

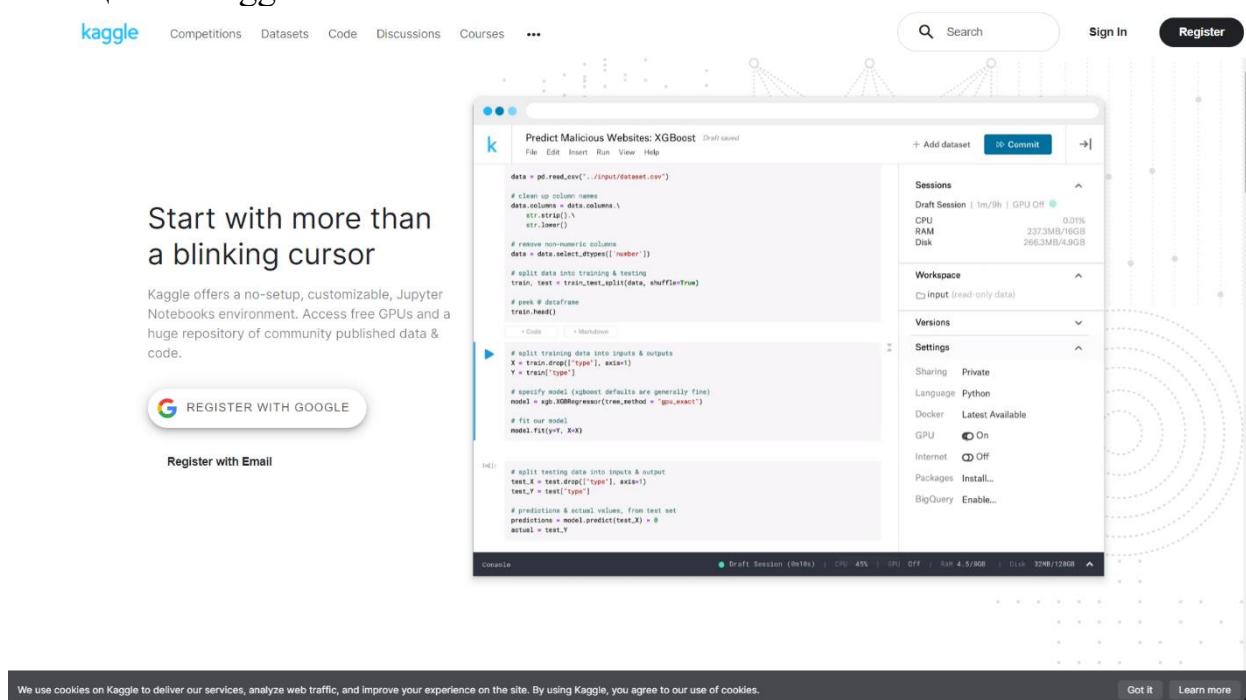
Первый взгляд и начало работы с Kaggle

Введение

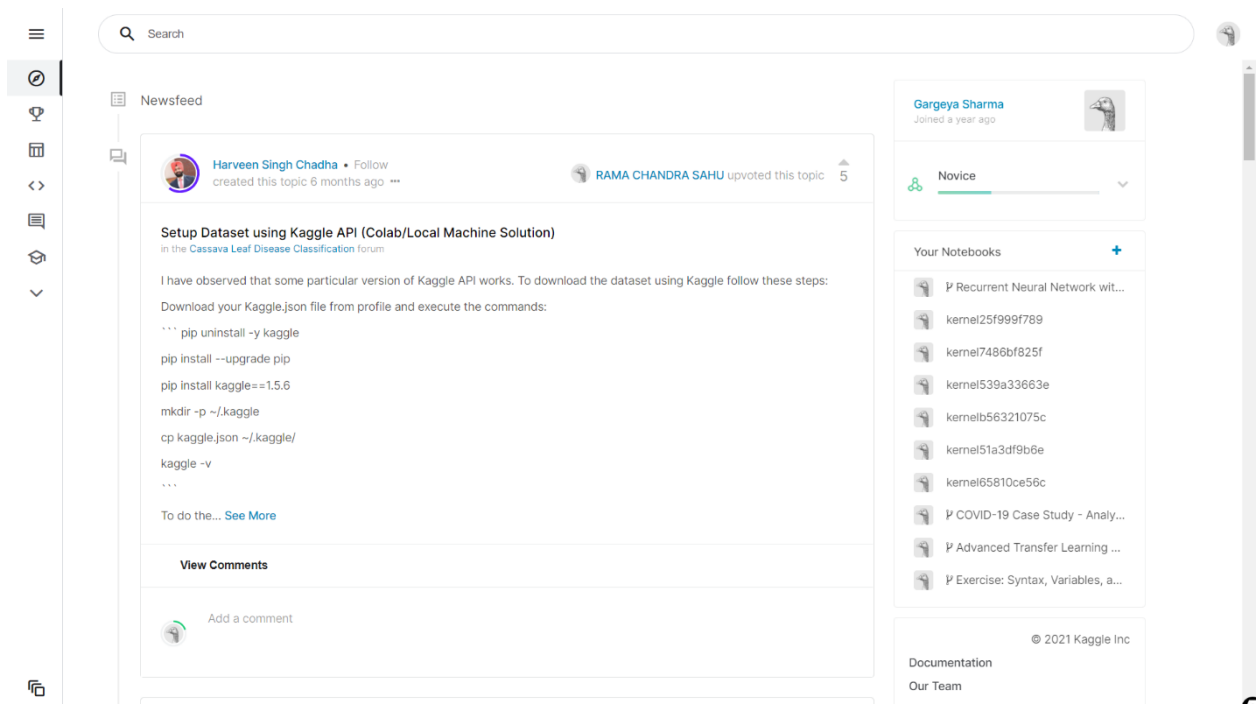
Сегодня в каждой карьере должно быть сообщество, группа людей, с которыми мы можем говорить о работе, ошибках, идеях и учиться. Kaggle - самое популярное и крупнейшее в мире сообщество специалистов по обработке и анализу данных. Наличие такого сообщества помогает нам чувствовать свою принадлежность, что является одним из важнейших чувств для нашего социального взаимодействия и здоровья.

