РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ЭМИТ

Введение в анализ данных с Python

Зам. Декана ЭФ ЭМИТ РАНХиГС,
проф. кафедры Эконометрики и математической экономики ЭМИТ РАНХиГС
д.т.н. Шилин Кирилл Юрьевич

РАНХиГС каб. 419/3
email: kshilin@ranepa.ru

Программа курса

1 часть - 2 недели, 16 часов

Anaconda + Jupiter
DataCamp
GitHub
Основы программирования Python
Основы NumPy и Pandas
Парсинг данных

2 часть - 2 недели, самостоятельная работа

Курсы DataCamp

3 часть - 2 недели, 16 часов

Разбор Datasets (самостоятельное программирование под руководством преподавателя)

Аттестация - результаты DataCamp

Что такое анализ данных?

Анализ данных (современная интерпретация):

1. Представление данных в табличном виде
2. Очистка и заполнение пропущенных данных
3. Переформатирование данных
4. Комбинирование
5. Нормализация (сейчас уже в библиотеках машинного обучения)
6. Срезы данных
7. Преобразование данных (агрегирование)
8. Статистический анализ
9. Визуализация

Фактически - современный анализ данных, это подготовка данных для последующей обработки, например алгоритмами машинного обучения.

Современный анализ данных это Excel без ограничений

Траектория подготовки

Анализ данных
библиотека Pandas

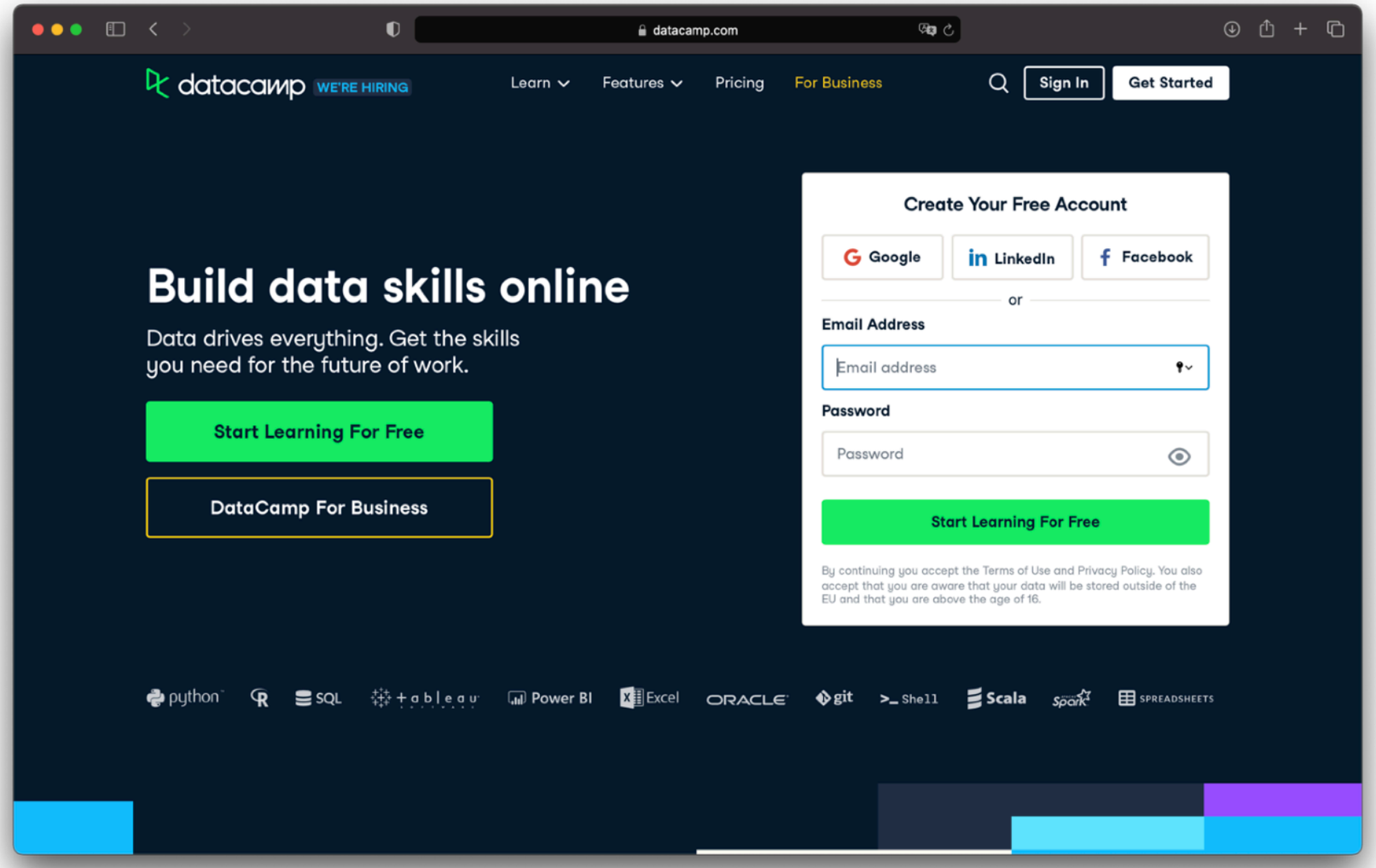
Классические методы машинного обучения (признаки создает человек)
библиотека Scikit-Learn

