



ОНЛАЙН-ОБРАЗОВАНИЕ

Базовые инструменты анализа данных в Python

Стройкова Ксения



Проверка звука

Напишите в чат:

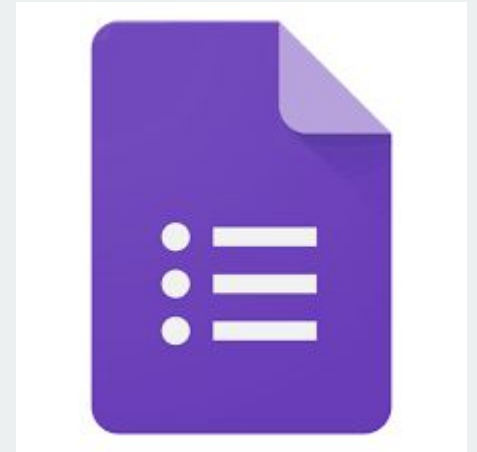
0 - ничего не слышно

1 - плохо слышно

2 - хорошо слышно



Опрос о слушателях курса




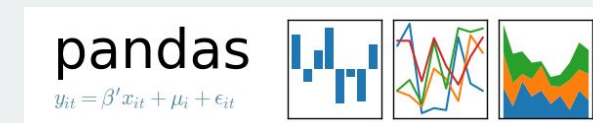
1. Выяснить входные знания
2. Выяснить мотивацию
3. Получить пожелания

<https://goo.gl/forms/kna8Ray3sZINcFma2>



Сегодня на занятии

1. О курсе
2. Введение в Python  python™
3. Стек научных библиотек в Python
4. API в sklearn
5. Домашнее задание



0 курсе

4 модуля

В каждом модуле 8 занятий (2 раза в неделю)

Основы машинного обучения

Применение машинного обучения, дополнительные алгоритмы

Применение машинного обучения в бизнесе

Работа с большими объемами данных

Домашние задания

В каждом модуле 4 ДЗ.

5й модуль - Проект

MVP с использованием машинного обучения

Необходимо скачать данные, предсказать некоторую характеристику.



Почему Python?

