

Data Science Day 1

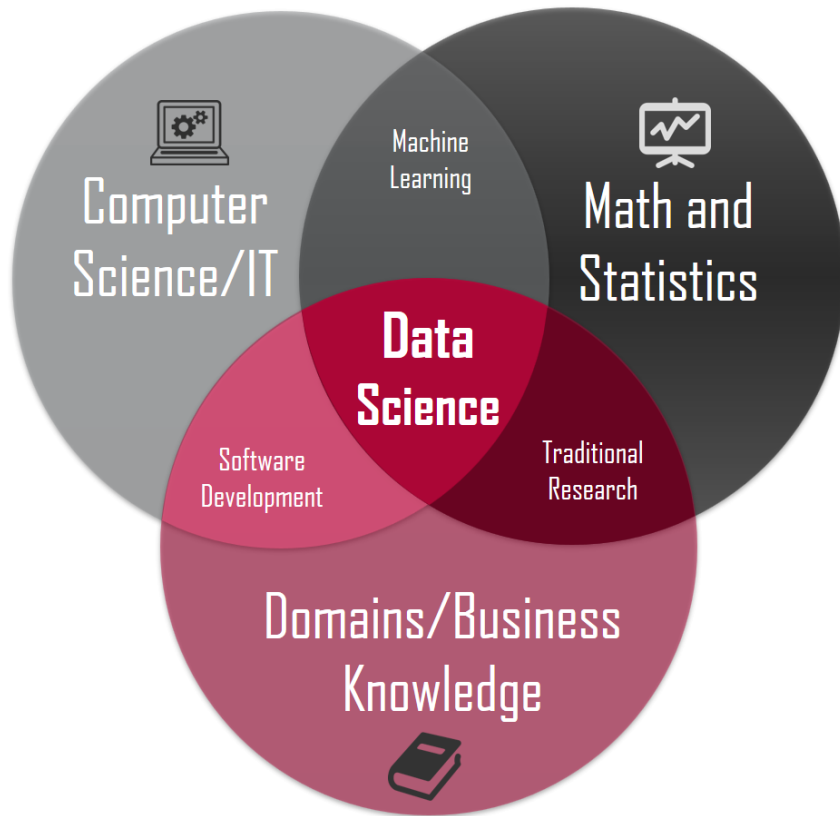
Правила 😊

- Занятия
 - 18-45 по 21-15 (Пн-Ср)
- В здании – пропускная система
- Кофе, чай, молоко, сливки – в кафетериях на 1,2,8 этажах
- Вода в кулерах – на каждом этаже
- Курение – только на улице
- Материалы курса -
https://github.com/Demolis/Infopulse/tree/master/DS_Mon_Wed_2019

Структура курса

- Знакомство
- Что такое Data Science, основные определения
- Основы языка R
- Data Mining и математическая статистика
- Машинное обучение
- Текстовый анализ
- Пространственный анализ
- GUI
- Big Data, что такое и как работать

