

Прикладная аналитика на R

Best Practice по работе с данными с помощью языка R

Длительность курса: 192 академических часа

1 Программирование на языке R

1 Знакомство с языком R. Установка рабочего окружения.

Рассматривается история создания языка R, сфера его использования, преимущества и недостатки.

Участники научатся настраивать рабочее окружения под основными ОС (Windows/Linux/macOS), устанавливать пакеты и пользоваться встроенной справкой, а также познакомятся с возможностями IDE RStudio.

Домашние задания

1 ДЗ №1.1. Эссе о целях и задачах

Цель: Слушатели получат понимание экосистемы языка, разберутся в базовом синтаксисе, научатся устанавливать пакеты и пользоваться встроенной справкой, а также установят на свои ОС (Mac, Windows, Linux) IDE RStudio и разберутся с ее функционалом

Напишите эссе о том, с какими ограничениями текущего используемого вами инструментария вы планируете справиться, изучив R; какие задачи вы хотели бы научиться решать в рамках данного курса.

2 Базовый синтаксис языка. Типы и структуры данных.

Слушатели освоят интерактивное использование R, изучат основные типы и структуры данных, а также получат вводную информацию об организации кода в свои проектах.

Домашние задания

1 ДЗ №1.2. Практика по синтаксису и структурам данных

Цель: Целью данного урока является подготовка и объяснение слушателям необходимой информации для того чтобы разобраться в типах и структурах данных, основных библиотеках языка, пользоваться поиском для установки проблемных библиотек и находить пути решения проблем. Студенты увидят как парсить текстовый документ и в дальнейшем научиться верстать документы в красивые презентации в формате HTML и PDF

Выполните задания из файла `hw_1_2.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

3 **Управляющие конструкции. Векторизация.**

Слушатели изучат управляющие конструкции языка R и научатся использовать векторизованные вычисления.

Домашние задания

1 ДЗ №1.3. Практика по управляющим конструкциям

Цель: Целью урока является показать и рассказать почему привычные управляющие конструкции языка такие как циклы и условные блоки не подходят для анализа данных, т.к слишком медлительны. Поэтому целью урока будет написание и создание векторов для эффективного решения задач и использование функций.

Выполните задания из файла `hw_1_3.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

4 Написание и использование функций.

Слушатели научатся использовать функции из сторонних пакетов и писать свои собственные функции для автоматизации повторяющихся операций, а также получат базовые сведения об обработке исключений.

Домашние задания

1 ДЗ №1.4. Практика по функциям

Цель: Слушатель научиться писать массивы кода в функции для более лаконичного и читабельного представления данных, также студенты разберутся со встроенными функциями без использования сторонних пакетов

Выполните задания из файла `hw_1_4.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

5 **Объектно-ориентированное программирование в R.**

Участники изучат основные объектные системы языка R и особенности диспетчеризации для S3/S4-классов (генерические функции вместо методов, определяемых внутри класса).

Домашние задания

1 ДЗ №1.5. Практика по ООП

Цель: В данной домашней работе студент разрабатывает класс S3 (в качестве бонуса и доп. практики - любой другой), пишет методы для реализованного класса. Целью задания является получение опыта ООП в R, а также понимания, как работают классы в R. Последнее очень пригодится при работе с объектами S3/S4, написанными другими программистами R.

Выполните задания из файла `hw_1_5.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

6 Окружения и области видимости. Основы функционального программирования.

Слушатели ознакомятся с иерархией окружений и узнают, как в R происходит поиск объектов по именам; освоят такие элементы функционального программирования, как создание функций-замыканий и использование функций высшего порядка, применяющих заданную функцию к элементам списков.

Домашние задания

1 ДЗ №1.6. Практика по функциональному программированию

Цель: В данной работе студенты учатся работать с областями видимости функций, применяют подход функционального программирования для решения задач. Полученные умения и знания пригодятся для улучшения структуры, читабельности и эффективности кода, призванного выполнять повторяющиеся задачи.

