1 Azure ML Studio

1.1 Регистрация

Регистрация на платформе Azure ML Studio (https://studio.azureml.net/) осуществляется во кладке `Sign up`. При регистрации следует выбрать «Free Workspace» и пройти авторизацию с помощью аккаунта Microsoft (это может быть учетная запись от Office 365, полученная через ИСУ, или любая другая) либо создать новую учетную запись.

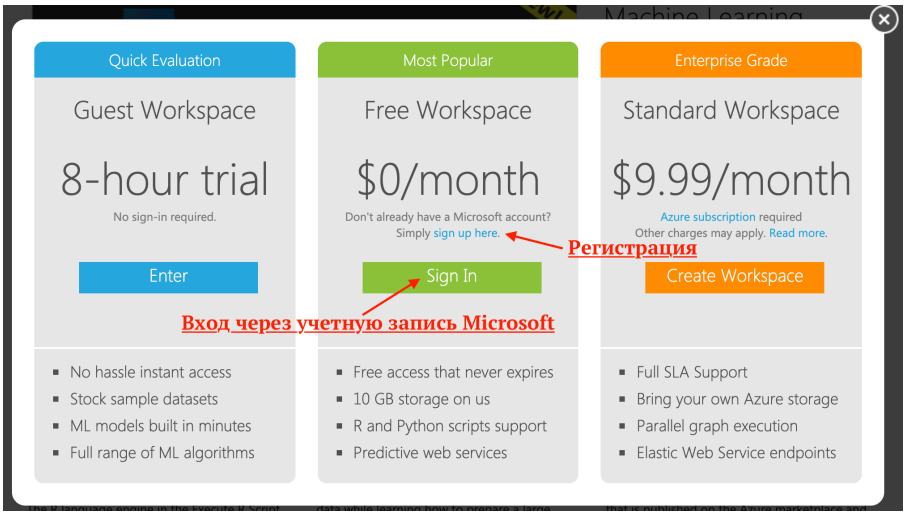


Рис. 1: Первый шаг регистрации.

После авторизации, Вы попадаете на главную страницу.

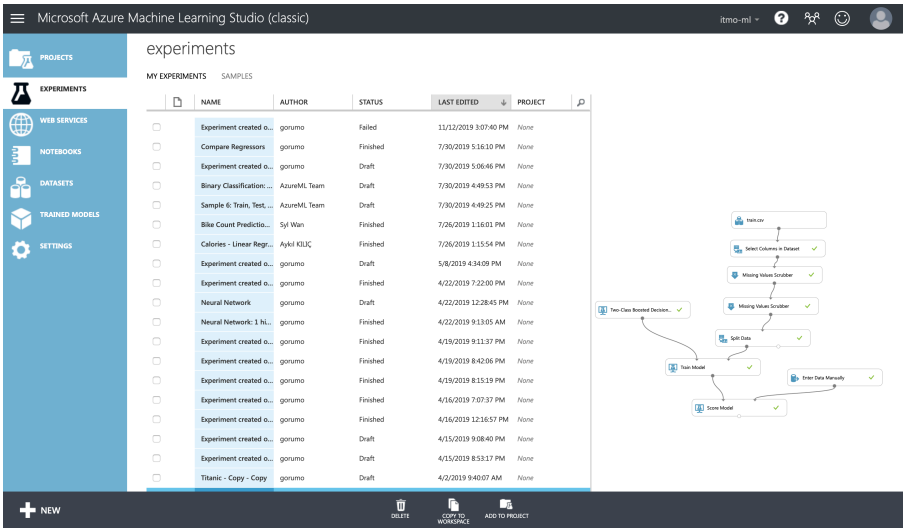


Рис. 2: Главная страница Azure ML Studio.