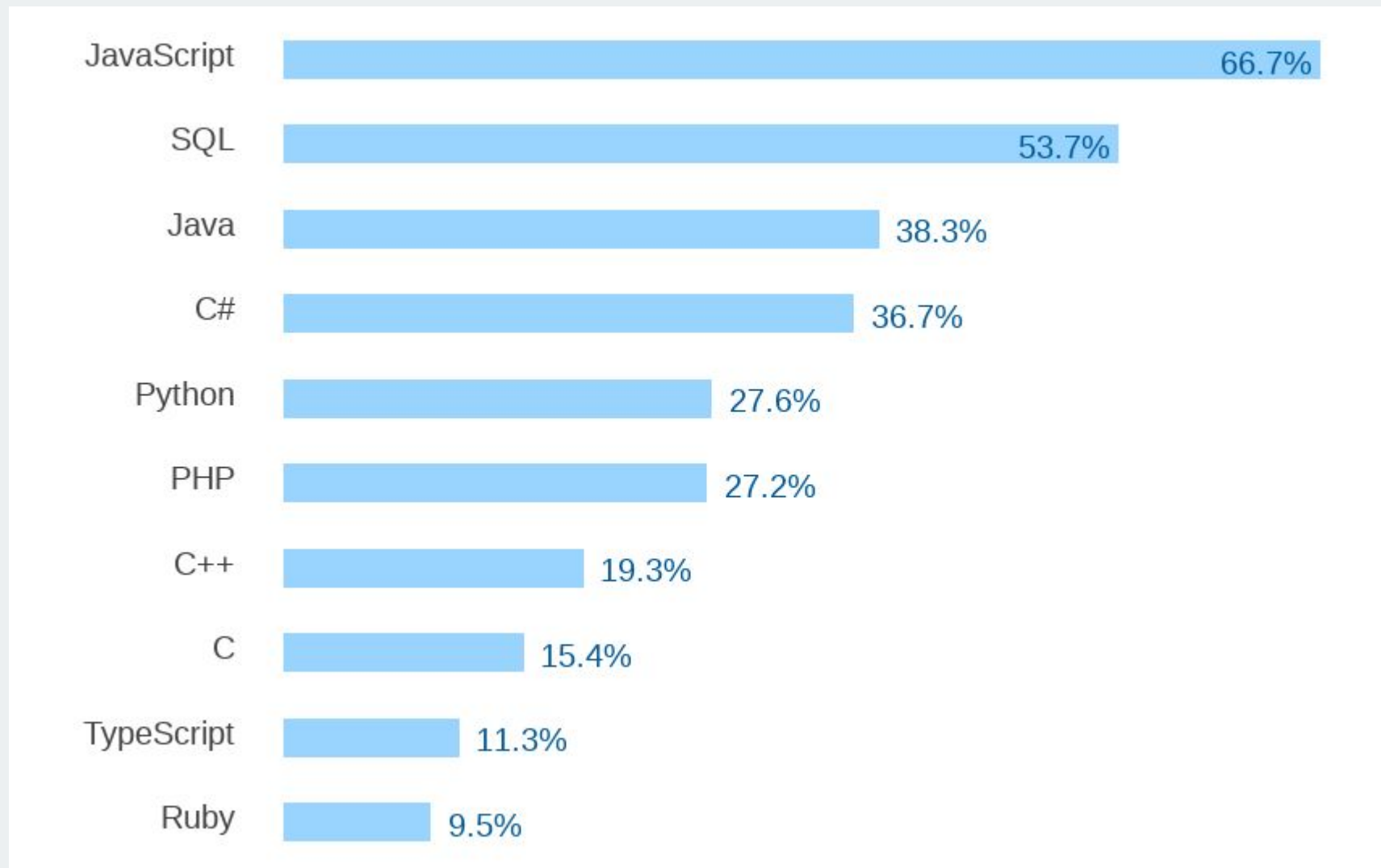
**Python is powerful... and fast;
plays well with others;
runs everywhere;
is friendly & easy to learn;
is Open.**

- **Web and Internet Development**
- **Database Access**
- **Desktop GUIs**
- **Scientific & Numeric**
- **Education**
- **Network Programming**
- **Software & Game Development**

