

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ МОСКОВСКОЙ ОБЛАСТИ**

**Государственное бюджетное профессиональное образовательное**

**учреждение Московской области**

**«Люберецкий техникум имени Героя Советского Союза,**

**лётчика-космонавта Ю.А. Гагарина»**

**КУРСОВАЯ РАБОТА**

**по МДК.01.01 «Разработка программных модулей »**

**по МДК.01.03 «Разработка мобильных приложений »**

специальность

09.02.07 «Информационные системы и программирование»

**Тема:** Data Science

Группа 185

Выполнил/а/студен\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_            Городенцев Сергей Витальевич

                                    (подпись)                         (ФИО полностью)

Руководитель                                                         Жирнова Юлия Витальевна

                                             (подпись)                           (ФИО полностью)

Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Люберцы-2021 год

**Содержание**

[Введение 3](#_heading=h.gjdgxs)

[1.](#_heading=h.30j0zll) Теоретическая часть 4

[1.1.](#_heading=h.1fob9te) Актуальность 4

[1.2.](#_heading=h.3znysh7) Теория 6

[2.](#_heading=h.tyjcwt) Практическая часть 11

2.1. Введение в Phyton13

[2.1.](#_heading=h.3dy6vkm) Рассмотреть принцип работы 11

[2.2.](#_heading=h.1t3h5sf) Пример Data Science 13

[Список литературы 19](#_heading=h.44sinio)

**Введение**

Современный мир теперь нельзя представить без информационных систем. Они встречаются постоянно в нашей жизни. Информационные системы очень помогают в жизни и значительно упрощают ее.

Каждая организация стремиться минимизировать затраты времени, материальных, трудовых ресурсов в ходе своей деятельности и упростить процесс обработки информации. Эти задачи можно решить с использованием информационных систем.

Data Science – наука, ищущая способы получить представления об объектах через данные.

Цель данной курсовой работы: понять, что такое и как оно востребована эта профессия в наши дни Data science.

Для выполнения поставленной цели необходимо решить данные задачи:

1.Провети анализ основных инструментов разработки Data science.

2.Рассмотреть принцип работы.

3.Разработать Data science.

**Теоретическая часть**

**1.1Актуальность**

Наука о данных — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных.

В 2010-х годах объемы данных стали расти по экспоненте. Свою роль сыграл целый ряд факторов — от повсеместного распространения мобильного интернета и популярности соцсетей до всеобщей оцифровки сервисов и процессов. В итоге профессия дата-сайентиста быстро превратилась в одну из самых популярных и востребованных. Еще в 2012 году позицию дата-сайентиста журналисты назвали самой привлекательной работой XXI века (The Sexiest Job of the XXI Century). Развитие Data Science шло вместе с внедрением технологий Big Data и анализа данных. И хотя эти области часто пересекаются, их не следует путать между собой. Все они предполагают понимание больших массивов информации. Но если аналитика данных отвечает на вопросы о прошлом (например, об изменениях в поведениях клиентов какого-либо интернет-сервиса за последние несколько лет), то Data Science в буквальном смысле смотрит в будущее. Специалисты по DS на основе больших данных могут создавать модели, которые предсказывают, что случится завтра. В том числе и предсказывать спрос на те или иные товары и услуги.  
  
Вот лишь несколько примеров отраслей, которые используют Data Science для решения своих задач:  
  
**онлайн-торговля и развлекательные сервисы**: рекомендательные системы для пользователей;  
**здравоохранение**: прогнозирование заболеваний и рекомендации по сохранению здоровья;  
**логистика**: планирование и оптимизация маршрутов доставки;  
**digital-реклама**: автоматизированное размещение контента и таргетирование;  
**финансы**: скоринг, обнаружение и предотвращение мошенничества;  
**промышленность**: предиктивная аналитика для планирования ремонтов и производства;  
**недвижимость**: поиск и предложение наиболее подходящих покупателю объектов;  
**госуправление**: прогнозирование занятости и экономической ситуации, борьба с преступностью;  
**спорт**: отбор перспективных игроков и разработка стратегий игры.  
И это лишь самый краткий и поверхностный список использования Data Science. Количество различных кейсов с использованием «науки о данных» увеличивается с каждым годом в геометрической прогрессии.  
Каждый интернет-пользователь и просто потребитель ежедневно десятки раз сталкивается с продуктами и решениями, в которых применяются инструменты Data Science. К примеру, аудио-сервис Spotify использует их, чтобы лучше подбирать треки для пользователей в соответствии с их предпочтениями. То же самое можно сказать о предложении фильмов и сериалах на видео-стримингах, таких как Netflix. А в Uber науку о данных рассматривают как инструмент для предиктивной аналитики, прогнозирования спроса, улучшения и автоматизации всех продуктов и клиентского опыта.

**1.2 Теория**

Data science использует разные инструменты для создания своих данных. Часто используются такие инструменты как: Python, R, SAS, Java, C/C++.

Data Science и принятие решений на основе данных — это итерационный, а не линейный процесс. Тем не менее, стандартный цикл анализа данных обычно включает в себя следующие этапы:

**Планирование** - Определение задач и потенциальных результатов проекта;

**Построение модели данных**  Для построения моделей машинного обучения исследователи данных зачастую используют различные библиотеки с открытым исходным кодом или средства, работающие в базах данных. Пользователям часто требуются API, чтобы упростить получение данных, их профилирование, визуализацию или разработку функций. Для этого им требуются правильные инструменты, а также доступ к правильным данным и другие ресурсы, такие как вычислительные мощности;

**Оценка модели:**  Исследователям данных необходимо добиться высокого процента точности для своих моделей. Только после этого их можно будет с уверенностью использовать. При оценке моделей обычно формируется сложный набор метрик и визуализаций, позволяющих измерить точность работы моделей с актуальными данными, а также ранжировать их по времени, чтобы добиться оптимального поведения в продуктивной среде. При оценке моделей учитывается не только их производительность, но и ожидаемое базовое поведение;

**Объяснение моделей:**  Не всегда возможно объяснить внутреннюю механику результатов работы моделей машинного обучения понятным человеку языком, но способность делать это приобретает все большее значение. Исследователям данных требуются автоматически формируемые пояснения того, как определяется относительный вес и факторы важности, которые используются при формировании прогноза, а также подробное объяснение прогнозов, выдаваемых конкретными моделями;

**Развертывание модели:**  Применение обученной модели машинного обучения в нужных системах зачастую является сложным и трудоемким процессом. Его можно упростить путем реализации моделей в виде масштабируемых и безопасных API или использования моделей машинного обучения, работающих в базах данных;