В этой статье мы рассмотрим Kaggle как целое сообщество и Kaggle как платформу: все его различные инструменты, сервисы и ресурсы, доступные для нас, чтобы **изучать**, а также **практиковать науку о данных**.

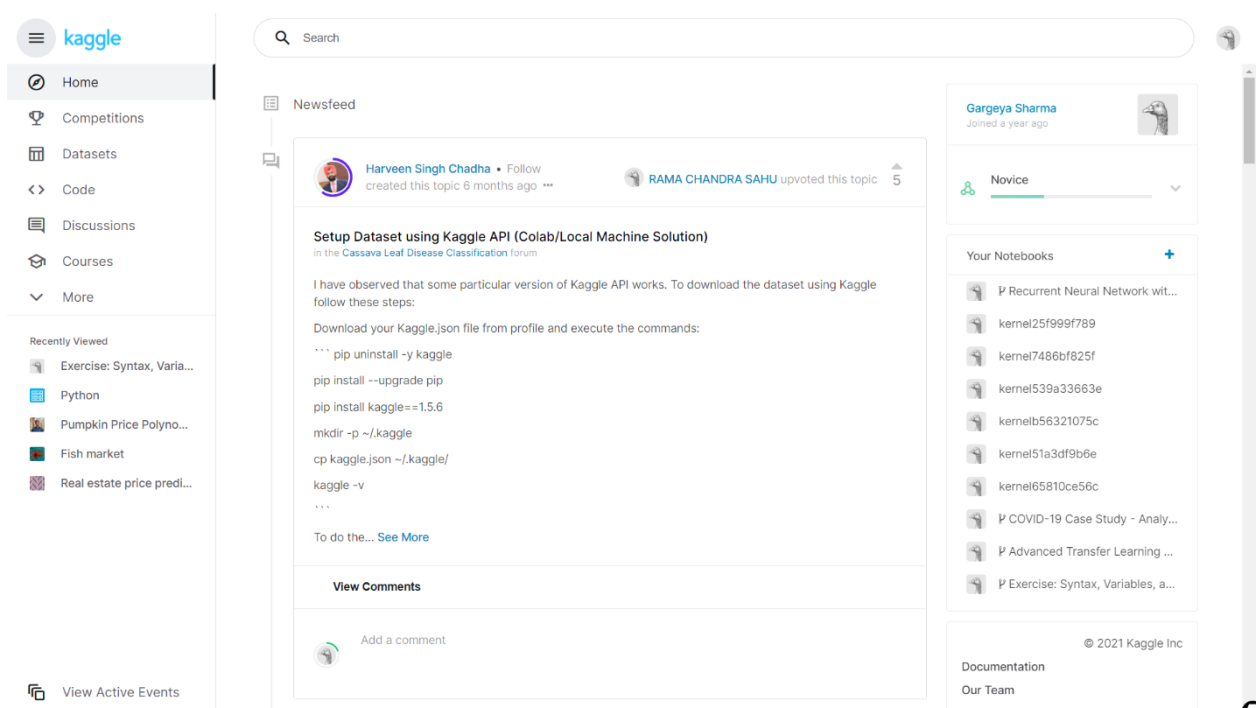
Давайте посмотрим на интерфейс, который мы получаем при первом посещении Kaggle.



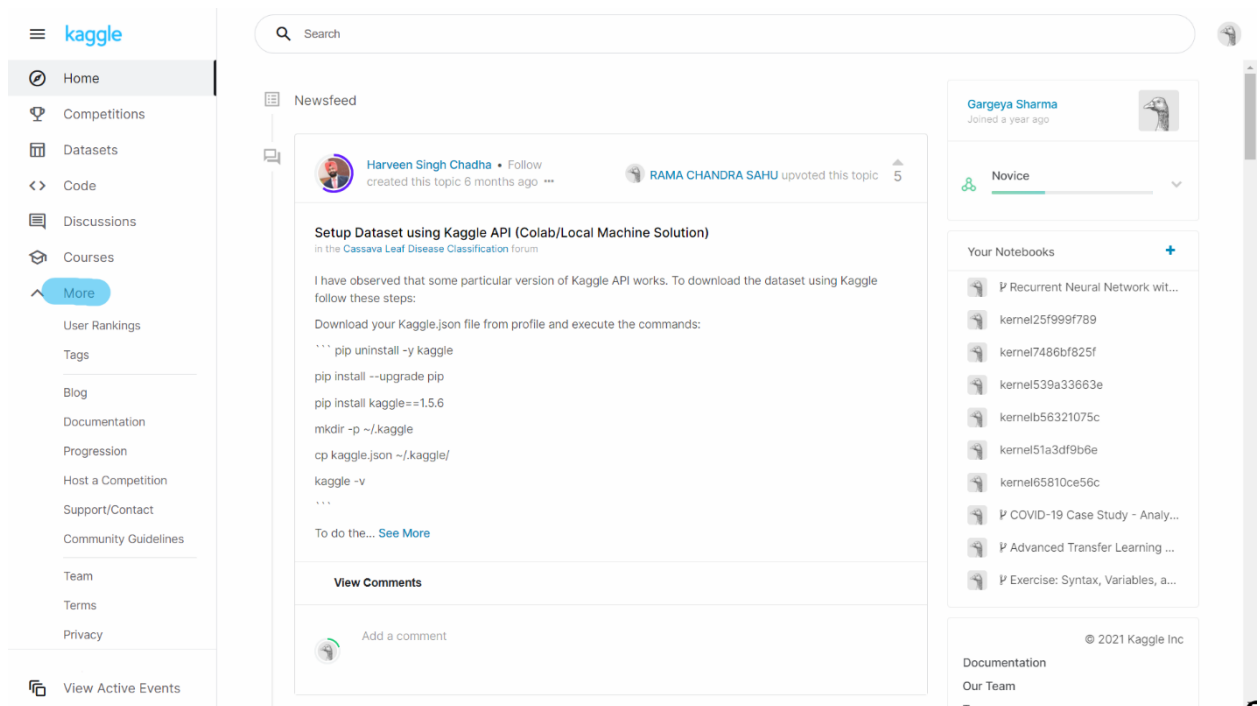
Прежде чем мы начнем использовать Kaggle, нам нужно создать учетную запись, а затем войти в систему, вы можете увидеть оба варианта в правом верхнем углу. Когда вы закончите с этим, это может выглядеть следующим образом:



Некоторые из видимых здесь вещей могут отличаться у вас, потому что интерфейс персонализирован в соответствии с тем, кто и как использовал Kaggle до сих пор с момента регистрации.