<https://www.python.org/about/>



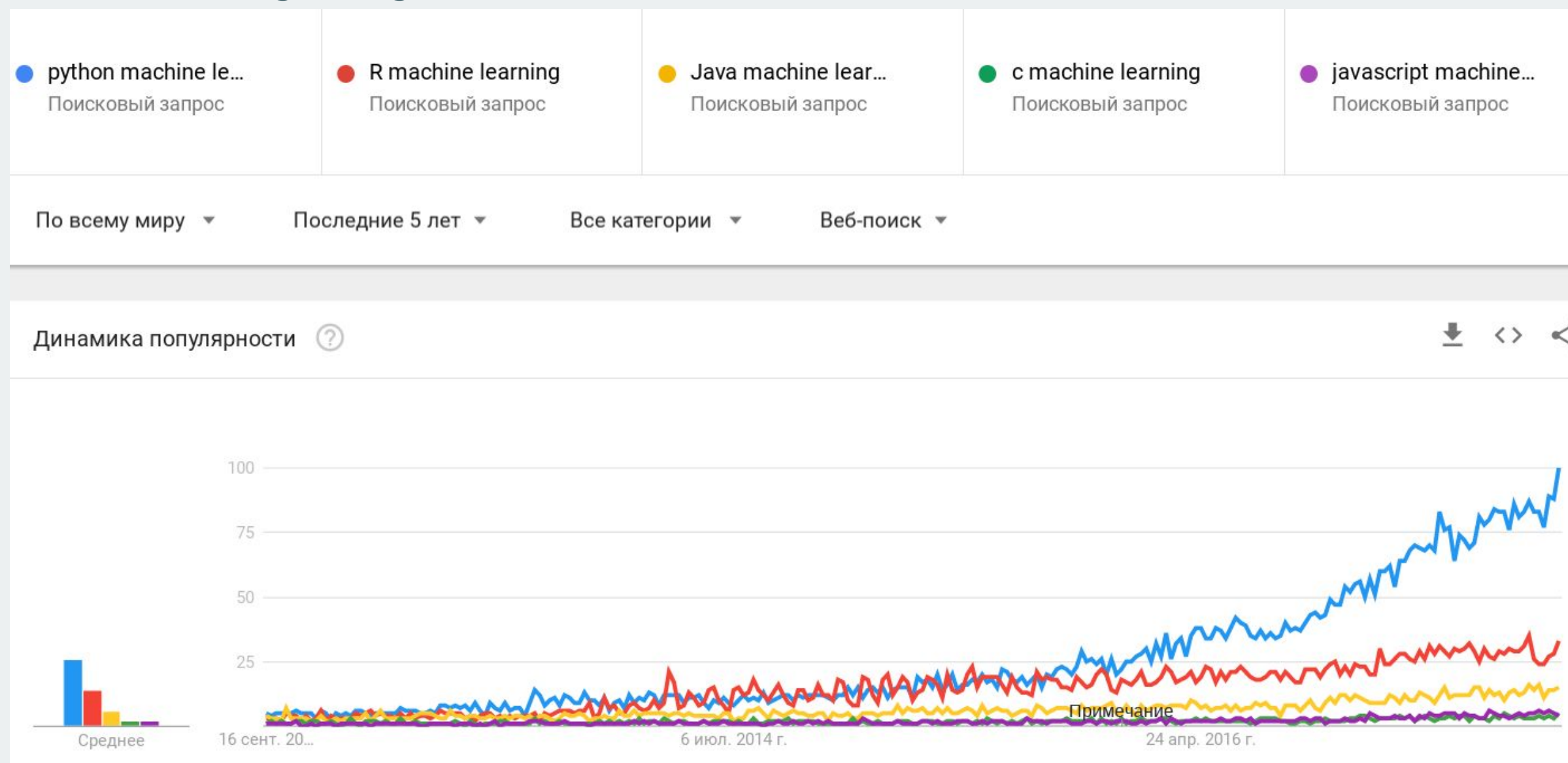
Почему Python на самом деле?



<https://insights.stackoverflow.com/survey/2017>



Почему Python на самом деле?



<https://goo.gl/vGMRcB>



ML библиотеки на Python

- NumPy
- SciPy
- Pandas
- Matplotlib, Seaborn, Bokeh, Plotly
- Sci-Kit-Learn
- Keras, Tensorflow, Theano
- NLTK, Gensim
- ...

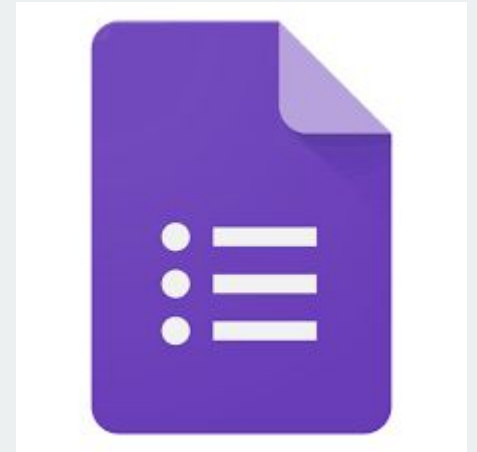
<https://goo.gl/vGMRcB>



Опрос

1. Программирование на Python
2. Стек научных библиотек в Python
3. API в sklearn
4. Базовые термины в машинном обучении