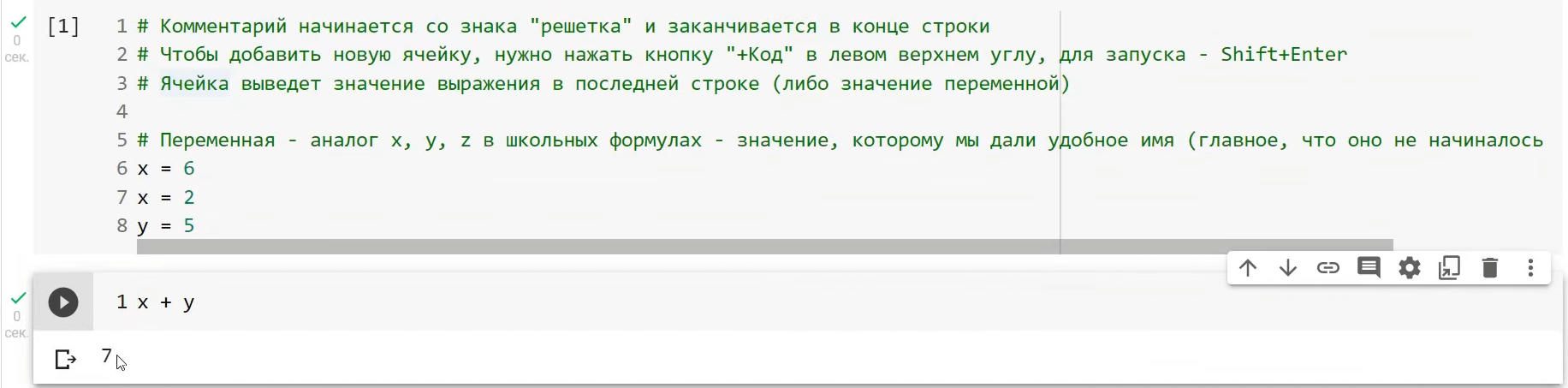
**Мониторинг моделей:**  К сожалению, все не заканчивается развертыванием модели. Чтобы обеспечить надлежащую работу моделей, после развертывания необходимо вести постоянный мониторинг. Спустя какое-то время данные, на которых производилось обучение моделей, могут перестать быть релевантными для будущих прогнозов. Например, киберпреступники постоянно реализуют новые способы взлома учетных записей.

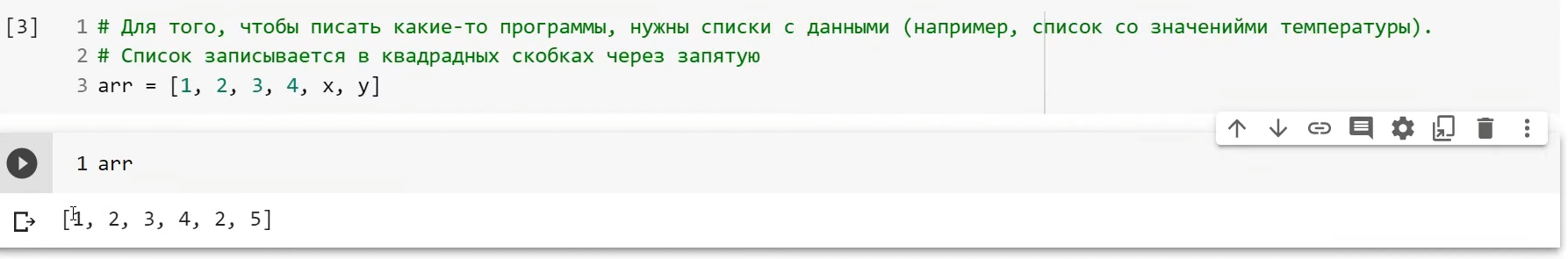
**2 Практическая часть**

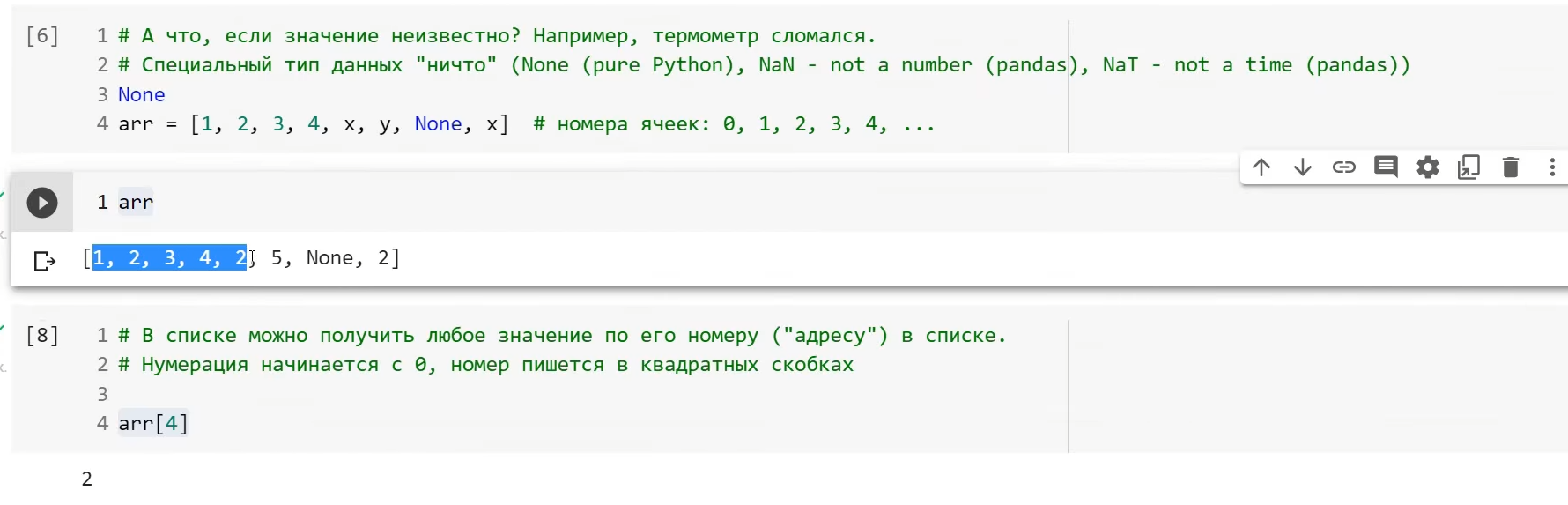
**2.1 Введение в Phyton.**

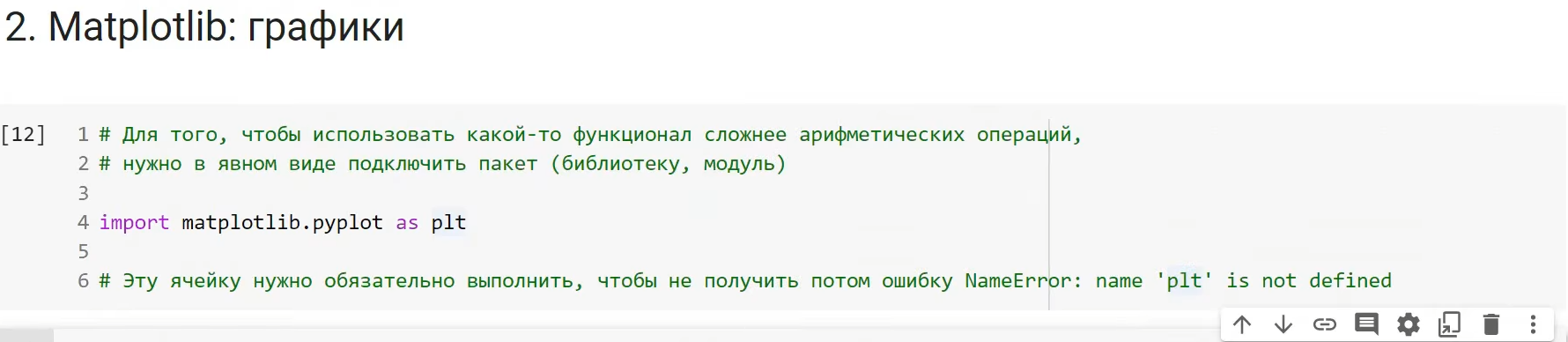
Рассмотрим изучение Data Science на языке Phyton, на платформе googl.colab.



