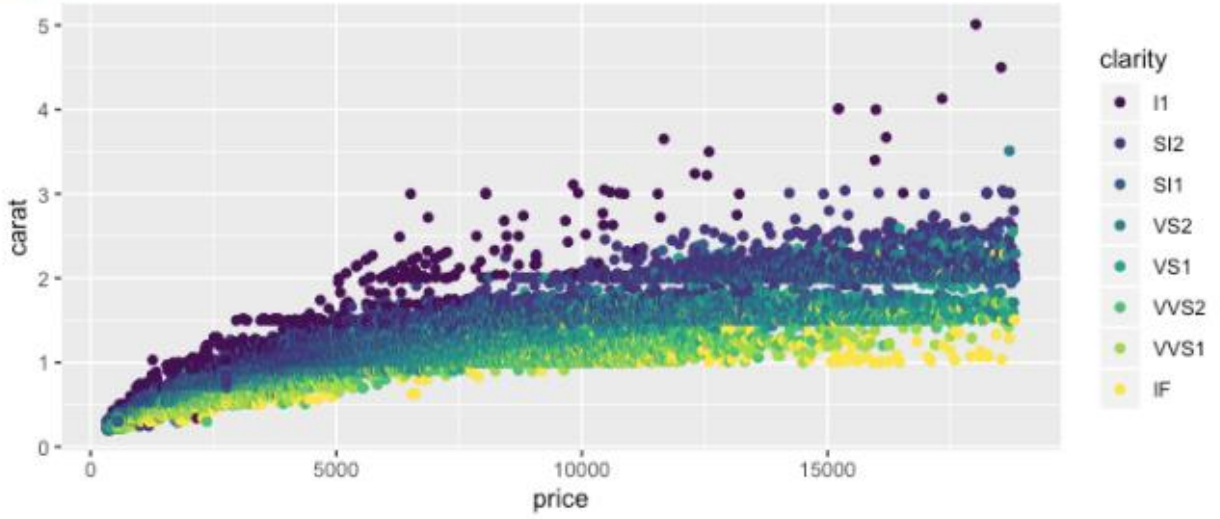
Задания:

Задание 1. С помощью любой из изученных библиотек напишите программный код, строящий следующие графики по набору данных diamonds.

Вариант – 1

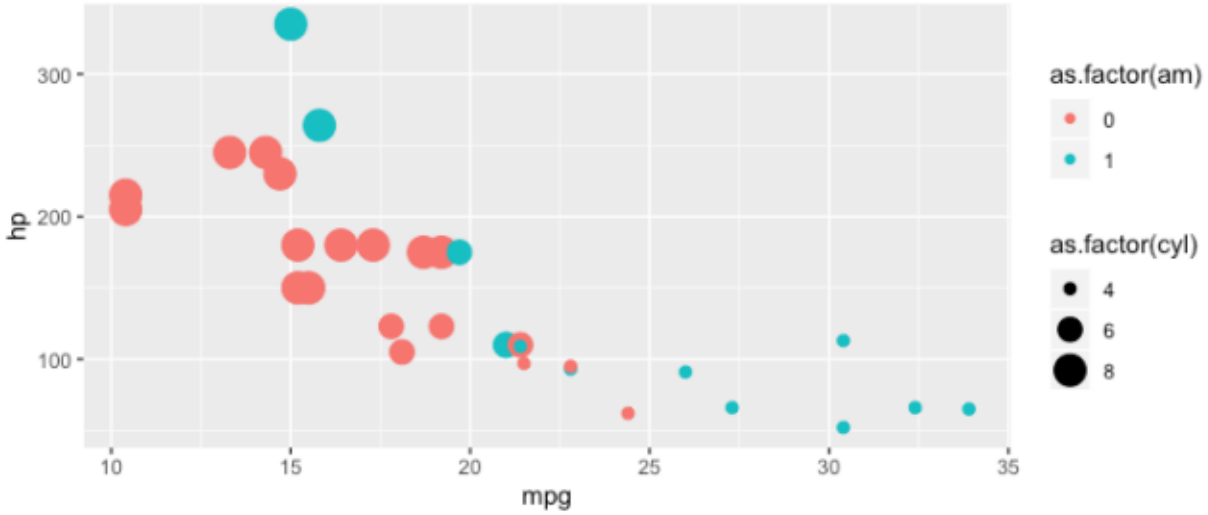


Вариант – 4

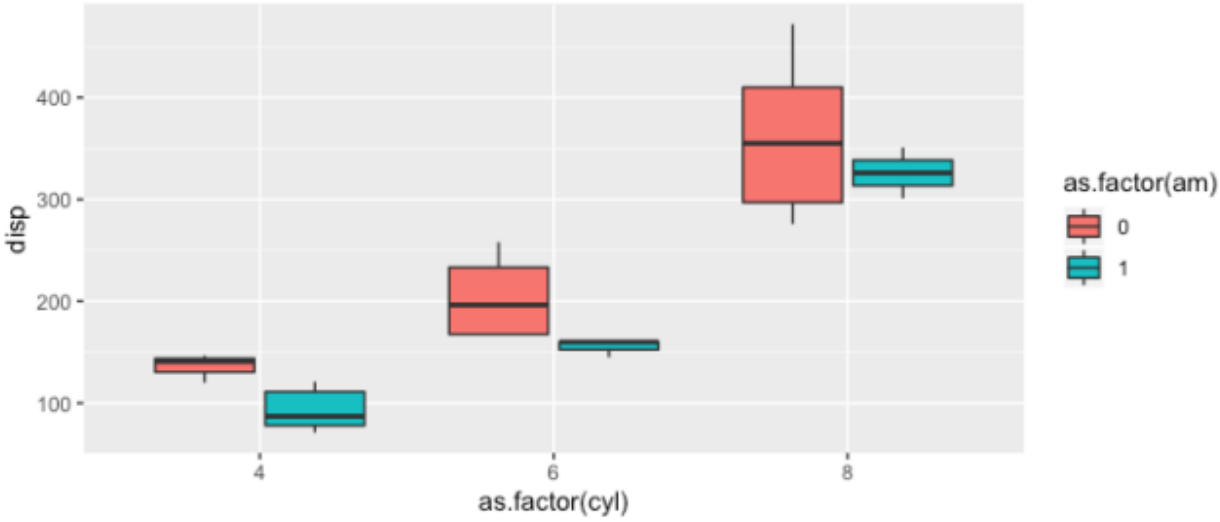


Задание 2. Напишите программный код, строящий следующие графики по набору данных mtcars.