1.2 Datasets

Раздел меню `Datasets` позволяет просмотреть загруженные Вами данные. Открытые и загруженные датасеты можно найти во вкладке `Samples`. При добавлении нового датасета рекомендуется использовать формат дан-

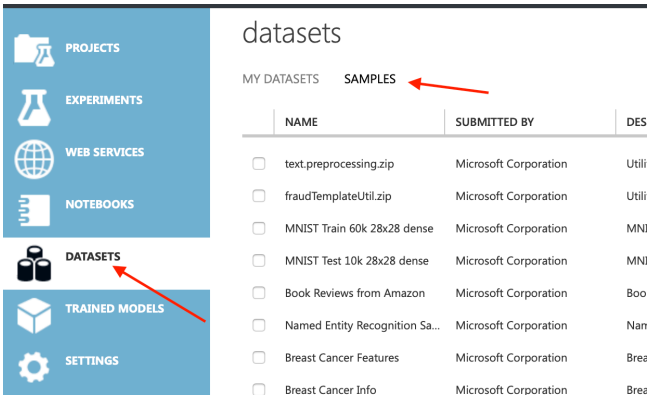
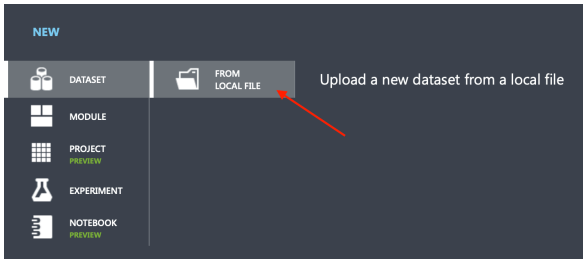
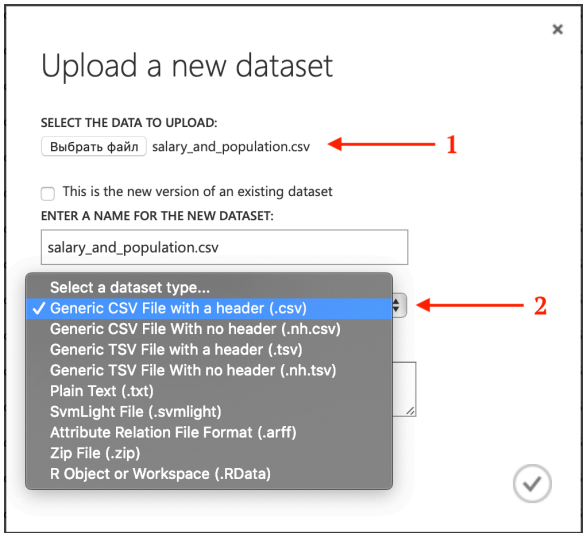


Рис. 3: Наборы данных.

ных CSV, разделителем колонок **обязательно** должна быть запятая. Для загрузки данных необходимо нажать кнопку `+ NEW` в левом нижнем углу, после чего нажать «`FROM LOCAL FILE`» и выбрать CSV файл данных в пункте `SELECT THE DATA TO UPLOAD`. После загрузки файла необходимо указать его формат, пункт `SELECT A TYPE FOR THE NEW DATASET`. В рамках курса мы будем работать с CSV файлами, содержащими заголовки (пункт `Generic CSV File with a header (.csv)`), или с файлами без заголовков (пункт `Generic CSV File with no header (.nh.csv)`). После выбора типа файла остается только нажать «галочку» для подтверждения и начала загрузки.



(a) Загрузка файла данных.



(b) Выбор типа файла.

Рис. 4: Импорт наборов данных в Azure ML Studio.

Статус загрузки данных отображается в нижней часте экрана.

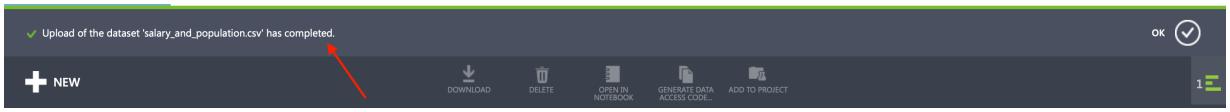


Рис. 5: Загрузка данных.