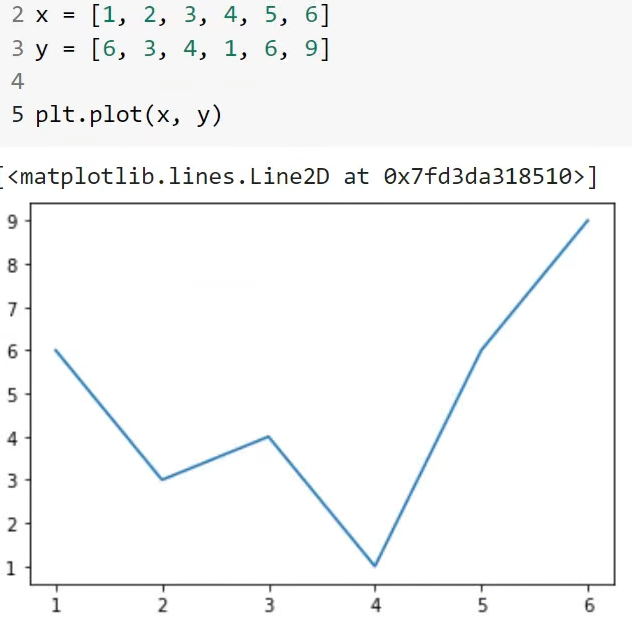


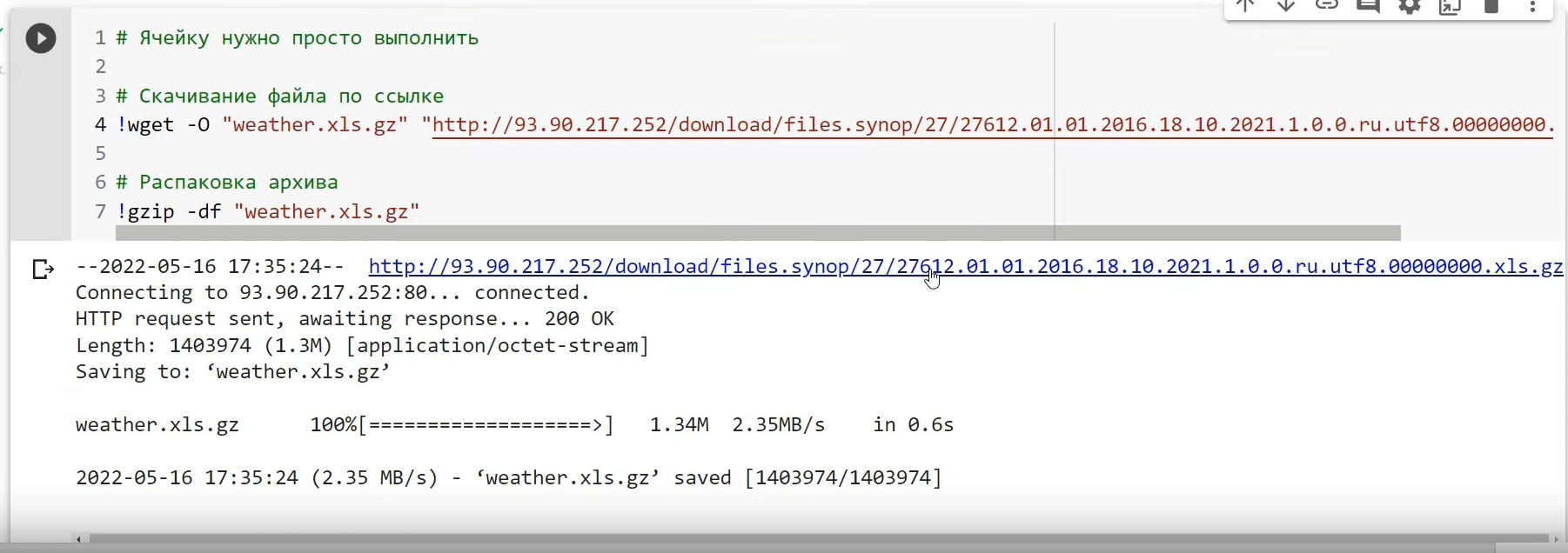




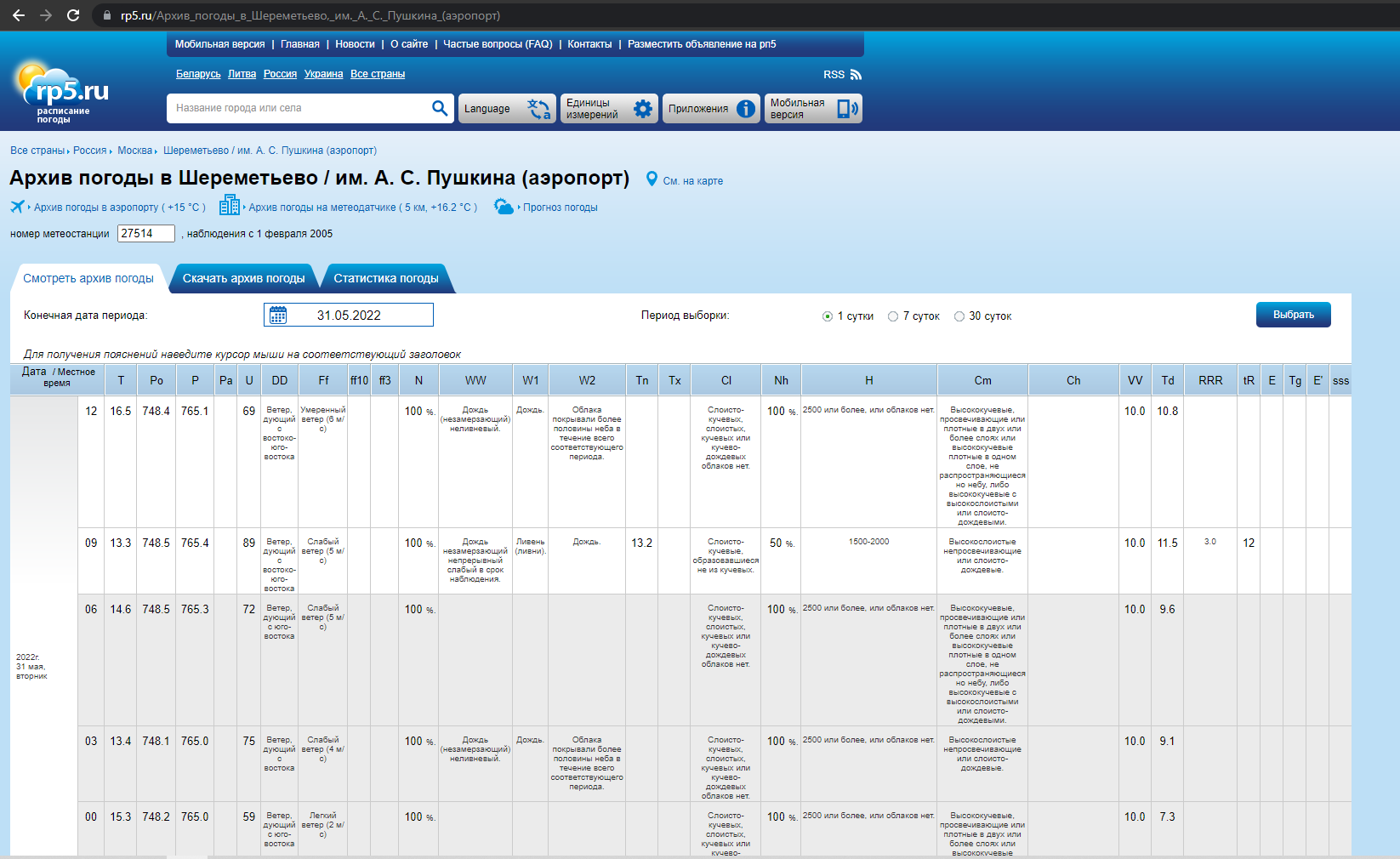
**2.1 Принцип работы**  


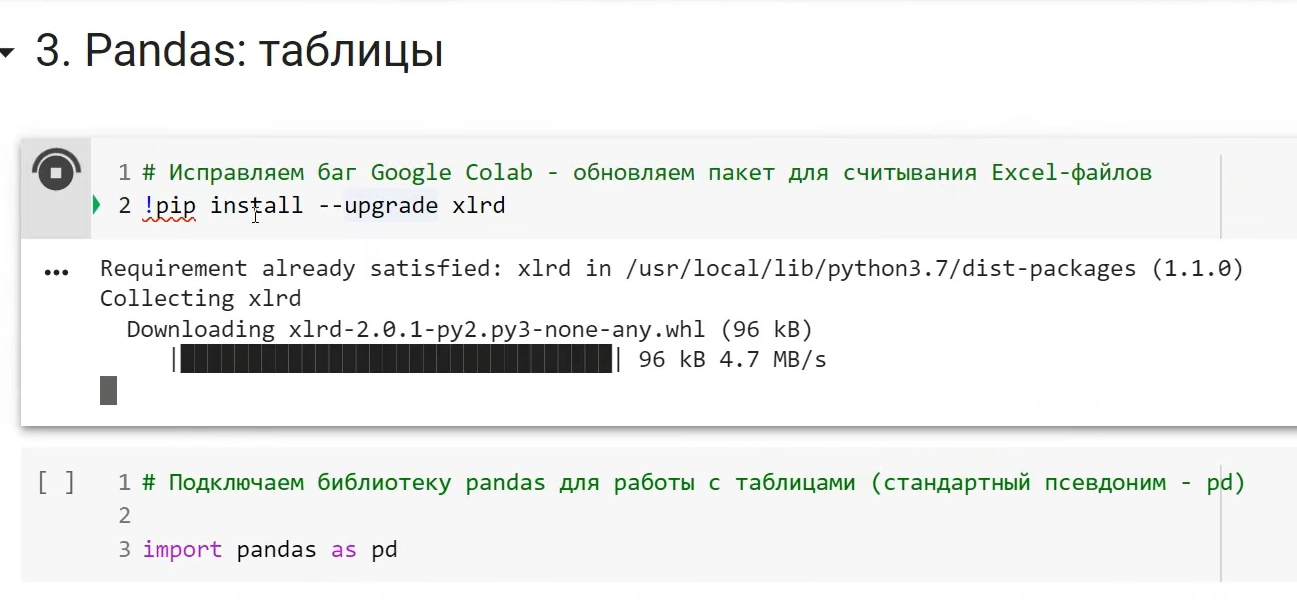
**Matplotlib** — [библиотека](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B0_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)) на языке программирования [Python](https://ru.wikipedia.org/wiki/Python) для визуализации данных [двумерной](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0#%D0%94%D0%B2%D1%83%D0%BC%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0_(2D)) и [трёхмерной](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0#%D0%A2%D1%80%D1%91%D1%85%D0%BC%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0) графикой. Получаемые изображения могут быть использованы в качестве иллюстраций в публикация.



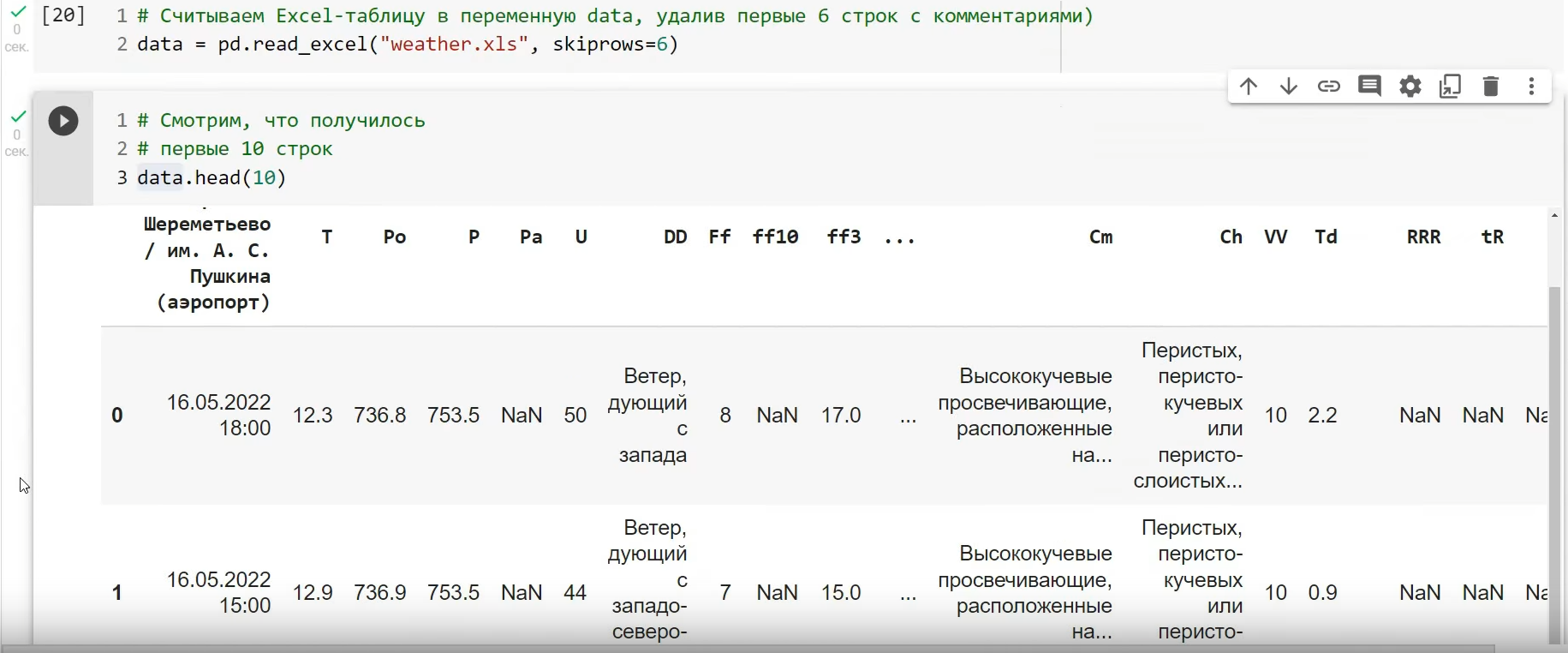


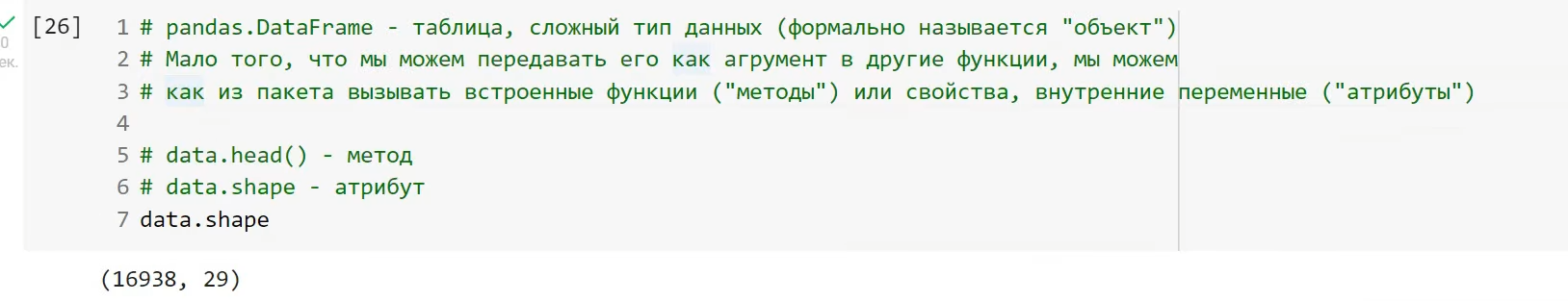
По ссылке мы перешли на сайт, с которого будем брать данные и работать с ним дальше.