<https://goo.gl/forms/VPvQi8PXHRJi2Acq2>



Сгенерируем датасет студентов ВУЗа

Пользователь

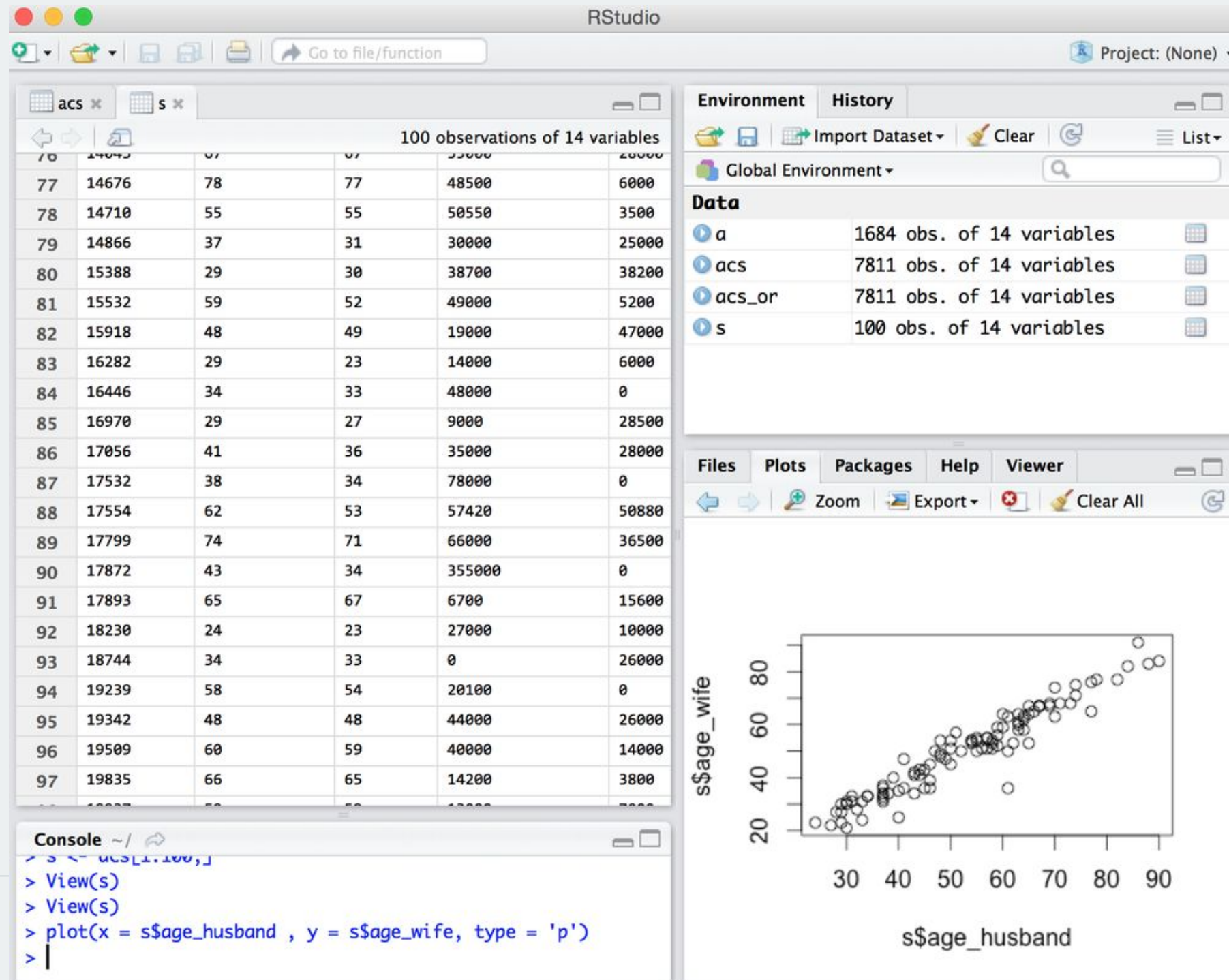
- Имя
- Возраст
- Пол
- Дата зачисления
- Факультет
- Средний балл
- Наличие медали