Продвинутый градиентный бустинг
библиотека XGBoost (Google) или CatBoost (Yandex)

Глубокое обучение (нейросеть) (признаки создает компьютер)
библиотеки TensorFlow и Keras или PyTorch

Вероятностное программирование (Байесовские методы)
библиотеки PyMC3

DataCamp



Практика программирования на R и Python. Около 350 курсов

для преподавателя бесплатно, достаточно написать письмо

DataCamp

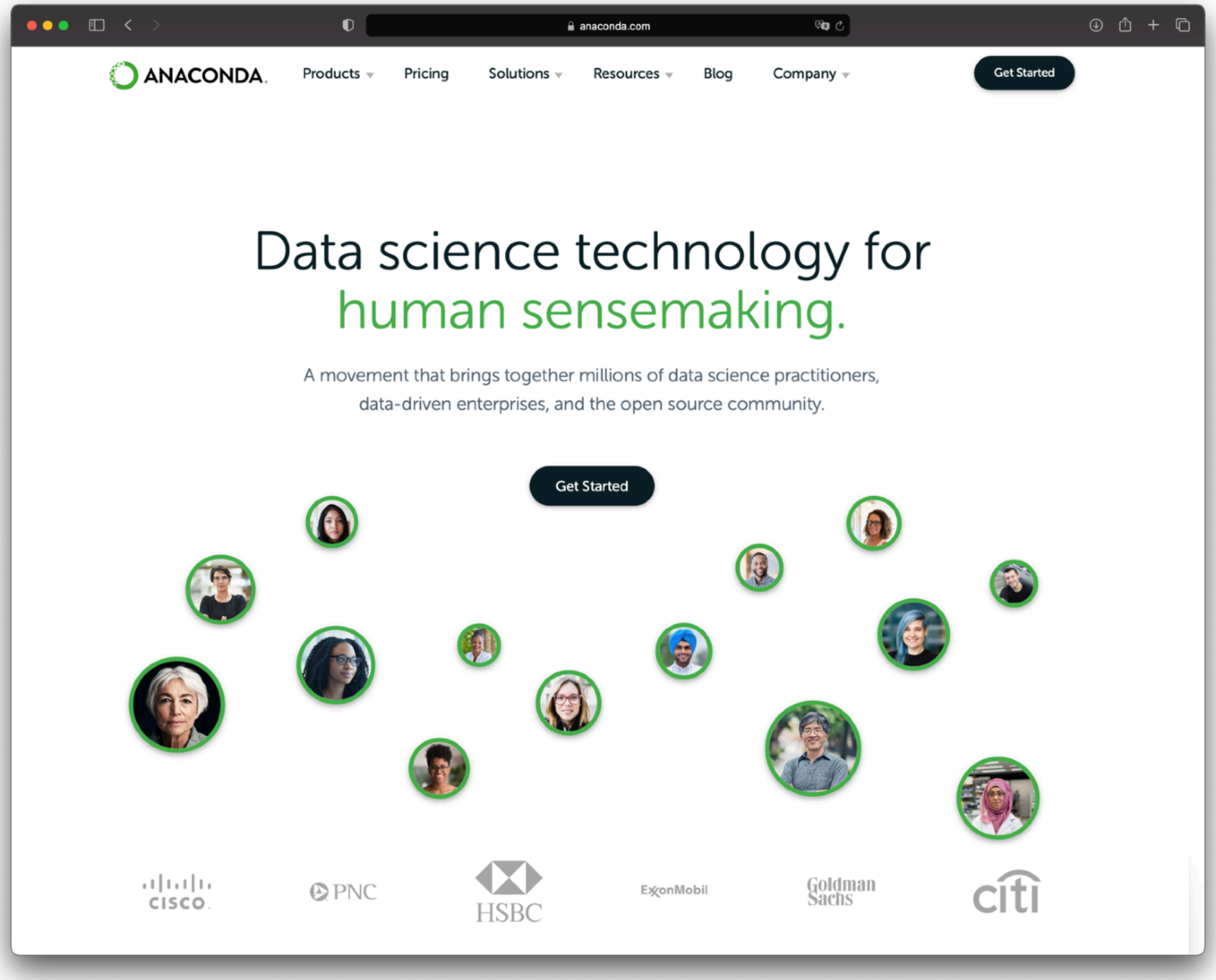
Обязательные курсы:

1. Introduction to Python
2. Intermediate Python
3. Introduction to Data Science in Python
4. Data Manipulation with pandas
5. Manipulating Time Series Data in Python
6. Introduction to Data Visualization with Seaborn

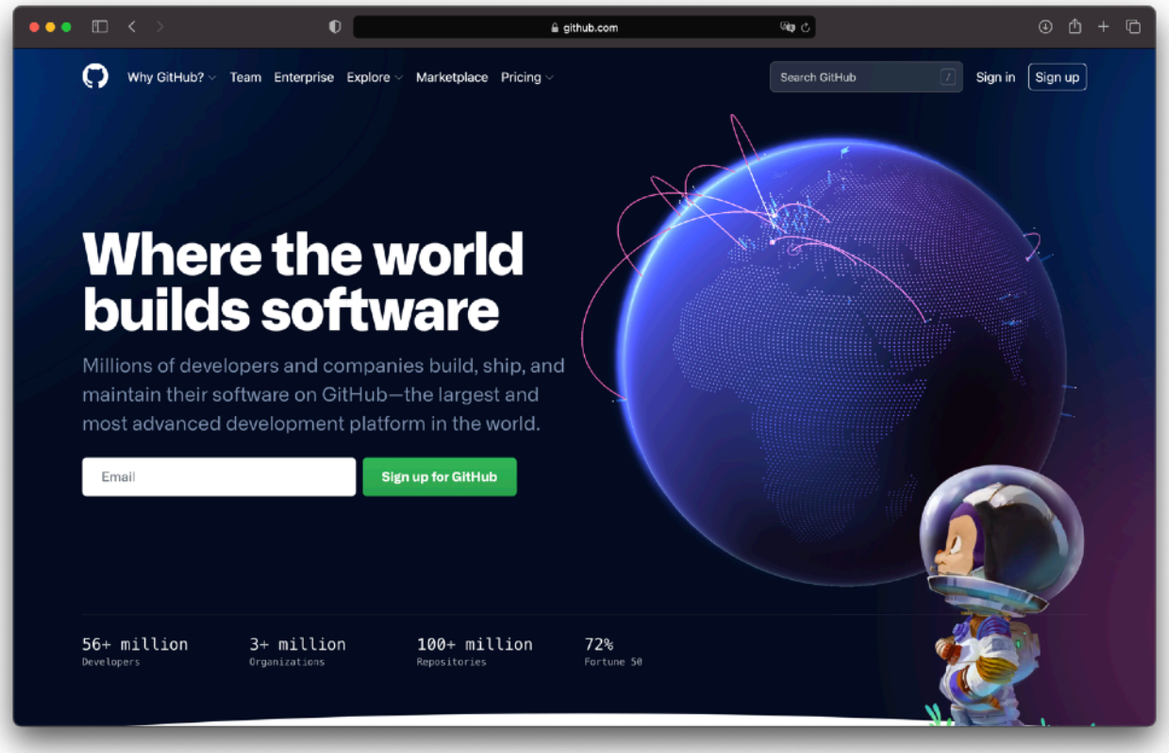
Рекомендовано:

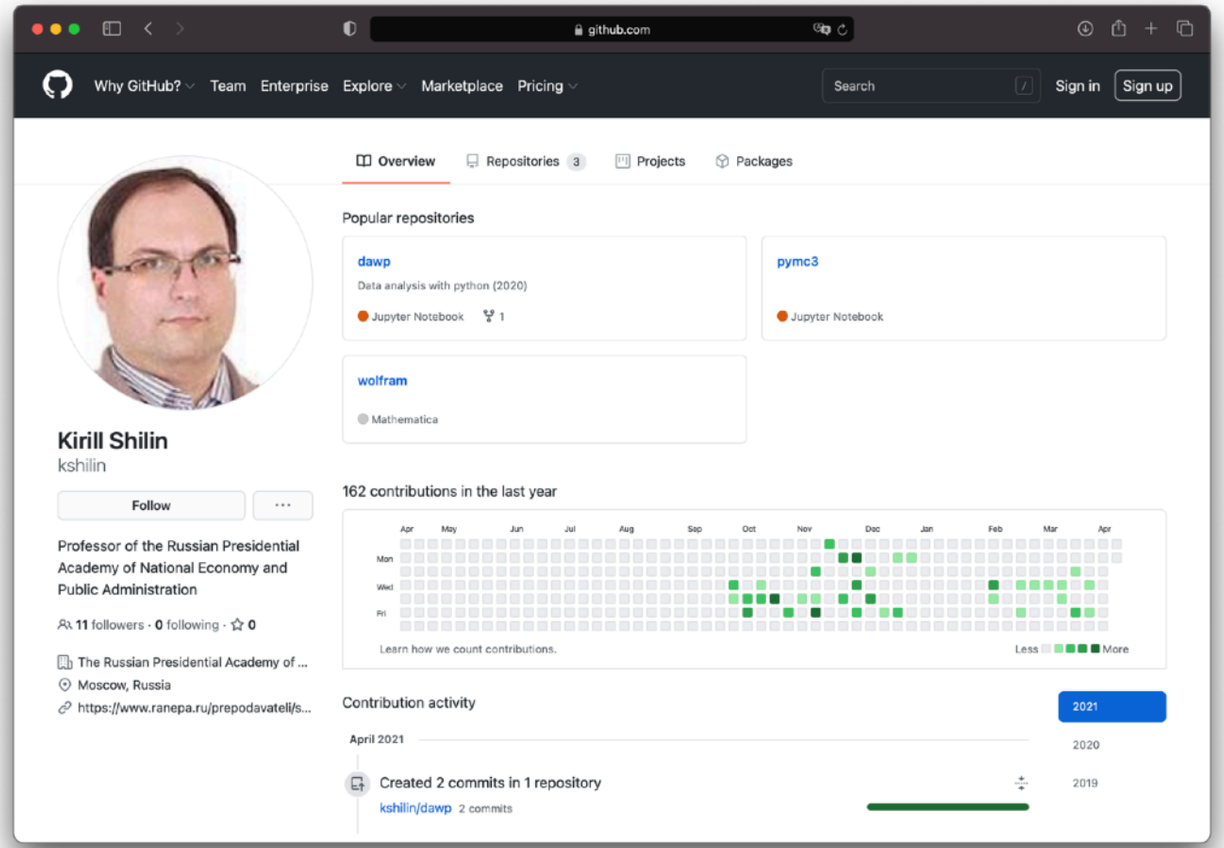
1. Introduction to Data Visualization with Matplotlib
2. pandas Foundations