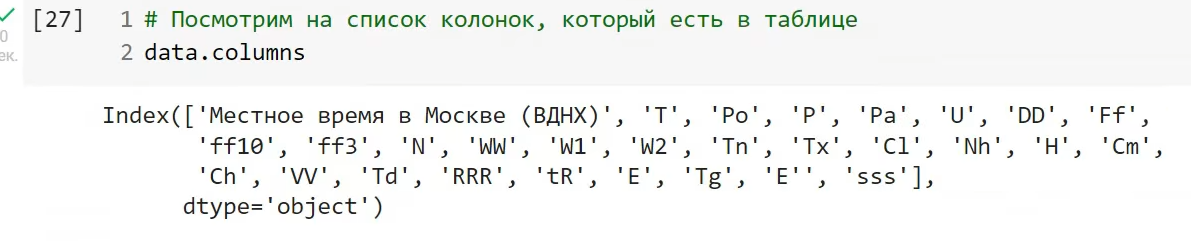


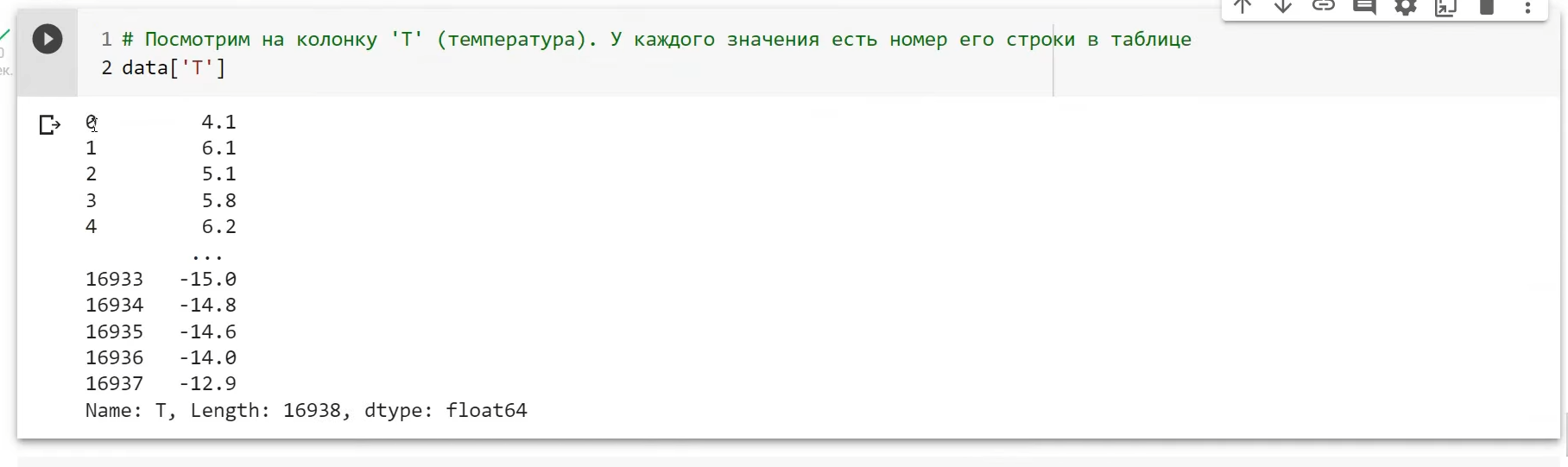
Pandas программа с помощью, которой мы считываем данные для графика.

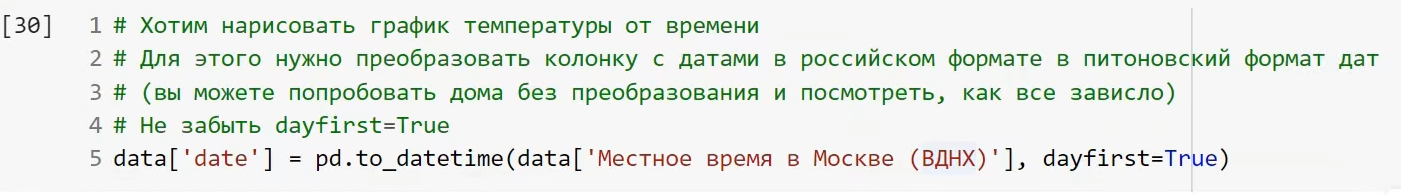


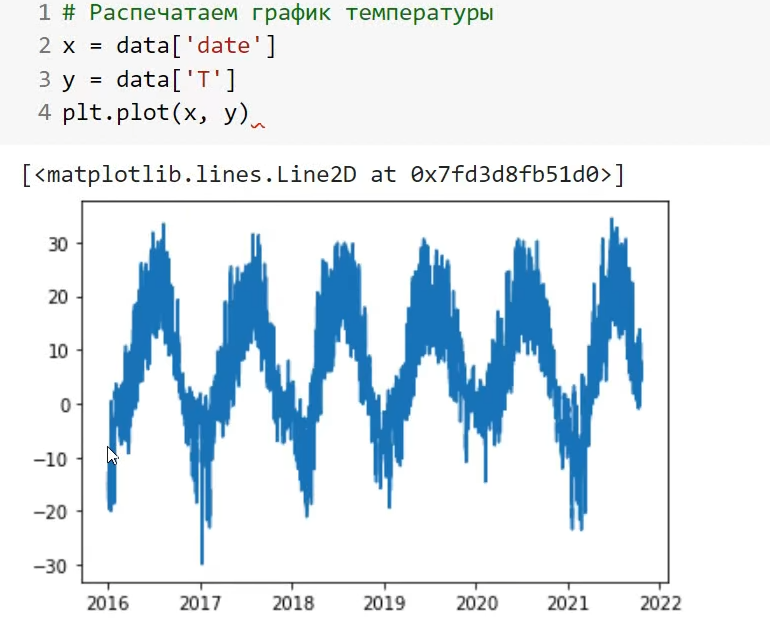


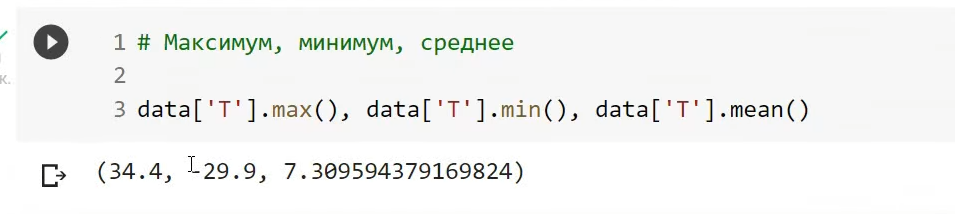
Размерность данных которые нам выдала программа за промежуток в 5 лет.