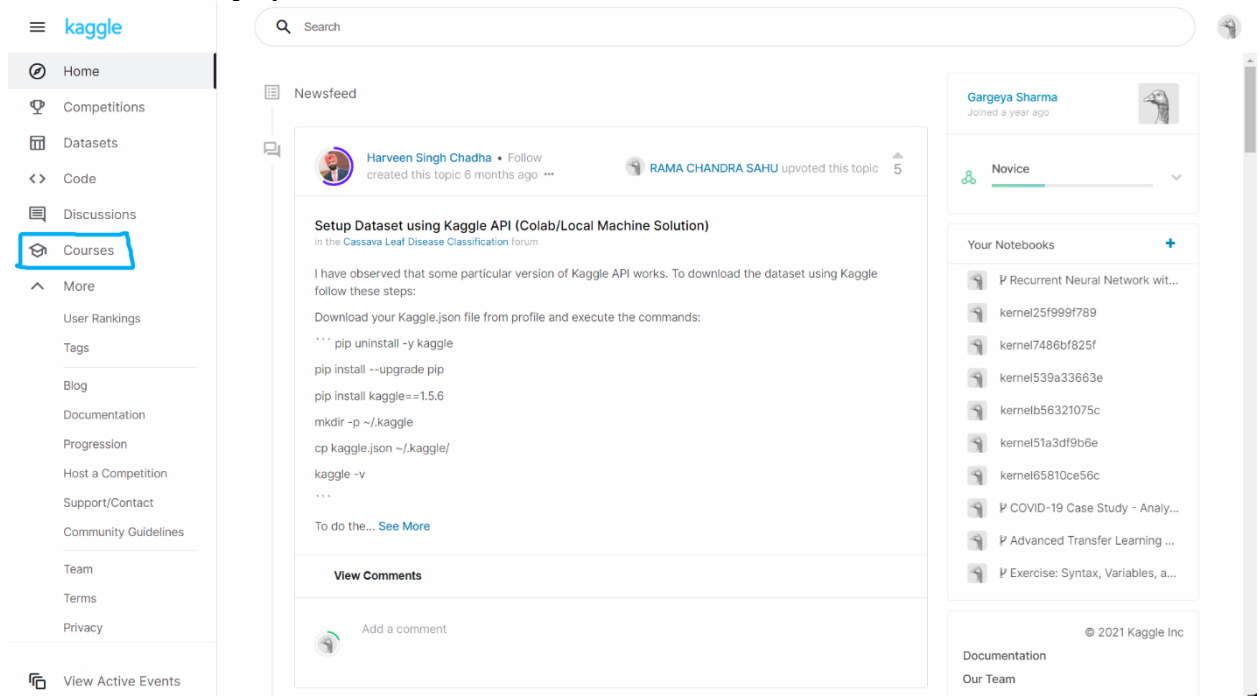
Как только нажимаешь «Home», это все, к чему я могу получить доступ из своей учетной записи Kaggle.



4 основные вещи, которые делают Kaggle «САМЫМ ЛУЧШИМ».

1. Доступны бесплатные курсы и бесплатные сертификаты.

Существует так много курсов по разным областям машинного обучения и науки о данных. После каждого урока доступны не только курсы, но и практические тетради (тетради с упражнениями) для практического изучения темы. Для получения бесплатного сертификата Kaggle необходимо выполнить все задания и упражнения.



The image shows the Kaggle website's 'Courses' section. On the left is a sidebar with navigation links: Home, Competitions, Datasets, Code, Discussions, Courses (selected), and More. Below these are 'Recently Viewed' items and a 'View Active Events' link. The main content area displays a list of courses, each with a colored icon, a title, a brief description, and a progress indicator. The courses listed are: Python, Intro to Machine Learning, Intermediate Machine Learning, Pandas, Data Visualization, Feature Engineering, Data Cleaning, Intro to SQL, Advanced SQL, Intro to AI Ethics, Intro to Deep Learning, Computer Vision, Geospatial Analysis, Machine Learning Explainability, and Microchallenges. A search bar is located at the top of the course list.

Course Title	Description	Progress
Python	Learn the most important language for data science.	9%
Intro to Machine Learning	Learn the core ideas in machine learning, and build your first models.	
Intermediate Machine Learning	Handle missing values, non-numeric values, data leakage, and more.	
Pandas	Solve short hands-on challenges to perfect your data manipulation skills.	
Data Visualization	Make great data visualizations. A great way to see the power of coding!	
Feature Engineering	Better features make better models. Discover how to get the most out of your data.	
Data Cleaning	Master efficient workflows for cleaning real-world, messy data.	
Intro to SQL	Learn SQL for working with databases, using Google BigQuery.	
Advanced SQL	Take your SQL skills to the next level.	
Intro to AI Ethics	Explore practical tools to guide the moral design of AI systems.	
Intro to Deep Learning	Use TensorFlow and Keras to build and train neural networks for structured data.	
Computer Vision	Build convolutional neural networks with TensorFlow and Keras.	
Geospatial Analysis	Create interactive maps, and discover patterns in geospatial data.	
Machine Learning Explainability	Extract human-understandable insights from any model.	
Microchallenges	Solve ultra-short challenges to build and test your skill.	

Есть еще несколько курсов, но с помощью этого я хотел показать вам, что на этих курсах существует такое разнообразие тем, что вам не нужно никуда идти, в любое время, чтобы почувствовать себя потерянным по какой-либо теме или проблеме, получить помощь отсюда.