1.3 Experiments

Раздел меню `Experiments` позволяет Вам просмотреть уже созданные вами эксперименты. Создать новый можно с помощью кнопки `+ NEW` в левом нижнем углу, после чего нажать «`Blank Experiment`». Кроме создания собственного эксперимента, в этом меню можно посмотреть коллекцию готовых примеров.

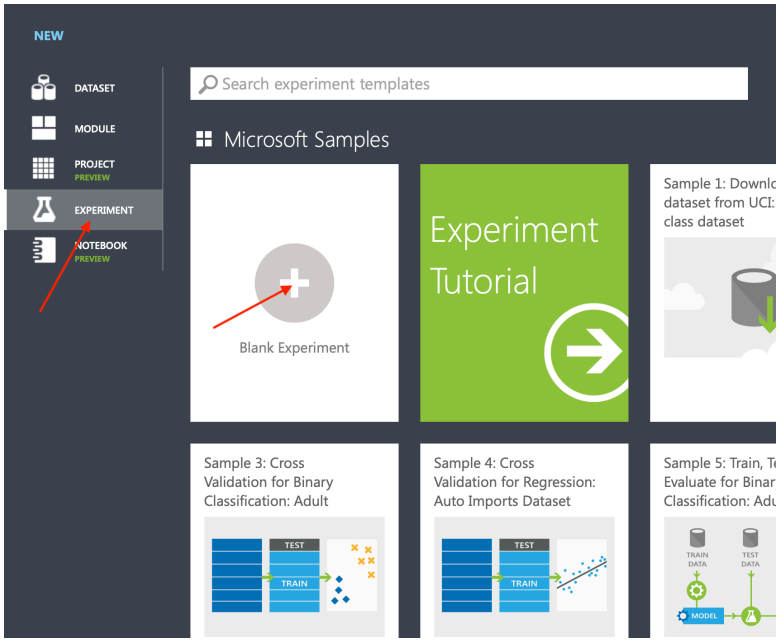


Рис. 6: Создание нового эксперимента.

Рабочее пространство эксперимента состоит из меню блоков в левой части экрана, панели настроек в правой части, рабочей области в центре, нижней панели. Кроме того, в верхней рабочей области можно сменить название эксперимента. Меню блоков содержит элементы для работы с данными, различные методы обработки данных, модели машинного обучения и многое другое, всем этим мы будем пользоваться. Рабочая область – это место, где мы будем строить цепочку для анализа данных, используя готовые блоки. Для добавления нового блока, его достаточно просто перетянуть на рабочую область. Панель настроек блока становится активна, если выбрать блок на рабочей области. Нижняя панель техническая: она позволяет запустить проект, сохранить проект и проч.