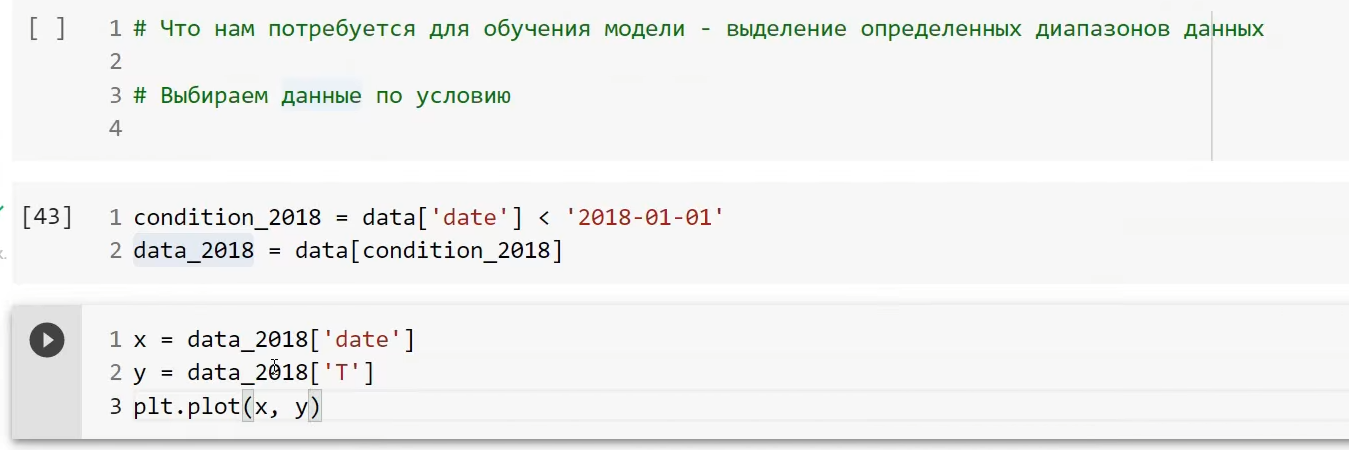


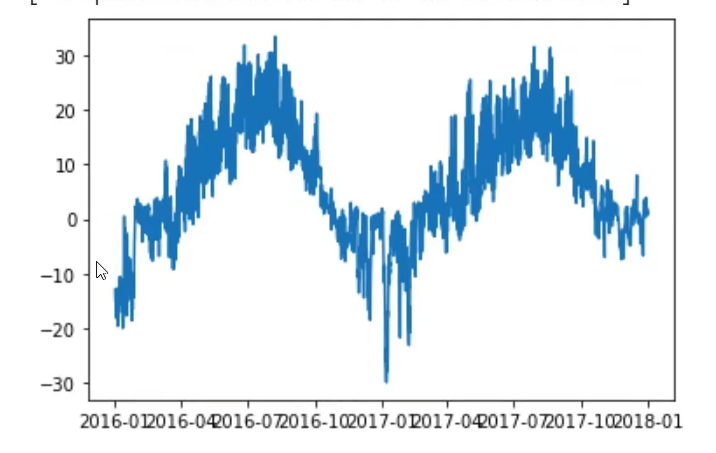












Тут мы обозначили диапазон в два года.

**Data Science**

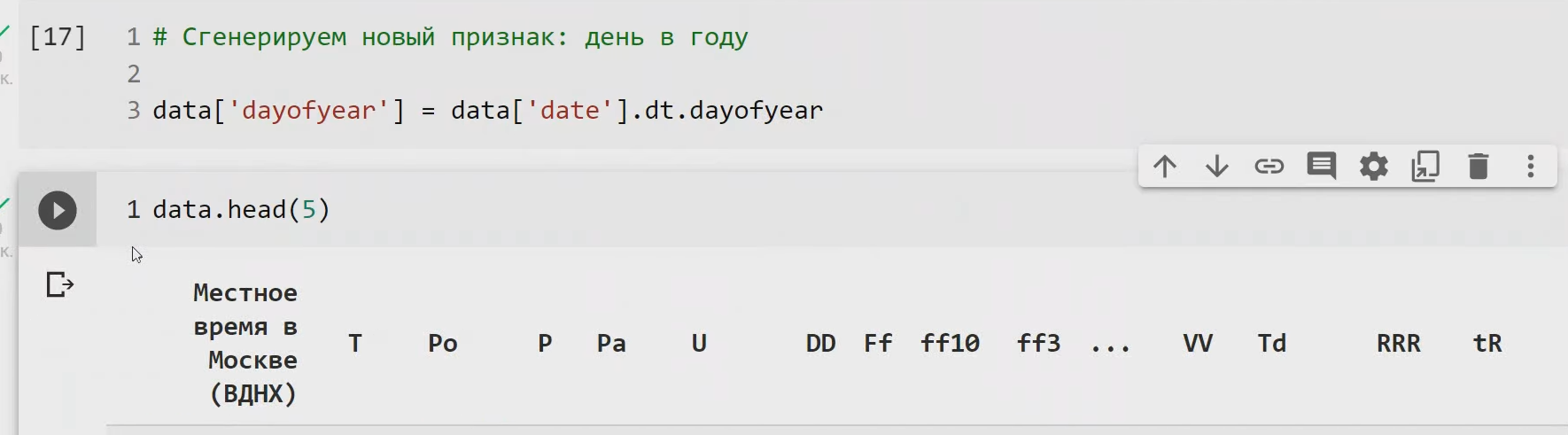
Работа **data scientist** – подобрать наиболее близкую функцию с помощью инструментов машинного обучения y=f(x)

Y – **target(**таргет, целевая переменная) – величина которую мы хотим прогнозировать. В данный момент это будет колонка с температурой [‘T’]

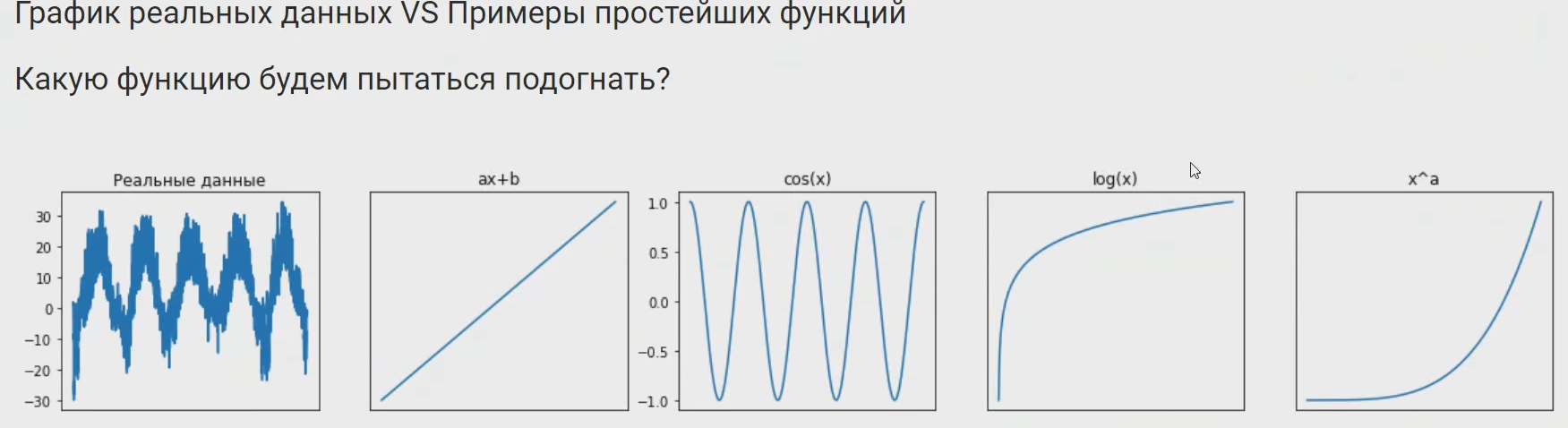
X -**features**(признаки) – факторы, по которым мы хотим прогнозировать что-то. В нашем случае день в году. Можно выбрать другие признаки. Это всё входит в работу **data scientist**.

Выберем модель – **Линейная регрессия**.

Чтобы обучить модель нужно пройти две предварительные стадии: 1)Сгенерировать новые признаки(если нужно); 2) Разделить данные на две выборки (2 набора записей) – для обучения модели и тестирования.