Позвольте мне показать вам, как, как выглядят эти курсы на одном примере:

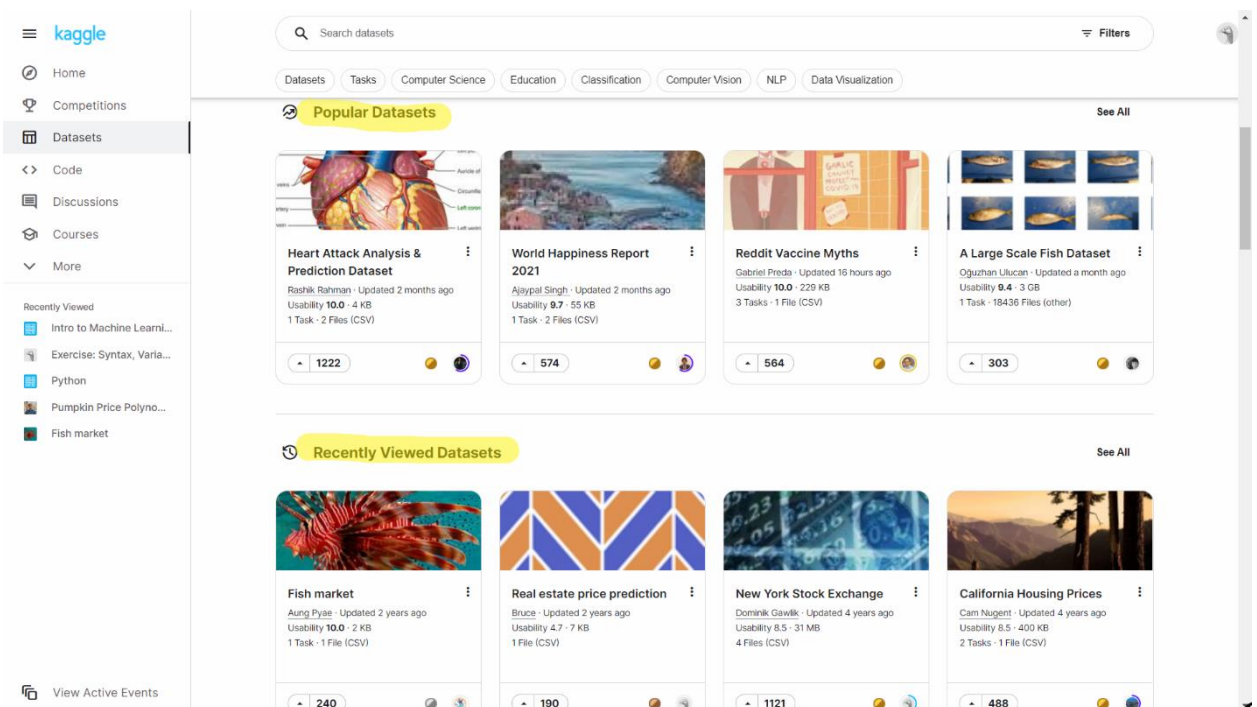
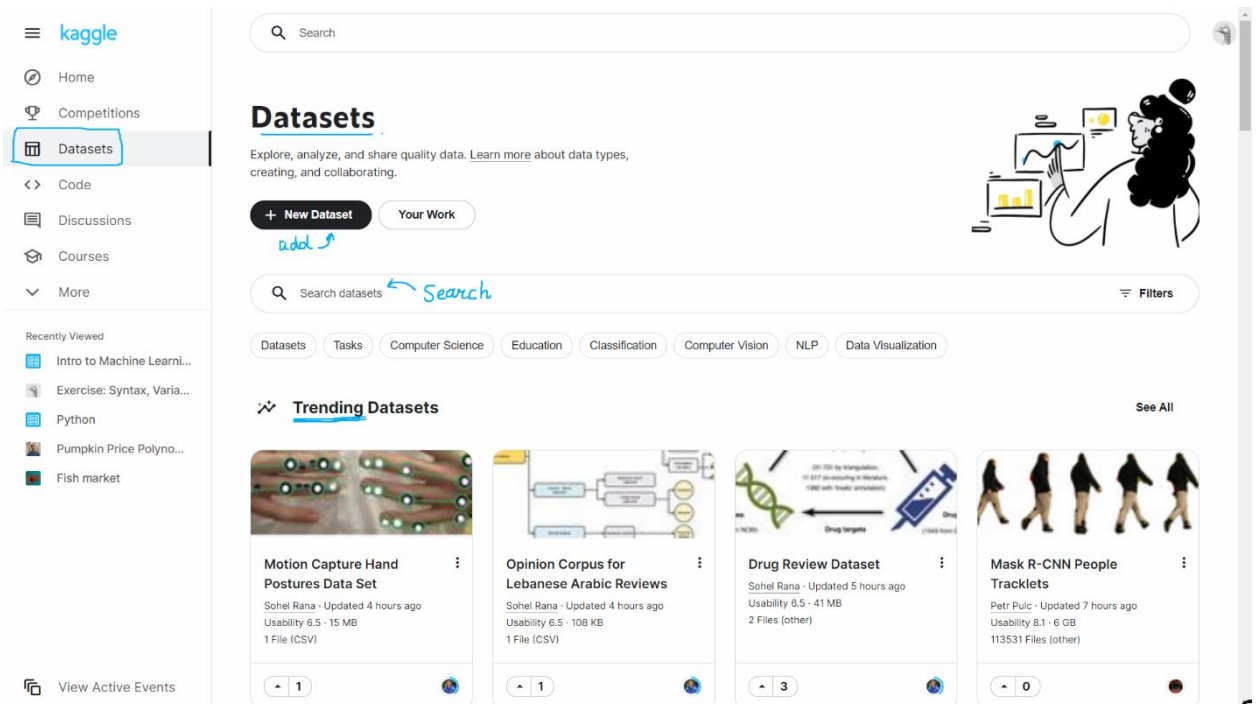
The screenshot displays the Kaggle 'Intro to Machine Learning' course interface. On the left, a sidebar contains navigation links: Home, Competitions, Datasets, Code, Discussions, Courses (selected), and More. Below these are 'Recently Viewed' items and 'View Active Events'. The main content area features a search bar, the course title 'Intro to Machine Learning', and a subtitle 'Learn the core ideas in machine learning, and build your first models.' A list of lessons follows, each with a title, description, and icons for tutorial and exercise. The lessons are: 1. How Models Work, 2. Basic Data Exploration, 3. Your First Machine Learning Model, 4. Model Validation, 5. Underfitting and Overfitting, 6. Random Forests, 7. Machine Learning Competitions, and a Bonus Lesson: Intro to AutoML. On the right, a 'Your Progress' section shows a 0% completion rate and a 'Begin today!' button. Below this is an 'Overview' section with a 3-hour duration and 8 lessons. A 'Prerequisite Skills' box lists Python and several Kaggle courses. The 'Instructor' section introduces Dan Becker, a Data Scientist with extensive experience in data science and machine learning.