Выполните задания из файла `hw_1_6.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

7 Контроль версий и воспроизводимые исследования (literate programming).

Участники изучат концепцию literate programming и ее реализацию в виде пакета knitr, освоят автоматическую генерацию отчетов в html/docx, научатся применять систему контроля версий git.

Домашние задания

1 ДЗ №1.7. Практика по github

Оформите результаты предыдущих домашних работ в виде репозитория на github. Напишите информативный файл README.

8 Что делать, если ничего не работает. Обзор экосистемы R.

Слушатели изучат эффективные методики поиска ответов на возникающие при изучении R вопросы, научатся задавать корректные вопросы на Stack Overflow, узнают о сообществах R-пользователей и ознакомятся с экосистемой языка R.

Домашние задания

1 ДЗ №1.8. Практика по нахождению и освоению специализированных пакетов

В ходе этого творческого задания вам предстоит самостоятельно найти и освоить пакет для решения какой-то нетривиальной для вас задачи (в контексте работы с R). Хорошим вариантом может быть работа с изображениями, например, последовательность действий от загрузки .jpeg до сохранения уменьшенного и обрезанного изображения с наложенным на него текстом. Результатом работы должен быть .Rmd-файл, иллюстрирующий решение поставленной задачи.

2 Загрузка и выгрузка данных.

1 Источники данных.

участники изучат основные источники данных, научатся загружать данные и выбирать оптимальный способ и формат данных для различных ситуаций.

Домашние задания

1 ДЗ №2.1. Практика по источникам данных

Выполните задания из файла `hw_2_1.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`).

Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown-скрипта` - `.Rmd`.

2 Работа с базами данных.

участники научатся загружать данные в R из популярных СУБД при помощи пакета dbi и сохранять данные в БД при помощи пакета RSQLite.

Домашние задания

1 ДЗ №2.2. Практика по базам данных

Выполните задания из файла hw_2_2.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки

(variable_name вместо variable.name).

Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

3 data.table - основы.

Участники изучат основы синтаксиса data.table, научиться считывать и сохранять табличные данные, выполнять отбор наблюдений и столбцов, а также использовать группировки.

Домашние задания

1 ДЗ №3.1. Практика по основам data.table

Выполните задания из файла hw_3_1.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name).

Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

4 **Текстовые и неструктурированные данные.**

Участники изучат базовые приемы работы с текстовыми данными, начнут изучать регулярные выражения, познакомятся с форматом JSON и научатся парсить XML и HTML при помощи пакетов `rvest` и `xml2`.

Домашние задания

1 ДЗ №2.4. Практика по неструктурированным и текстовым данным

Выполните задания из файла `hw_2_4.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`).

Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown-скрипта` - `.Rmd`.

3 Преобразования данных. Пакет data.table. Tidyverse.

1 Работа с API для получения данных.

студенты научатся загружать данные с веб-сайтов с использованием API, таких как FRED, Quandl, quantmod (финансовые и экономические данные) и Вконтакте (соц сети)

Домашние задания

1 ДЗ №2.3. Практика по работе с API

Выполните задания из файла hw_2_3.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name).

Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

2 **data.table** - переформатирование и объединение таблиц.

Слушатели изучат концепцию "опрятных данных" (tidy data), научатся выполнять преобразования в "широкий" формат из "длинного" и наоборот.

Домашние задания

- 1 ДЗ №3.3. Практика по переформатированию и объединению таблиц

Выполните задания из файла `hw_3_3.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`).

Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

3 Программирование с использованием `data.table`.

участники научатся использовать возможности `data.table` внутри написанных ими функций.

Домашние задания

- 1 ДЗ №3.4. Практика по программированию с использованием `data.table`

Выполните задания из файла `hw_3_4.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`).

Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown-скрипта` - `.Rmd`.