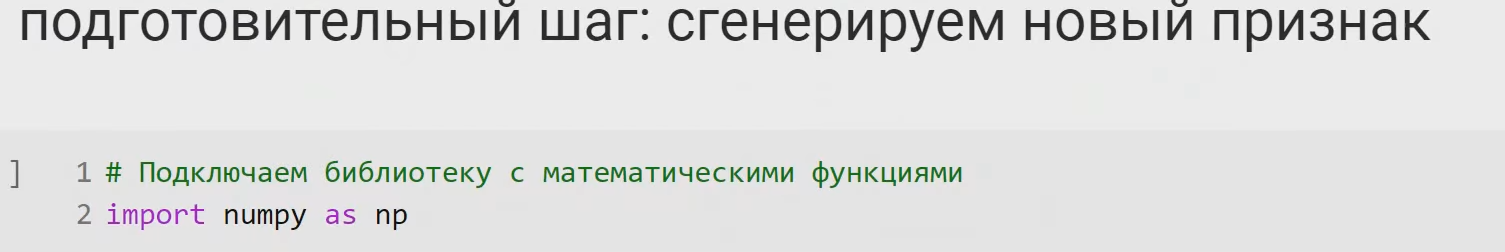


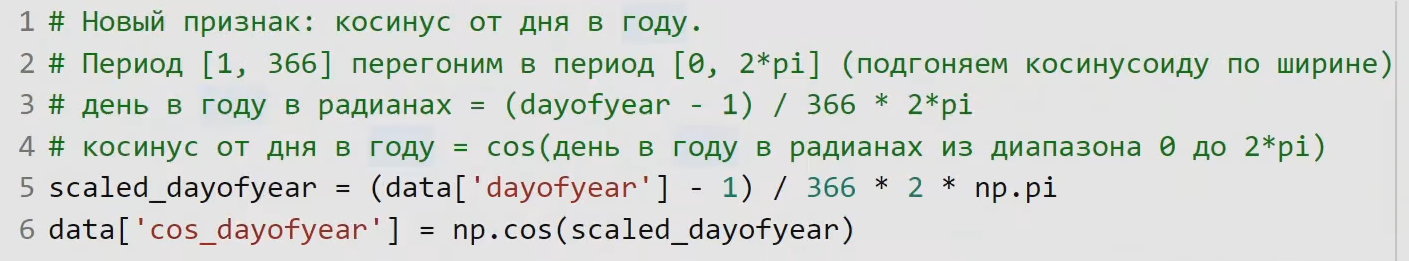
Пытаемся подобрать функцию z(x), которая просто вычисляется и максимально похожа на f(x). Самая простая модель – линейная регрессия, которая подбирает коэффициенты *z=ax+b.* Коэффициенты a и b подбираются автоматически



Данные похожи на косинус. Но его нужно подогнать к нашим данным.

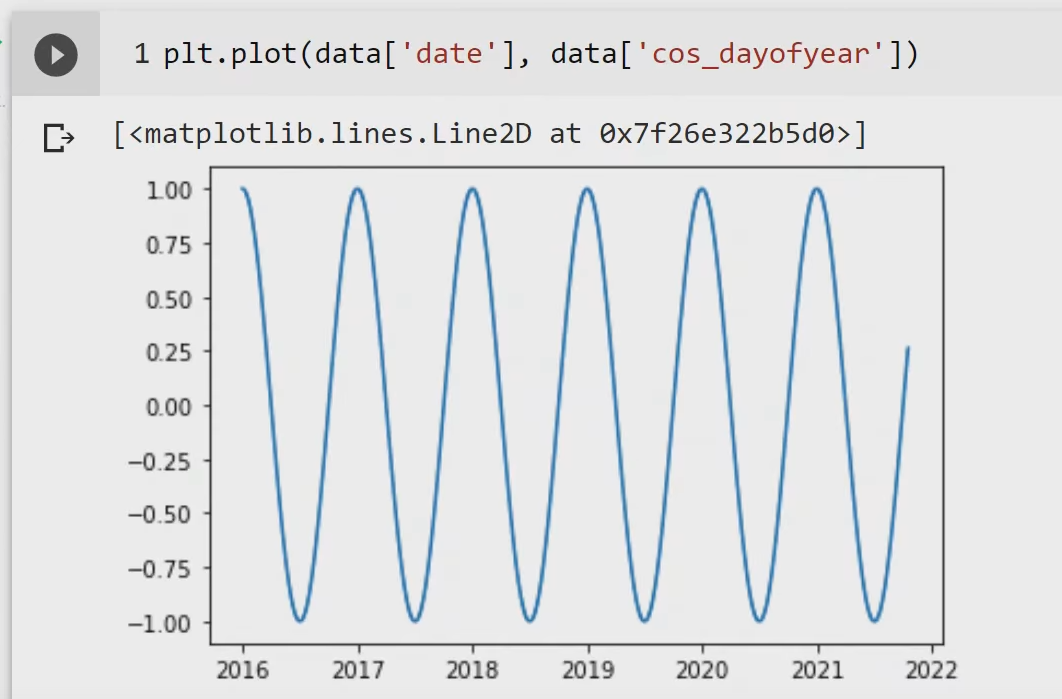
По периоду(растянуть) и по амплитуде (растянуть в высоту).