В конце каждого курса есть один дополнительный урок, который отличается по содержанию, но похож на вариант использования и понимание курса. В основном они включают какую-то известную или / или важную тему. Здесь у нас есть AutoML (от Google) для автоматизации машинного обучения.

2. Огромная коллекция общедоступных / предоставленных наборов данных для практики / работы над

Для решения любых задач в области науки о данных, машинного обучения или глубокого обучения нам нужны данные, и большую часть времени их большую часть времени. Вместо того, чтобы просматривать разные сайты для разных типов / размеров наборов данных, Kaggle предоставляет общее место для огромной коллекции всех этих наборов данных. Вы можете одним

щелчком мыши отказаться от их использования. Они чрезвычайно просты в использовании.



Вот что вы увидите, щелкнув «Наборы данных» на панели навигации. Вы можете искать определенный набор данных, импортировать / вносить свой собственный набор данных в сообщество или изучать или начать работу над набором данных, показанным на этой странице. (Трендовые наборы данных, популярные наборы данных, недавно просмотренные наборы данных) Для демонстрации я буду искать определенный набор данных («Набор данных солнечных пятен»). Посмотрим, как это выглядит.

The screenshot shows the Kaggle Datasets page. On the left is a sidebar with navigation links: Home, Competitions, Datasets (selected), Code, Discussions, Courses, and More. Below these are 'Recently Viewed' items. The main area has a search bar with 'sunspot time series' entered. Below the search bar are category filters: Datasets, Tasks, Computer Science, Education, Classification, Computer Vision, NLP, and Data Visualization. A '12 Datasets' badge is shown. The first dataset, 'Sunspots', is highlighted with a red checkmark and a red box around its '95' popularity score. Other datasets listed include 'Daily Sun Spot Data (1818 to 2019)', 'Kanzel Sunspot Data: May 1944 - Aug. 2020', 'Average Sun Spot Number', 'SunSpot', and 'Time Series Practice Datasets'.

Dataset Name	Author	Updated	Usability	File Type	Size	Popularity Score
Sunspots	Especuolide	Updated 4 months ago	Usability 8.2	1 File (CSV)	22 KB	95
Daily Sun Spot Data (1818 to 2019)	Abhinand	Updated 2 years ago	Usability 10.0	1 File (CSV)	681 KB	26
Kanzel Sunspot Data: May 1944 - Aug. 2020	David Jackson	Updated 9 months ago	Usability 3.8	1 File (CSV)	337 KB	2
Average Sun Spot Number	Sarah VCH	Updated 4 years ago	Usability 7.1	1 File (CSV)	2 KB	2
SunSpot	ThiruArasan Dayalan	Updated 3 years ago	Usability 4.1	1 File (CSV)	424 KB	1
Time Series Practice Datasets	Doug Crosswell	Updated 3 years ago	Usability 2.9	6 Files (CSV, other)	48 KB	4

Число в красном выделении - это количество людей, проголосовавших за наиболее релевантный / понравившийся вариант. Давайте изучим и подробно рассмотрим этот набор данных.

Есть много вещей, которые мы можем использовать, чтобы узнать больше об этих данных и немедленно начать работу.

- Вы можете скачать набор данных,
- создайте новый блокнот Kaggle с уже загруженным в него набором данных.
- Некоторые подробности о столбцах внутри данных.
- Действия, связанные с этими данными.
- И последнее, но не менее важное: все записные книжки, созданные и опубликованные до этой даты, используют эти данные.

kaggle

Home

Competitions

Datasets

Code

Discussions

Courses

More

Recently Viewed

Sunspots

Intro to Machine Learnl...

Exercise: Syntax, Varia...

Python

Pumpkin Price Polyno...

View Active Events

kaggle

Home

Competitions

Datasets

Code

Discussions

Courses

More

Recently Viewed

Sunspots

Intro to Machine Learnl...

Exercise: Syntax, Varia...

Python

Pumpkin Price Polyno...

View Active Events

Search

Dataset

Sunspots

Monthly Mean Total Sunspot Number - form 1749 to july 2018

Especuloide

updated 4 months ago (Version 3)

Data

Tasks

Code (46)

Discussion

Activity

Metadata

Download (70 KB)

New Notebook

Usability 8.2

License CC0: Public Domain

Tags earth and nature, astronomy

Description

Context

Sunspots are temporary phenomena on the Sun's photosphere that appear as spots darker than the surrounding areas. They are regions of reduced surface temperature caused by concentrations of magnetic field flux that inhibit convection. Sunspots usually appear in pairs of opposite magnetic polarity. Their number varies according to the approximately 11-year solar cycle.

Source: <https://en.wikipedia.org/wiki/Sunspot>

Content :

Monthly Mean Total Sunspot Number from 1749 to 2018

Start now, here!

Data Explorer

69.79 KB

Sunspots.csv

Sunspots.csv (69.79 KB)

Detail

Compact

Column

3 of 3 columns

Search

Data

Tasks

Code (46)

Discussion

Activity

Metadata

Download (70 KB)

New Notebook

Data Explorer

69.79 KB

Sunspots.csv

Sunspots.csv (69.79 KB)

Detail

Compact

Column

3 of 3 columns

About this file

Sunspots - Monthly Mean Total Sunspot Number

#

Date

Monthly Mean To...

0

3264

31Jan49

31Jan21

0

398

0

1749-01-31

96.7

1

1749-02-28

104.3

2

1749-03-31

116.7

3

1749-04-30

92.8

4

1749-05-31

141.7

5

1749-06-30

139.2

6

1749-07-31

158.8

7

1749-08-31

118.5

8

1749-09-30

126.5

9

1749-10-31

125.8

10

1749-11-30

264.3

11

1749-12-31

142.8

12

1750-01-31

122.2

13

1750-02-28

126.5

Range of values

Summary
1 file
3 columns

The screenshot shows the Kaggle interface for the 'Sunsports' dataset. The left sidebar contains navigation links: Home, Competitions, Datasets, Code, Discussions, Courses, and More. Below this is a 'Recently Viewed' section with items like 'Sunsports', 'Intro to Machine Learning', 'Exercise: Syntax, Variables', 'Python', and 'Pumpkin Price Polynomial'. The main content area shows the 'Sunsports' dataset page, which includes a header with the dataset name and description, tabs for Data, Tasks, Code (46), Discussion, Activity, and Metadata, and buttons for Download (70 KB) and New Notebook. The Activity tab is selected, showing activity stats (26.2k views, 4346 downloads, 0.17 download per view ratio, 40 total unique contributors) and a line chart of sunspot activity from June 2020 to April 2021. Below the chart are sections for Top contributors (745H1N, Mohsin Raza, Srinath Srinivasan) and Top contributor countries (a world map). Below the dataset page, a search for notebooks returns five results related to sunspot time series analysis using ARIMA, LSTM, and Keras.