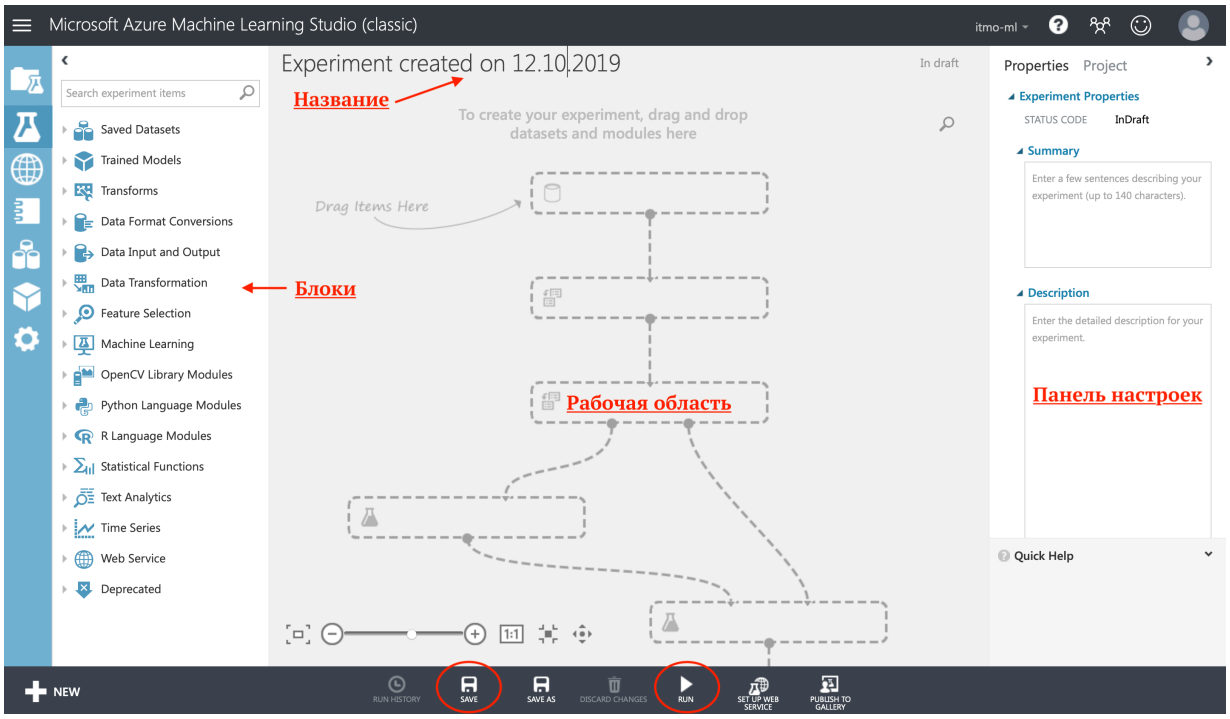


Рис. 7: Рабочее пространство.

Вкладка `Saved Datasets` содержит как загруженные ранее наборы данных, так и предустановленную коллекцию датасетов. Необходимый набор данных следует перетянуть на рабочую область. Как правило, любой новый эксперимент начинается с подгрузки данных. Просто перетяните данные из пункта `Saved Datasets`.

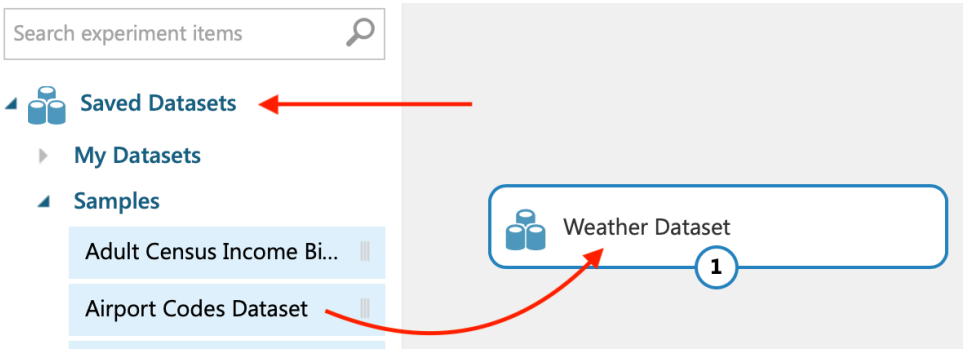


Рис. 8: Выбор данных.

По нажатию правой кнопки мыши на блоке, открывается контекстное меню параметров. Доступны стандартные действия (удалить, скопировать и т.д.), но также и возможности, доступные в зависимости от блока. Посмотреть на структуру блока данных можно нажав на пункт `Visualize`.