Data Science

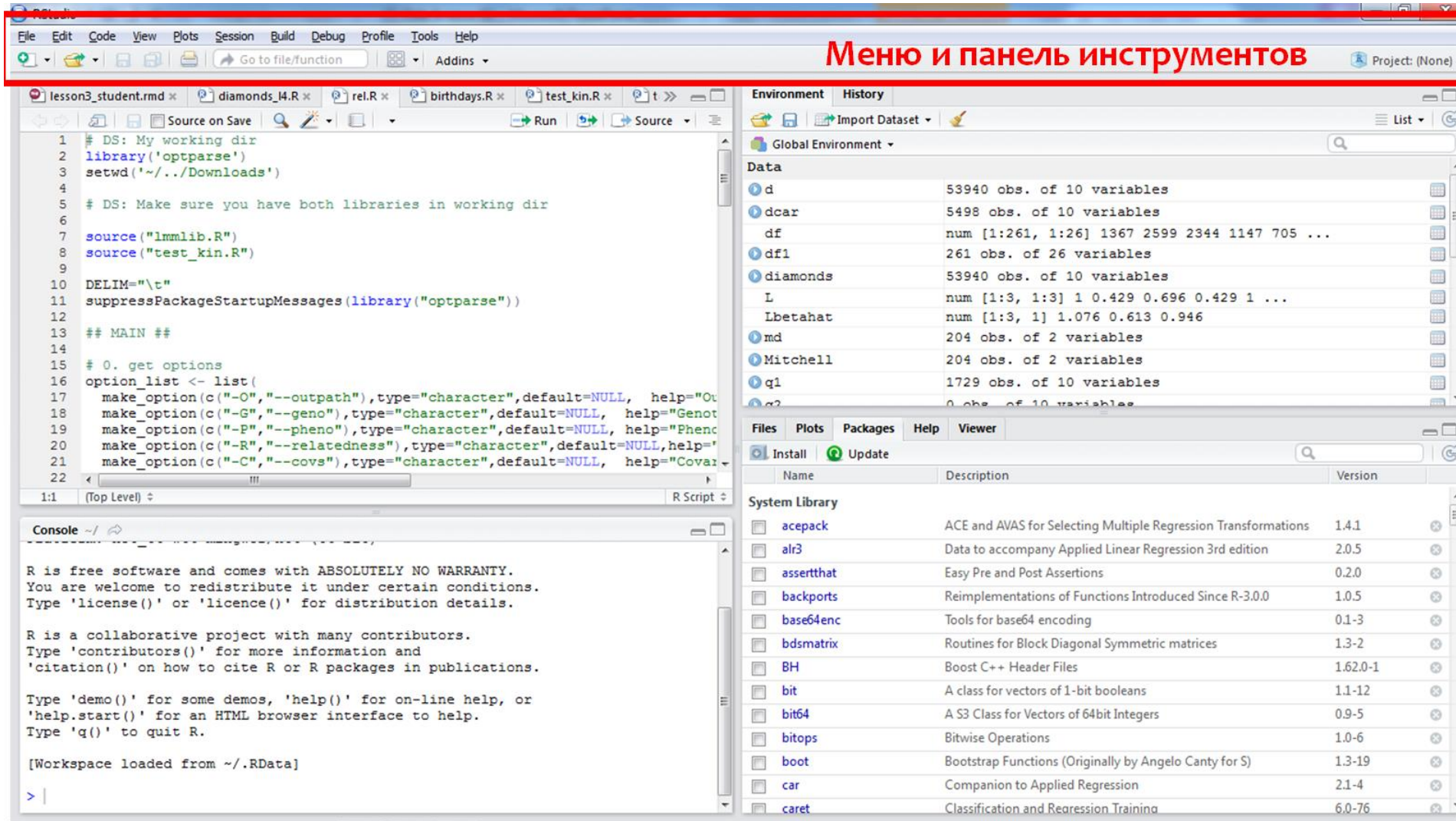


- R <https://cran.r-project.org/>
- RStudio
<https://www.rstudio.com/products/RStudio/>
- MS Excel

Характеристики R

- Высокого уровня
- Интерпретируемый (скриптовой)
- Не строго типизированный
- Регистрозависимый
- Кроссплатформенный
- Open Source

RStudio



Ctrl+L – очистить консоль

Ctrl+Shift+C – закомментировать

Ctrl+Enter – запустить выделенный код

Ctrl+Alt+R – запустить весь код

Основные типы данных

Типы данных:

- численные;
- логические;
- символьные.

Специальные объекты:

- Inf - бесконечность;
- NA - пропуск;
- NaN – не число.

Задания:

1. Даны стороны прямоугольника a и b . Найти его площадь $S = a \cdot b$ и периметр $P = 2 \cdot (a + b)$.
2. Даны два неотрицательных числа a и b . Найти их среднее геометрическое, то есть квадратный корень из их произведения.
3. Даны координаты двух противоположных вершин прямоугольника: (x_1, y_1) , (x_2, y_2) . Стороны прямоугольника параллельны осям координат. Найти периметр и площадь данного прямоугольника.
4. Дано целое число A . Проверить истинность высказывания: «Число A является нечетным».
5. Даны три целых числа: A , B , C . Проверить истинность высказывания: «Справедливо двойное неравенство $A < B < C$ ».
6. Даны три целых числа: A , B , C . Проверить истинность высказывания: «Хотя бы одно из чисел A , B , C положительное».
7. (ДЗ) Даны три целых числа: A , B , C . Проверить истинность высказывания: «Ровно одно из чисел A , B , C положительное».

Основные структуры данных

Структуры данных:

- массивы;
- матрицы;
- факторы;
- списки;
- data frame (таблицы данных);
- временные ряды.

Описательная статистика

`mean()` – среднее значение (**не путать с математическим ожиданием!!!!!!**)

`median()` – медиана

`sum()` – сумма

`range()` – разброс

`min()` – минимальное значение

`max()` – максимальное значение

`var()` – вариация (**не путать с дисперсией!!!!**)

`sd()` – стандартное отклонение

`quantile()` - квантили

`summary()` – вывод основной статистики

Элементарные графики:

`plot()` – точечный график

`hist()` – гистограмма

`barplot()` – коробчатая диаграмма

Задания

1. Изучить функцию `gnorm()`, разобраться зачем она, какие у нее параметры.
2. Создать 2 массива по 100 случайных чисел используя `gnorm` (параметры выбрать произвольно). Посчитать среднее значение и стандартное отклонение каждого массива, сравнить с параметрами функции.
3. Построить точечный график, где X – элементы 1 массива, в Y - второго.
4. Построить гистограммы элементов каждого массива.
5. Увеличить элементы 1 массива в 3 раза.
6. От всех элементов второго массива, которые больше среднего значения, отнять 18.