Вариант – 1



Вариант – 4



Задание 3. Постройте два произвольных графика, показав умение использовать различные настройки для набора данных по заданному варианту. Дайте описание набору данных и то, что показывает изображенный график.

№	Набор данных	№	Набор данных	№	Набор данных	№	Набор данных
1	CO2	8	Puromycin	15	iris	22	swiss
2	ChickWeight	9	Seatbelts	16	longley	23	trees
3	DNase	10	Theoph	17	mtcars	24	economics
4	LifeCycleSavings	11	ToothGrowth	18	quakes	25	faithfuld
5	Loblolly	12	USArrests	19	rock	26	midwest
6	Orange	13	freeny	20	stack.x	27	mpg
7	OrchardSprays	14	infert	21	stackloss	28	txhousing

Лабораторная работа № 3

1. Выберите датасет, согласно своего варианта (https://habr.com/ru/post/452392/). Можно выбрать наиболее интересный для себя.
2. Выполните первичный анализ данных, **используя только методы визуализации:**
   a. Выбросы
   b. Пропущенные значения (с анализом)
   c. Коллинерные признаки.
3. Выполните визуализацию зависимости целевой переменной/переменных от информативных признаков. Проанализируйте.
4. Выполните различные виды визуализации, согласно примерам из лекций. Можно добавить свои.
5. Сделайте общие выводы.

ВАРИАНТЫ:

1. Данные смертей и сражений из игры престолов — этот набор данных объединяет три источника данных, каждый из которых основан на информации из серии книг.
2. Глобальная база данных терроризма — Более 180 000 террористических атак по всему миру, 1970-2017.
3. Биткойн, исторические данные — данные биткойнов с интервалом в 1 минуту с избранных бирж, январь 2012 г. — март 2019 г.
4. FIFA 19 полный набор данных игроков — 18k + FIFA 19 игроков, ~ 90 атрибутов, извлеченных из последней базы данных FIFA.
5. Статистика видео YouTube — ежедневная статистика трендовых видео на YouTube.
6. Обзор показателей самоубийств с 1985 по 2016 год — Сравнение социально-экономической информации с показателями самоубийств по годам и странам.
7. Huge Stock Market Dataset — исторические дневные цены и объемы всех американских акций и ETF.
8. Индикаторы мирового развития — показатели развития стран со всего мира.
9. Kaggle Machine Learning & Data Science Survey 2017 — Большое представление о состоянии науки о данных и машинного обучения.
10. Данные о насилии и оружии — полный отчет о более чем 260 тыс. американских инцидентов с применением оружия в 2013-2018 гг.

Лабораторная работа № 4

1. Задание рассчитано на выполнение в группе – 2 человека (по желанию – можно индивидуально).
2. Изучить один из инструментов создания дашбордов, используя документацию и тестовые примеры:

- Вариант 1. Google Sheets (Excel)
- Вариант 2. Яндекс.Метрика
- Вариант 3. Google Analytics
- Вариант 4. Google Data Studio
- Вариант 5. Qlik
- Вариант 6. Tableau
- Вариант 7. Power Bi
- Вариант 8. Owox Bi

3. Построить свой дашборд, используя открытые данные любого выбранного вами сайта.
4. В отчет должно входить описание используемой системы, аргументация выбора средств визуализации, скриншоты дашборда и ссылка на дашборд.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет выставляется на основе представления и защиты индивидуального проекта. Студент выполняет презентацию, а также демонстрирует программный код. Вопросы по результатам могут задавать все студенты группы, не только преподаватель.

Проект может быть выполнен как индивидуально, так и в мини-группе (2-3 чел.), при условии, что объем работы также будет увеличен. В конце семестра по каждому проекту представляется мини-презентация о результатах работы.

Тематика индивидуального проекта связана с темой ВКР магистранта. Цель работы – использование методов визуализации в своей научной работе.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle»

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Мастицкий, С.Э., В.К. Шитиков. Статистический анализ и визуализация данных с помощью R. – Москва : ДМК Пресс, 2015. — 496 с.
– Мастицкий, С.Э. Визуализация данных с помощью ggplot2 . – Москва : ДМК Пресс, 2017. — 222 с.
– Роберт, И. R в действии. Анализ и визуализация данных в программе R : руководство. – Москва : ДМК Пресс, 2014. — 588 с.
– Сузи, Р. А. Язык программирования Python. – Москва : Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. — 350 с.
– Маккинни, У. Python и анализ данных. – Москва : ДМК Пресс, 2020. — 540 с.

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Программное обеспечение – средства программирования на Python3 и R:

- Python3

- RStudio.

б) информационные справочные системы:
– Лань, электронно-библиотечная система
– IPRBooks, электронно-библиотечная система
– TOMSK STATE UNIV Electronic Resources

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

Марухина Ольга Владимировна, канд. техн. наук, доцент кафедры теоретических основ информатики ТГУ.