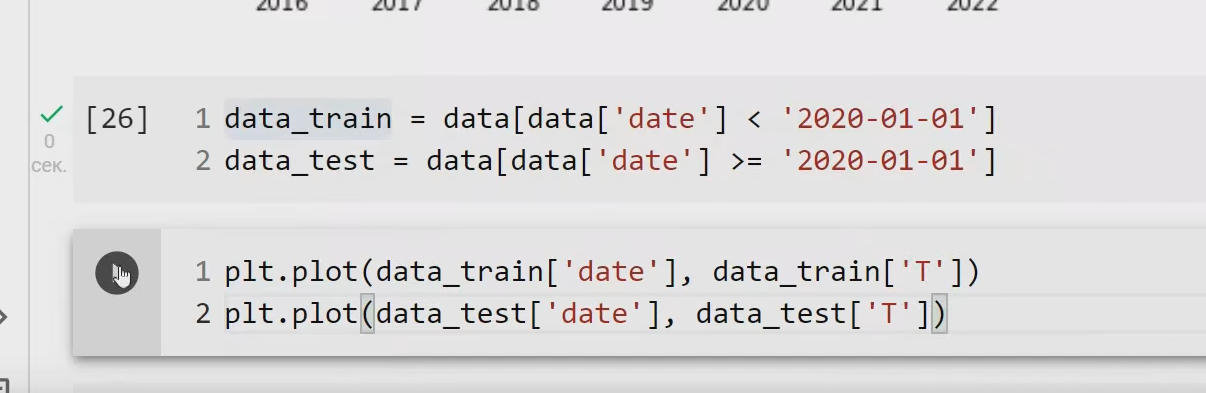


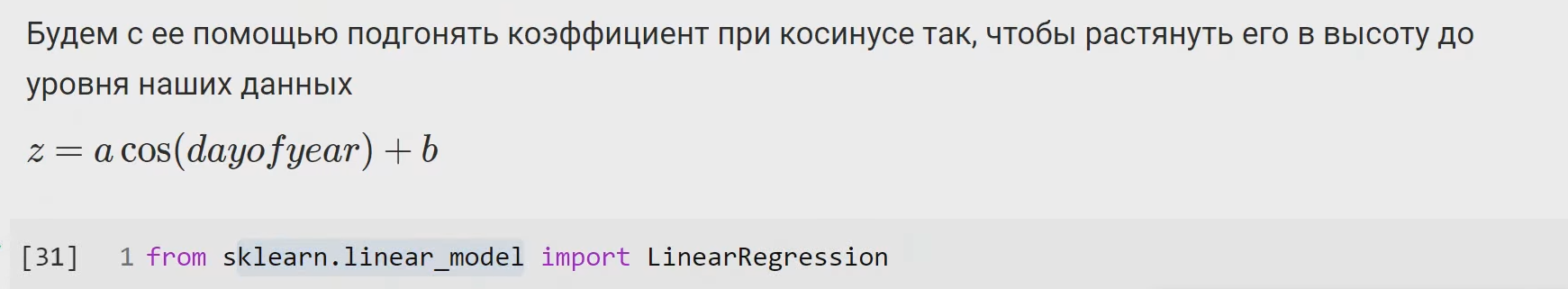


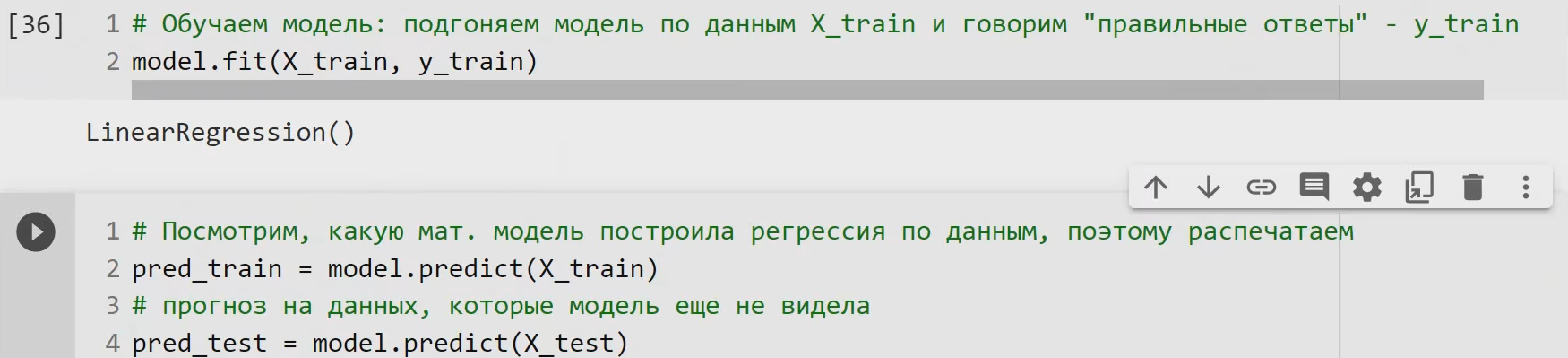
Теперь нам следует разделить данные на Test и Train.

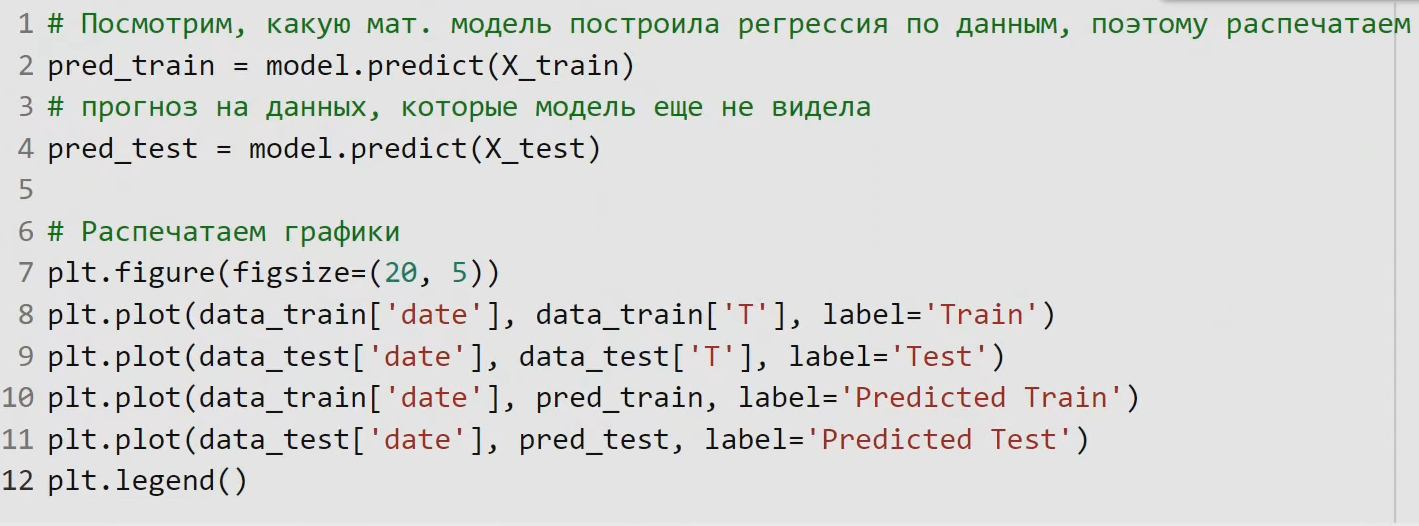
На обучающей выборке модель погоняет свои внутренние параметры с помощью статистики(train). На тестовой, тестируем качество модели.

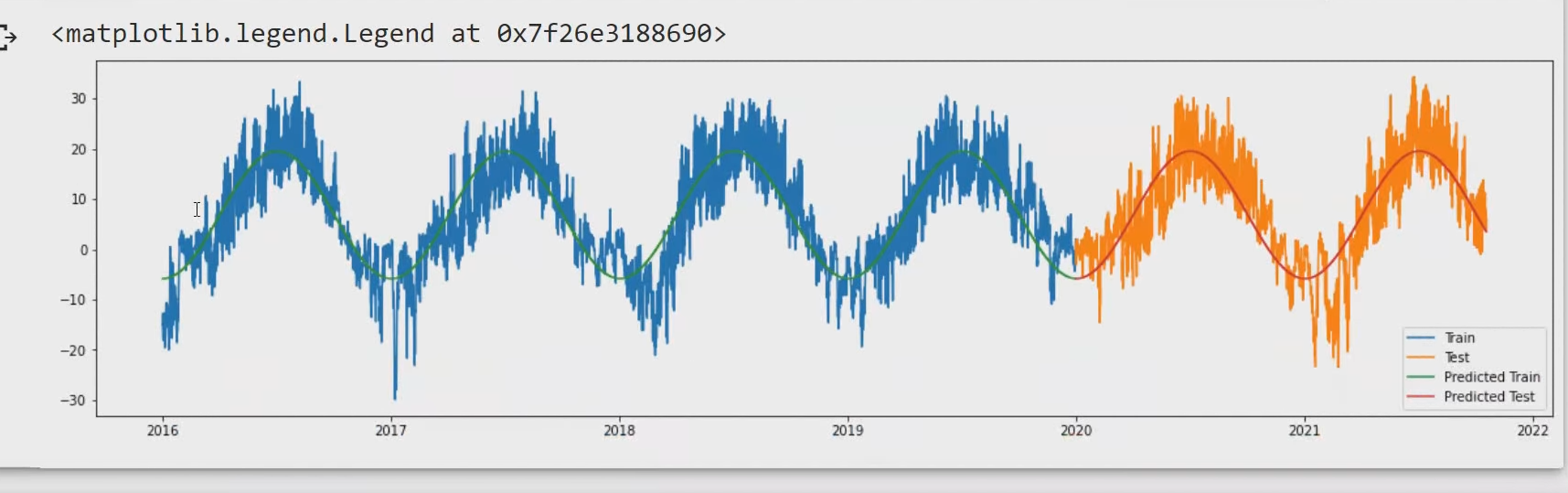












Вот что по итогу у нас получилось.

Список литературы

<https://rp5.ru>

<https://matplotlib.org>

https://colab.research.google.com