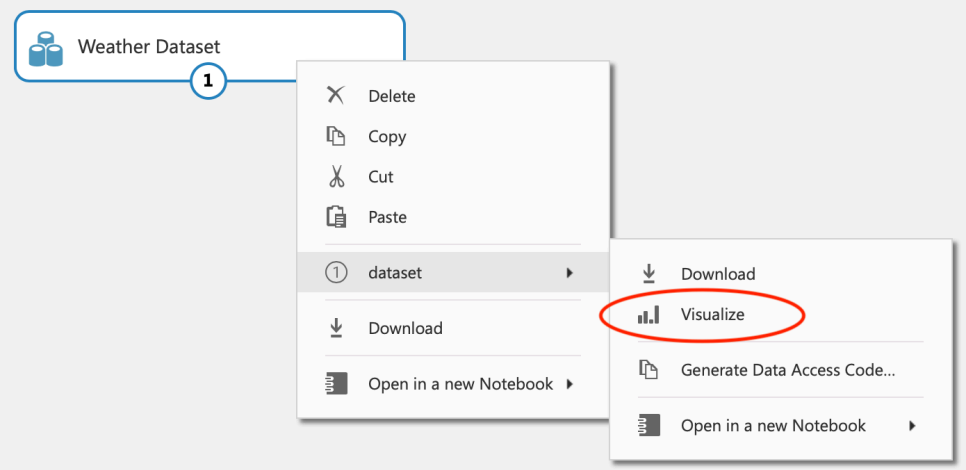


Рис. 9: Контекстное меню параметров блока.

В новом окне открывается сам набор данных, при этом CSV файл будет автоматически представлен в виде колонок (еще раз отметим, что разделителем должна быть запятая). У Вас имеется возможность ознакомиться как с самими данными, так и с основными описательными статистиками. Нажимая на названиях столбцов, в правой части окна отображается дополнительная информация. Для строковых данных доступны:

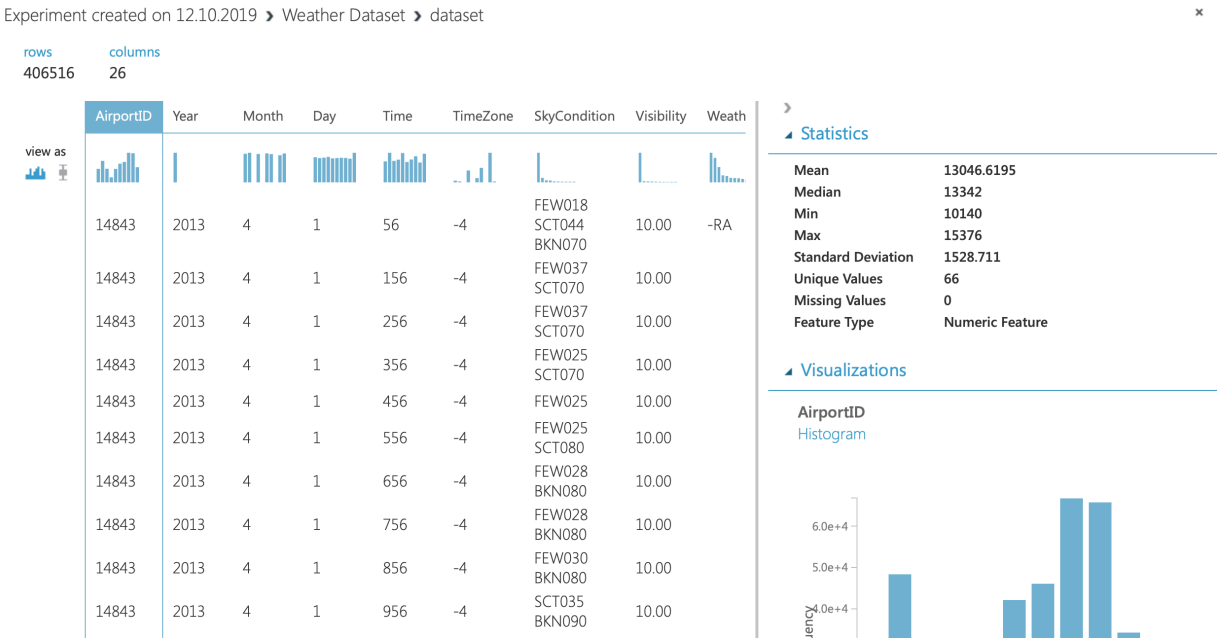


Рис. 10: Окно просмотра данных.

- Unique Values – число уникальных значений;
- Missing Values – число пропусков в столбце.