3. Соревнования по науке о данных / машинному обучению / глубокому обучению

Хотя я не участвовал ни в одном из них, мне нравится, как мы решаем задачу в реальном времени вместе с сообществом Kaggle и выигрываем потрясающие денежные призы (если участвуем в этом конкретном соревновании). Я определенно хочу участвовать когда-нибудь в ближайшее время, надеюсь, изображения мотивируют вас. Необязательно, чтобы это могли делать только крупные компании или богатые предприятия. Вы тоже можете это сделать. Есть определенные протоколы, которым нужно следовать, и вуаля, у вас есть собственное соревнование, организованное вами.

- Home
- Competitions
- Datasets
- Code
- Discussions
- Courses
- More

Recently Viewed

- Sunspots
- Intro to Machine Learn...
- Exercise: Syntax, Varia...
- Python
- Pumpkin Price Polyno...

View Active Events

- Home
- Competitions
- Datasets
- Code
- Discussions
- Courses
- More

Recently Viewed

- Sunspots
- Intro to Machine Learn...
- Exercise: Syntax, Varia...
- Python
- Pumpkin Price Polyno...

View Active Events

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [documentation](#) or learn about [InClass competitions](#).

[+ Host a Competition](#)
[Your Work](#)

Your Competitions

Active
Closed
Pinned
Hosted

Live competitions you've joined go here - why don't you enter one to [get started!](#)

All Competitions

Active (Not Entered)
Completed
InClass

All Categories
Default Sort

	SIIM-FISABIO-RSNA COVID-19 Detection Identify and localize COVID-19 abnormalities on chest radiographs Featured • 2 months to go • Code Competition • 240 Teams	\$100,000
	Coleridge Initiative - Show US the Data Discover how data is used for the public good Featured • 24 days to go • Code Competition • 1256 Teams	\$90,000
	CommonLit Readability Prize Rate the complexity of literary passages for grades 3-12 classroom use Featured • 2 months to go • Code Competition • 1560 Teams	\$60,000
	Bristol-Myers Squibb - Molecular Translation Can you translate chemical images to text?	\$50,000