4 Концепция `tidyverse`. Манипуляции данными с помощью `dplyr`.

знакомство с `tidyverse`. 6 типов `join`-ов в `dplyr`.
Пайплайны.

Домашние задания

- 1 Домашняя работа
-

5 **data.table** - ключи и модификация по ссылке.

участники научатся повышать скорость работы с данными в пакете `data.table` при помощи создания индексов, а также изучат оператор `:=`, позволяющий модифицировать таблицы без создания их копий.

Домашние задания

1 ДЗ №3.2. Практика по ключам и модификации по ссылке

Выполните задания из файла `hw_3_2.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`).

Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown`-скрипта - `.Rmd`.

1 Грамматика графики. Основы ggplot2.

Участники узнают, что такое грамматика графики, как устроено послойное создание графиков в ggplot2, а также научиться создавать наиболее часто используемые графики при помощи данной библиотеки.

Домашние задания

1 ДЗ №4.1. Практика по основам ggplot2

Выполните задания из файла hw_4_1.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода <https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name). Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

2 Сложные графики с использованием ggplot2.

Слушатели научатся создавать сложные типы графиков, включая "фасеточные", и освоят использование дополнительных цветовых палитр и тем.

Домашние задания

1 ДЗ №4.2. Практика по сложным графикам

Выполните задания из файла hw_4_2.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода <https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown-скрипта` - `.Rmd`.

3 Пакеты, расширяющие возможности ggplot2.

Участники изучат создание специализированных графиков, используя возможности пакетов GGally, owplot и patchwork. Также будет рассмотрено создание анимированных графиков при помощи ganimate.

Домашние задания

- 1 ДЗ №4.3. Практика по пакетам, расширяющим возможности ggplot2

Выполните задания из файла hw_4_3.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в

именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name). Домашнее задание

должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

- 4 **Интерактивная визуализация с plotly.** Участники познакомятся с основами интерактивной визуализации при помощи plotly и узнают, когда уместно использовать интерактивные графики.

Домашние задания

- 1 ДЗ №4.4. Практика по интерактивной визуализации

Выполните задания из файла hw_4_4.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода <https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name). Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

-
- 5 **Фреймворк Shiny. Интерактивные дашборды.** знакомство с Shiny и пакетами, расширяющими его функционал. Построение интерактивного дашборда, визуализация сырых данных, позволяющая рассказать историю с помощью данных

5 Введение в машинное обучение на языке R

Обзор методов глубокого обучения.

1 О чем говорят ваши данные и что с ними делать.

участники узнают что такое статистические выводы, машинное обучение и какие инструменты можно использовать.

Домашние задания

1 ДЗ №5.1. Практика по постановке и формализации бизнес-задач

Выполните задания из файла `hw_5_1.Rmd`, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (`variable_name` вместо `variable.name`). Домашнее задание должно быть выполнено в `r markdown` и без ошибок рендериться в `html`. Расширение файла `markdown-скрипта` - `.Rmd`.

2 **Алгоритмы машинного обучения (наивный Байес, деревья решений, k-means, логистическая регрессия). Рекомендательные системы и специальные задачи ML.**

формулировать задачи анализа данных на языке R, относящиеся к разным классам машинного обучения.

Домашние задания

1 ДЗ №5.2. Практика по статистике и теории вероятностей

Выполните задания из файла hw_5_2.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода <https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name). Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

3 **Применение машинного обучения на реальных задачах**

участники познакомятся с реальными кейсами ML применяемыми в различных индустриях (Oil&Gas, retail, healthcare, finance, risk mgm, trading)

Домашние задания

1 ДЗ №5.3. Практика по статистическим критериям

Выполните задания из файла hw_5_3.Rmd, который находится в разделе Материалы этого занятия.