Для числовых данных доступны:

- Mean – выборочное среднее $\overline{X}$;
- Median – выборочная медиана Me;
- Min/Max – минимальное и максимальное значение;
- Standard Deviation – среднеквадратическое отклонение σ;
- Unique Values – число уникальных значений;
- Missing Values – число пропусков в столбце.

2 Сборка простого эксперимента

Разберем некоторые блоки, которыми мы будем пользоваться далее. Начнем с того, что многие данные будут избыточны и нам потребуется выбирать только некоторый поднабор исходного набора данных – только часть столбцов и/или часть строк.

2.1 Select Columns in Dataset

Блок `Select Columns in Dataset`, который можно найти по поиску, либо выбрать вручную отвечает за отбор столбцов.

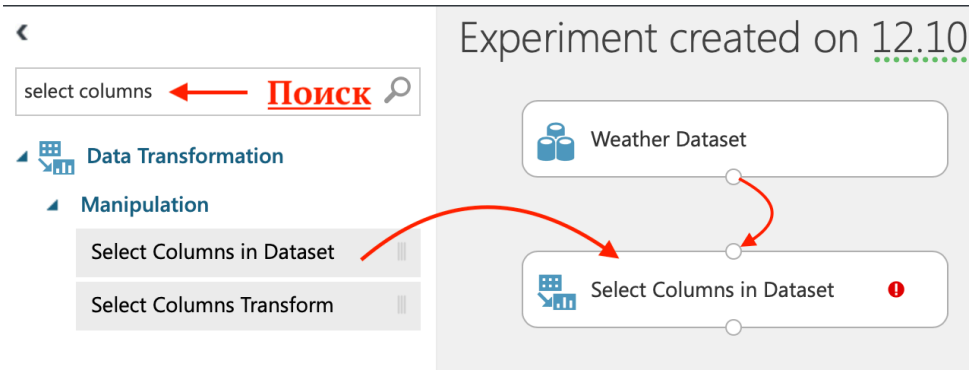


Рис. 11: Блок выбора столбцов.

Выбрав блок, в правой части экрана становится активным меню настроек (меню выбора необходимых столбцов). Оператор `Launch column selector` позволяет выбрать необходимые столбцы.

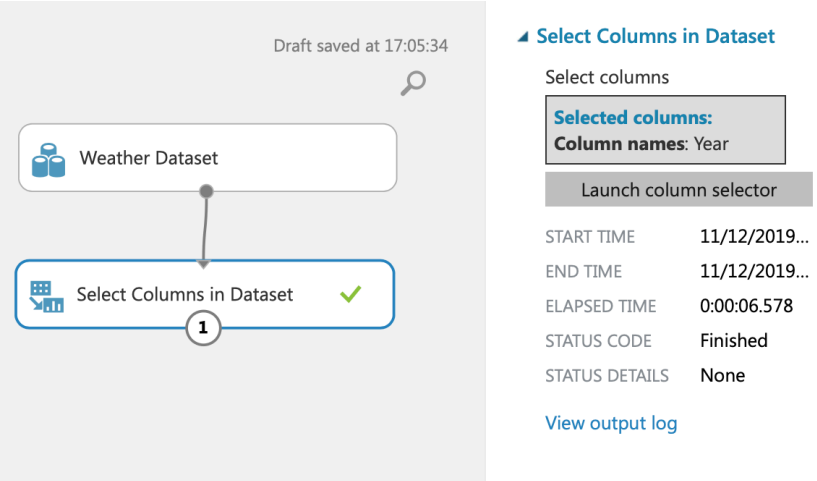


Рис. 12: Настройки блока выбора столбцов.

В окне `Select columns` можно выбрать колонки таблицы, перетащив имена столбцов в правую область.