Anaconda + Jupiter



Знакомство с GitHub



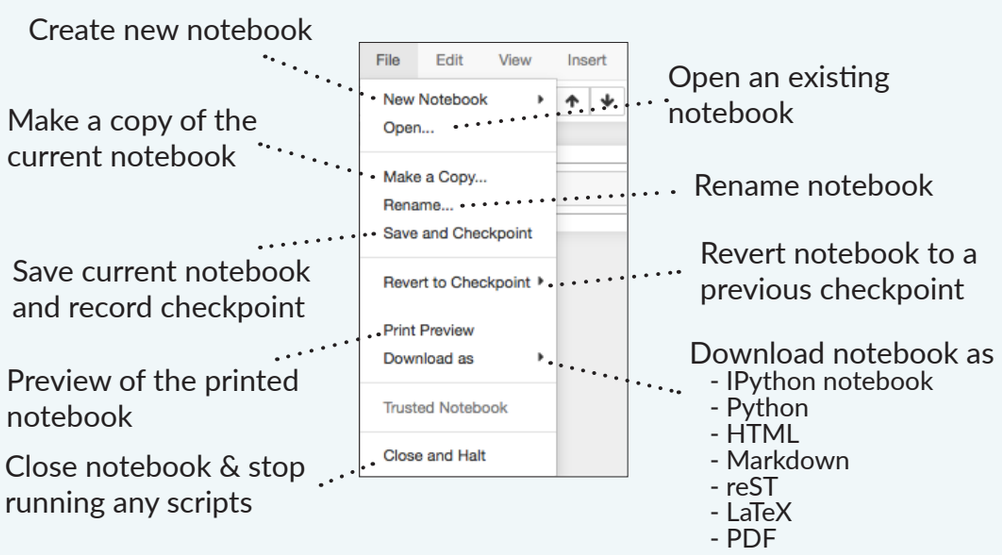


Jupyter

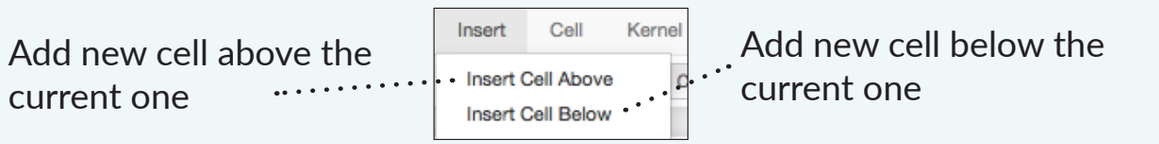
Шпаргалка (Cheat Sheet) для Jupyter Notebook [ссылка]

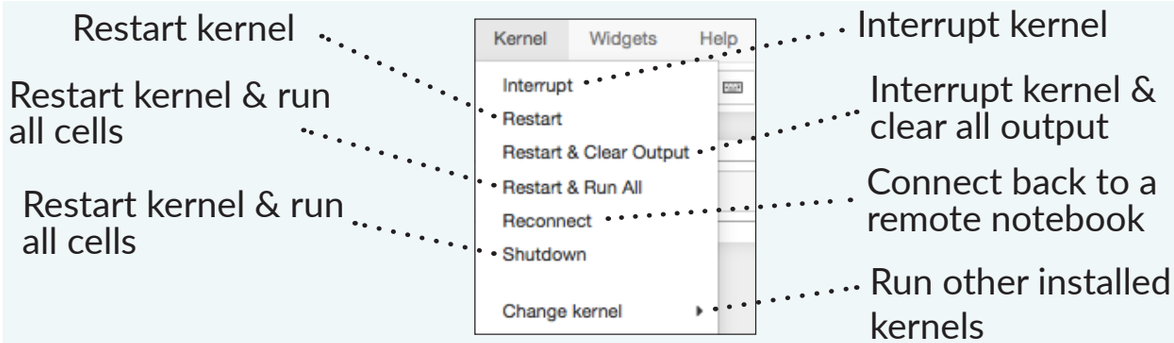
Command Mode:

Saving/Loading Notebooks



Insert Cells





Jupyter

Основные возможности:

1. Поля для ввода кода
2. Поблочное выполнение кода
3. Поля для ввода заголовков
4. Поля для ввода текста с формулами LaTeX

Внимание! команда для выполнения блока **<Shift>+<Enter>**

Ввод данных в режиме **Markdown**

Заголовки

Заголовок 1 уровня: **#** или **<h1>…</h1>**
Заголовок 2 уровня: **##** или **<h2>…</h2>**
Заголовок 3 уровня: **###** или **<h3>…</h3>**

Формула TeX: в выделенной строке **$$**\int_0^\pi \sin^2(x) dx**$$**
внутри строки **$**\int_0^\pi \sin^2(x) dx**$**

Конец абзаца: В конце абзаца поставить **
**