При работе используйте гайд по стилю оформления кода

<https://google.github.io/styleguide/Rguide.xml>, в именах переменных используйте нижнее подчеркивание вместо точки (variable_name вместо variable.name). Домашнее задание должно быть выполнено в r markdown и без ошибок рендериться в html. Расширение файла markdown-скрипта - .Rmd.

1 «Постановка и формализация бизнес-задачи».

участники узнают о важности выбора корректных метрик качества в аналитических задач, а также о важности предварительной оценки целесообразности применения методов аналитики или машинного обучения.

Домашние задания**1 Проектная работа**

Порядок выполнения работы:

1. Выбрать тему и обосновать ее актуальность, сформулировать цели и задачи исследования. Сформулировать гипотезы, подлежащие проверке на этапе статистического анализа.

2. Выбрать источник данных, задокументировать версию набора данных и/или дату выгрузки, осуществить выгрузку данных для локального использования. Поиск данных можно начать с <https://www.google.com/publicdata/directory>. Представленные по этому адресу наборы данных доступны для скачивания в CSV на сайтах <https://data.worldbank.org/> и <https://ec.europa.eu/eurostat/en/web/government-finance-statistics/statistics-illustrated>. Задача со звездочкой: загрузить данные с использованием API при помощи <https://cran.r-project.org/web/packages/RSocrata/index.html>. Данные по России можно найти, например, на сайте ВШЭ http://sophist.hse.ru/data_access.shtml.

3. Сохранить выгруженные данные для дальнейшего локального использования в двух вариантах: в CSV и в БД MonetDBLite или другую по собственному выбору таким образом, чтобы их было удобно повторно загружать в R. Убедиться,

что при сохранении и повторной загрузке целостность данных не нарушается. Задача со звездочкой: освоить сохранение в бинарный формат .fst и загрузку из него.

4. Описать таблицу или таблицы данных (количество строк и столбцов, описание и типы переменных). Вывести основные описательные статистики: для количественных переменных - среднее, стандартное отклонение, медиана, минимум и максимум; для категориальных - частота и доля в процентах.

5. Выполнить предварительную обработку данных (при необходимости объединить таблицы, выполнить агрегирование и другие преобразования). Сохранить обработанные данные в отдельный CSV и/или путем добавления новой таблицы к созданной в пункте 3 БД.

6. Выполнить визуализацию данных с двумя целями: понять структуру данных для дальнейшего анализа и рассказать при помощи графиков определенную историю, адресатом которой может быть как (воображаемый) руководитель проекта, так и просто заинтересованный читатель.

7. Статистический анализ. Выполнить проверку сформулированных ранее гипотез с использованием подходящих статистических методов.

8. Выводы. Если на некоторые из поставленных вопросов ответить в ходе исследования не удалось, следует постараться объяснить, почему.

- 2 **«Предиктивная аналитика и статистические выводы.»** участники повторят основные понятия теории вероятностей, такие как уровень значимости и мощность. Будет рассмотрено точечное и интервальное оценивание параметров в контексте типичных аналитических задач.
-
- 3 **«Проверка гипотез.»** участники познакомятся с основами A/B-тестирования, со статистическими критериями для сравнения долей и средних, с общей линейной моделью и с бутстрепом.
-
- 4 **« Рандомизация. Расчет размера выборки. »** участники изучат виды рандомизации и причины, делающие использование рандомизации крайне желательным способом формирования выборок; научатся рассчитывать размеры выборок для типичных дизайнов экспериментов.

- 1 **«Вводное занятие по проектной работе.»** это занятие поможет определиться с темой проекта, его объемом и используемыми данными, а также с форматом представления результатов работы.

Домашние задания

1 Проект

Цель: Утверждение темы проекта, его задач и результатов

- 2 **«Консультация по проектной работе.»** слушатели курса получают комментарии относительно прогресса проектной работы, ответы на вопросы, рекомендации по реализации.
-

- 3 **«Оценка проектных работ»** разбор проектов, комментарии и выставление оценок.