All Competitions

Active (Not Entered)
Completed
InClass

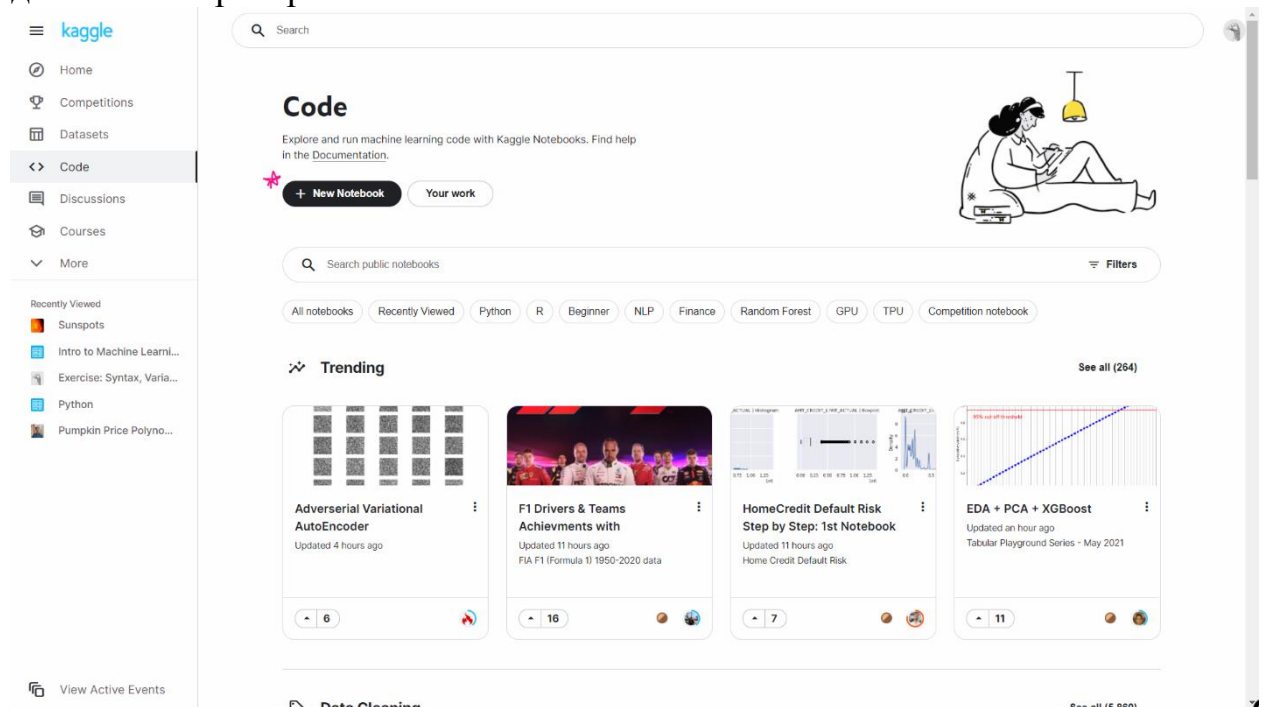
All Categories
Reward

	Passenger Screening Algorithm Challenge Improve the accuracy of the Department of Homeland Security's threat recognition algorithms Featured • 3 years ago • 518 Teams	\$1,500,000
	Zillow Prize: Zillow's Home Value Prediction (Zestimate) Can you improve the algorithm that changed the world of real estate? Featured • 3 years ago • 3770 Teams	\$1,200,000
	Data Science Bowl 2017 Can you improve lung cancer detection? Featured • 4 years ago • 1972 Teams	\$1,000,000
	Deepfake Detection Challenge Identify videos with facial or voice manipulations Featured • a year ago • Code Competition • 2265 Teams	\$1,000,000
	Heritage Health Prize Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012) Featured • 8 years ago • 1350 Teams	\$500,000
	GE Flight Quest Think you can change the future of flight? GE Quests • 8 years ago • 172 Teams	\$250,000
	Flight Quest 2: Flight Optimization, Milestone Phase Optimize flight routes based on current weather and traffic. GE Quests • 8 years ago • 129 Teams	\$250,000
	Flight Quest 2: Flight Optimization, Main Phase Optimize flight routes based on current weather and traffic. GE Quests • 7 years ago • 121 Teams	\$220,000
	Flight Quest 2: Flight Optimization, Final Phase Final Phase of Flight Quest 2	\$220,000

4. Блокноты Kaggle (Код)

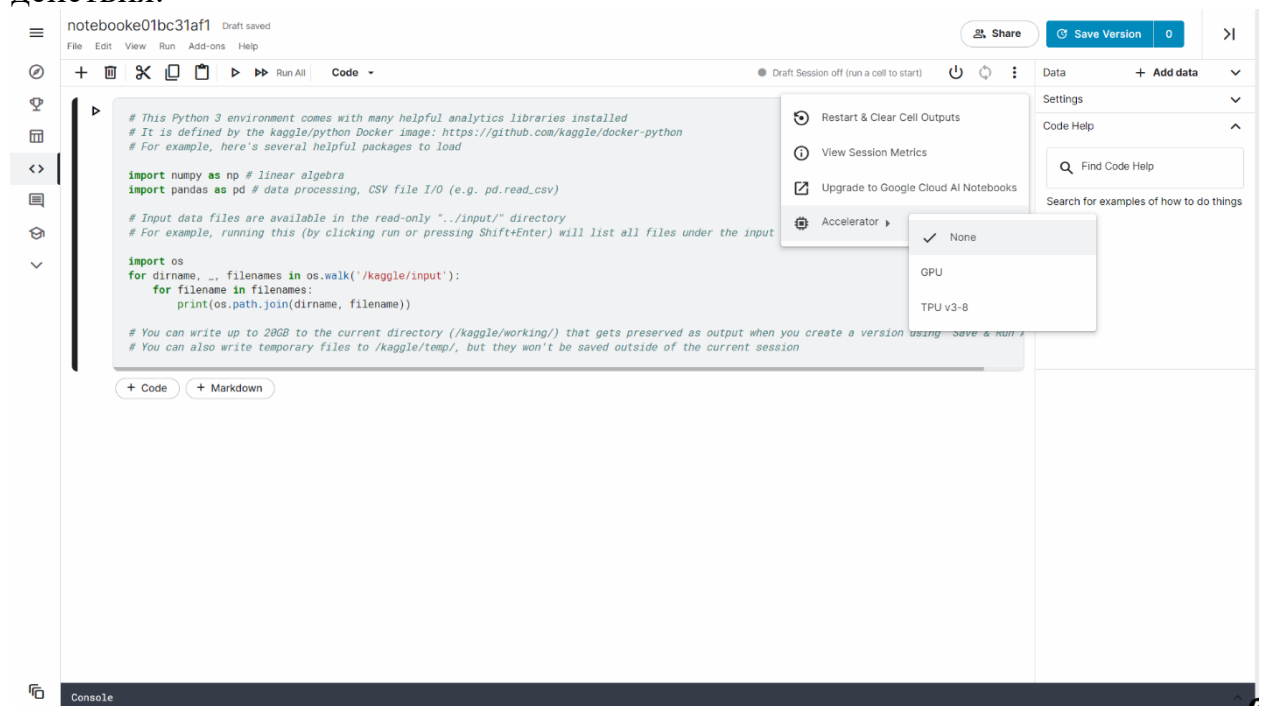
Для любой задачи, связанной с наукой о данных или информатикой, мы должны написать хоть какой-то код. Kaggle предоставляет нам свою собственную среду Notebook с определенным ограничением того, сколько мы можем хранить на них (в совокупности для каждой учетной записи), сколько часов доступного графического процессора и сколько часов доступно TPU. Они полностью интегрированы со всеми сервисами Kaggle и могут использоваться независимо, как любая другая среда для ноутбуков (Datalore, Google Colab, Jupyter и т.д.), что означает, что вы можете использовать их для своей практики, соревнований Kaggle, курсов Kaggle, анализа некоторых

Kaggle / или набор данных, не относящийся к Kaggle, и многое другое. Вы должны их проверить.