Сгенерируем датасет студентов ВУЗа



Другие инструменты - R





Другие инструменты - R

RStudio

Project: (None)

acs x

7811 observations of 14 variables

	household	age_husband	age_wife	income_husband	income_wife
1	48	64	62	11000	29200
2	218	63	64	100000	31000
3	279	56	51	31000	0
4	612	71	68	51700	8800
5	947	37	33	16600	26000
6	1373	86	91	77500	30000
7	1733	67	67	8400	4800
8	1858	70	74	73670	11000
9	1947	33	31	55050	600
10	1962	41	47	42000	36000

Displayed 1000 rows of 7811 (6811 omitted)

Environment History

Import Dataset Clear List

Global Environment

Data

- a 1684 obs. of 14 variables
- acs 7811 obs. of 14 variables
- acs_or 7811 obs. of 14 variables

Files Plots Packages Help Viewer

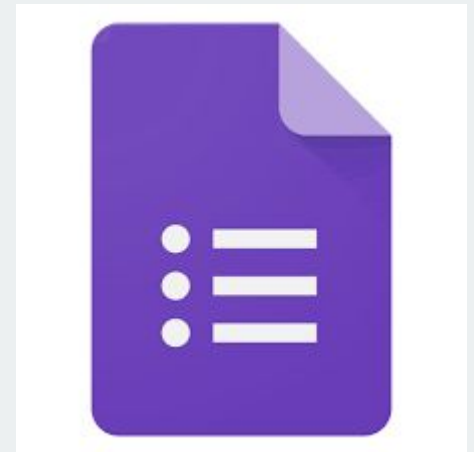
Zoom Export Clear All

Console ~/

```
> mean(acs$age_husband)
[1] 54.31776
> median(acs$age_wife)
[1] 53
> quantile(acs$age_wife)
 0% 25% 50% 75% 100%
19  40  53  63  95
> min(acs$age_wife)
[1] 19
> max(acs$age_wife)
[1] 95
> var(acs$age_wife)
[1] 220.527
> sd(acs$age_wife)
[1] 14.85015
> |
```



Опрос по пройденному материалу



1. Программирование на Python
2. Стек научных библиотек в Python
3. API в sklearn
4. Базовые термины в машинном обучении

<https://goo.gl/forms/YS2sB4z0EaVzyFUn1>



Домашнее задание



Материалы для дальнейшего изучения

1. <http://docs.python-guide.org/en/latest/writing/structure/>
2. <https://www.kdnuggets.com/2017/02/5-career-paths-data-science-big-data-explained.html>
3. <https://monkeylearn.com/blog/gentle-guide-to-machine-learning/>
4. <https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/>

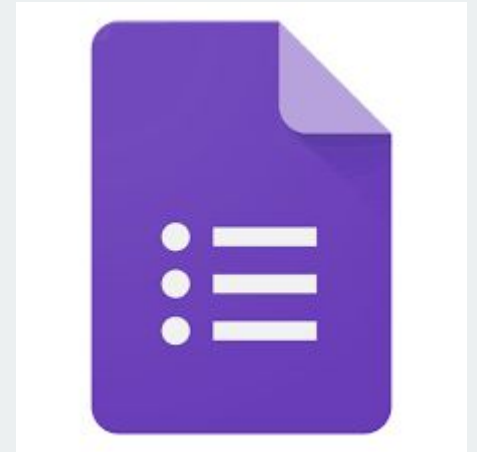


Roadmap

1. Знание языка программирования (Python, R, Java, Scala, Kotlin, ...)
2. Организация проекта, принятая в языке (тестирование, сборка, раскладка)
3. Поиск данных для анализа с бизнес-целью (учебные данные, подготовленные данные, API, неочищенные данные)



Обратная связь



Нам важно ваше мнение

<https://otus.ru/polls/schedule/BigData>





**Спасибо
за внимание!**