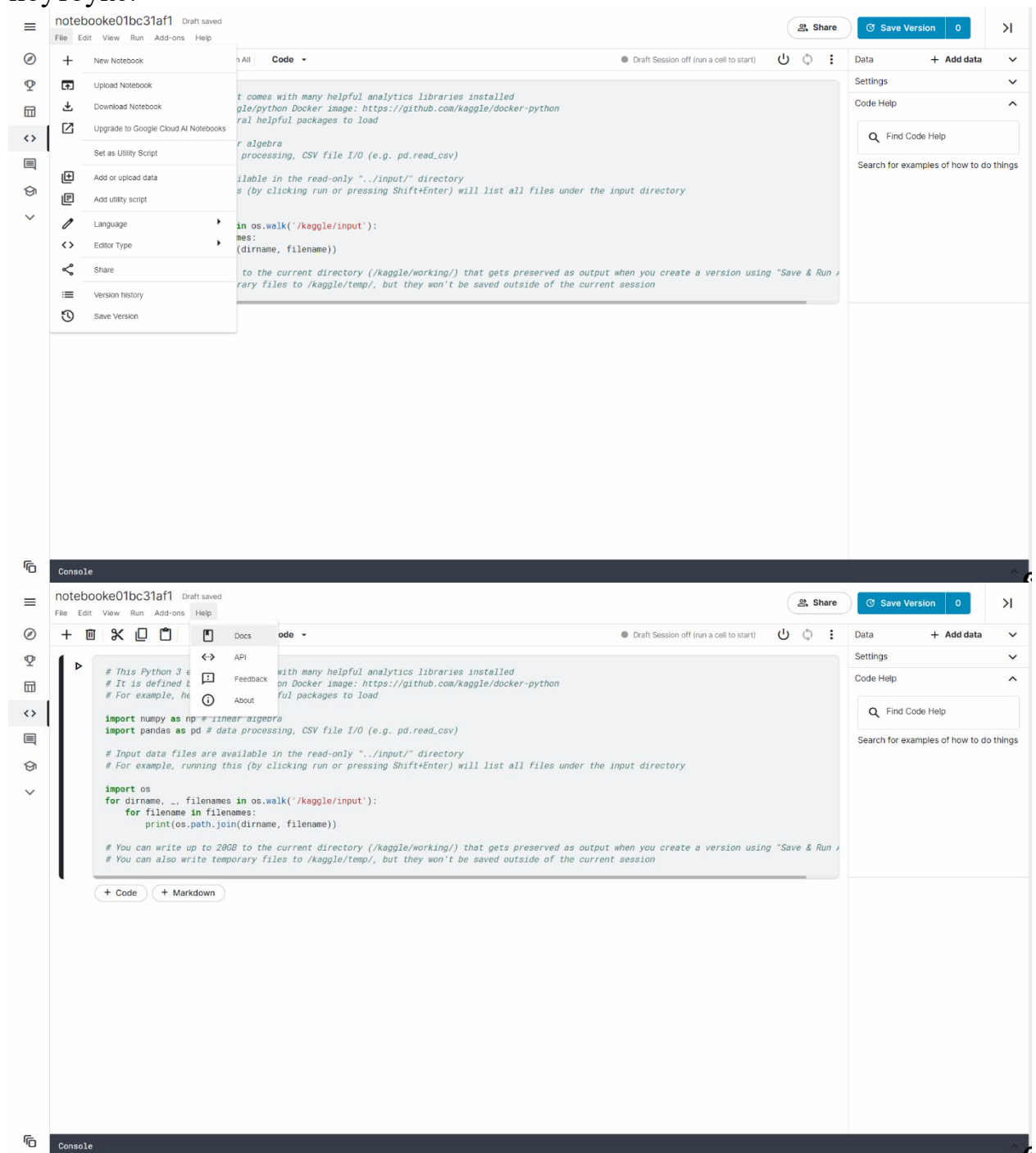


Нажав на эту черную кнопку, вы создаете свою записную книжку или открываете чужую записную книжку, которую хотите прочитать и изучить / сравнить. Все эти видимые записные книжки являются общедоступными, что означает, что ваши записные книжки не будут видны никому, если вы сами этого не сделаете.

Чтобы переключиться с CPU на GPU или TPU, выполните следующие действия:



Это большинство функциональных возможностей, доступных вам в этом ноутбуке:



Давайте посмотрим, как использовать их с данными (импортированными / взятыми непосредственно из Kaggle / загруженными с URL-адреса и т.д.) и начнем работать над вашими задачами в области науки о данных.

The image shows a Kaggle notebook interface. The top part displays a code cell with Python code for loading data. The bottom part shows the 'Add Data' modal, which is used to search for and upload datasets. The modal includes a search bar, a list of datasets, and an 'Add' button for each dataset.

Code Cell:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

Add Data Modal:

The modal shows a search bar and a list of datasets. The datasets listed are:

- Reddit Vaccine Myths
- A Large Scale Fish Dataset
- MusicNet Dataset
- Wikibooks Dataset
- Careerbuilder Job Listing 2020
- Twitter Edge Nodes
- Famous Iconic Women

Each dataset entry includes a thumbnail, a brief description, and an 'Add' button. The modal also includes a 'Search by URL' option and a 'Search Datasets' button.

Здесь покажем, как использовать набор данных «Солнечные пятна», который мы видели ранее. Начните с поиска.

The image consists of two screenshots of the Kaggle Notebook interface. The top screenshot shows the 'Add Data' modal with a search for 'sunspot'. The modal displays a list of datasets, including 'Sunsports', 'Daily Sun Spot Data (1818 to 2019)', 'SunSpot', 'Kanzel Sunspot Data: May 1944 - Aug. 2020', 'Average Sun Spot Number', 'SIDC/SILSO Daily Sunspot ISN', and 'Monthly Sunspots'. The bottom screenshot shows the notebook code with a red circle around the 'sunsports' dataset in the 'Data' panel on the right, with the word 'load' written next to it.

Теперь данные успешно загружены. Выделение на изображении выше - это каталог, в котором он хранится. Давайте посмотрим на небольшой код *pandas*, чтобы узнать, как импортировать набор данных.

The screenshot shows a Kaggle notebook titled 'notebook01bc31af1'. The code cell [1]: contains Python code for loading and processing data. The output cell [7]: shows the result of the code execution, which is a pandas DataFrame. The DataFrame has columns 'Unnamed: 0', 'Date', and 'Monthly Mean Total Sunspot Number'. The data is as follows:

Unnamed: 0	Date	Monthly Mean Total Sunspot Number
0	1749-01-31	96.7
1	1749-02-28	104.3
2	1749-03-31	116.7
3	1749-04-30	92.8
4	1749-05-31	141.7

Handwritten annotations in pink include 'Path' pointing to the file path in the code and 'data' pointing to the output table. The right sidebar shows the 'Data' section with 'input (69.79 KB)' and 'output (44.1MB / 19.6GB)' listed.

Последнее, что вы можете сделать после завершения своего проекта / работы, - это поделиться им с сообществом на Kaggle. Это важный шаг, потому что, делаясь своими идеями, своей работой, мы расширяем возможности, доступные сообществу, и поддерживаем друг друга.

Слева от Большой синей кнопки в правом верхнем углу вы увидите кнопку «Поделиться». Щелкните по нему и выберите «Публичный» в раскрывающемся меню.

The screenshot shows the same Kaggle notebook interface, but with the 'Share with others' dialog box open. The dialog box has a title 'Share with others' and three options: 'Private' (Only collaborators can view this notebook), 'Public' (Anyone at Kaggle can find and view under the Apache 2.0 License), and 'Private' (Only collaborators can view this notebook). The 'Public' option is selected. There is also an 'Add collaborators' button and 'Cancel' and 'Save' buttons at the bottom.