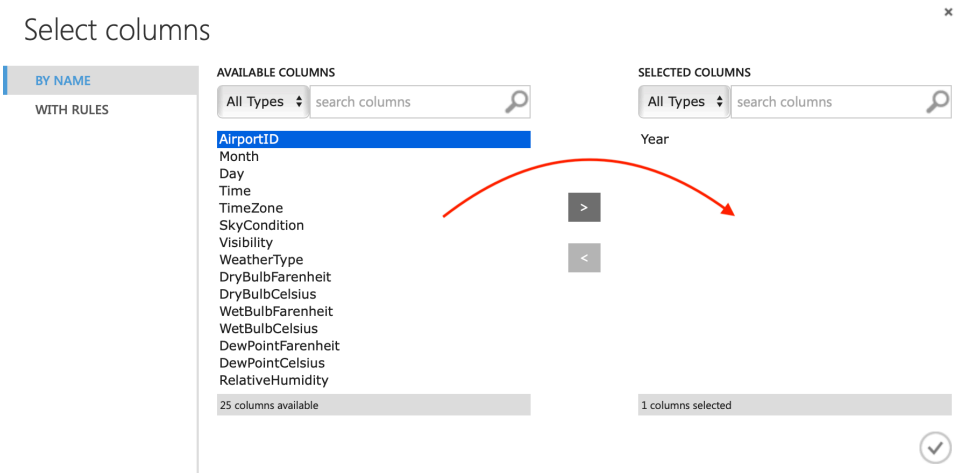


Рис. 13: Выбор столбцов из набора данных.

2.2 Apply SQL Transformation

Блок `Apply SQL Transformation` универсален, и позволяет работать с набором данных как с таблицей в базе данных, то есть осуществлять запросы к этой таблице. Запрос по-умолчанию (в поле `SQL Query Script`)

```
select * from t1;
```

позволяет получить все данные из подключенного набора данных. Имя таблицы `t1` используется для данных, подключенных к первому входу блока `Apply SQL Transformation`.

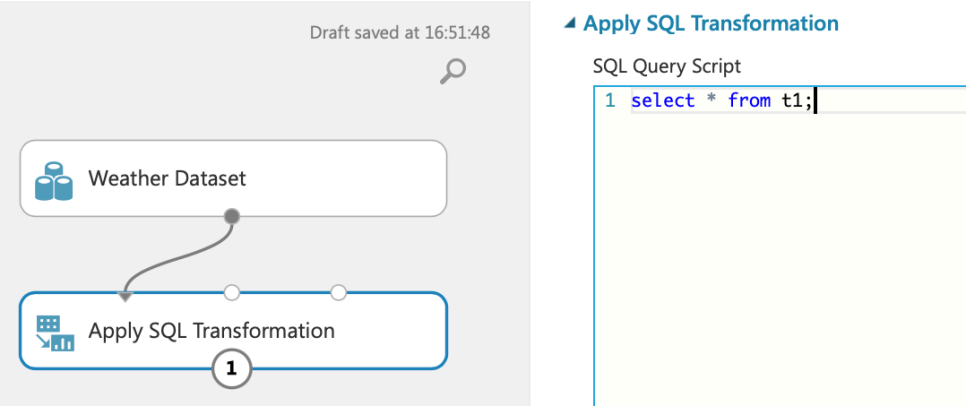


Рис. 14: Исполнение запросов к набору данных.

Приведем несколько примеров часто используемых запросов. Отберем из набора данных лишь определенные строки. К примеру все аэропорты с идентификатором 14843:

```
select * from t1 where AirportID = 14843;
```

или данные только за четвертый и пятый месяц:

```
select * from t1 where Month in (4, 5);
```

Чтобы исключить часть данных воспользуемся командой отрицания `not`. Исключим информацию о четвертом и пятом месяце:

```
select * from t1 where Month not in (4, 5);
```

При работе со строковыми данными не забывайте использовать одинарные кавычки. Например, исключим из набора данных информацию о чистом небе (`SkyCondition` принимает значение `CLR`):

```
select * from t1 where SkyCondition not in ('CLR');
```

2.3 Compute Elementary Statistics

Добавив блок `Compute Elementary Statistics` и соединив, как показано на рисунке, выберем параметр медиана (Median). Медиана будет найдена для всех числовых столбцов.

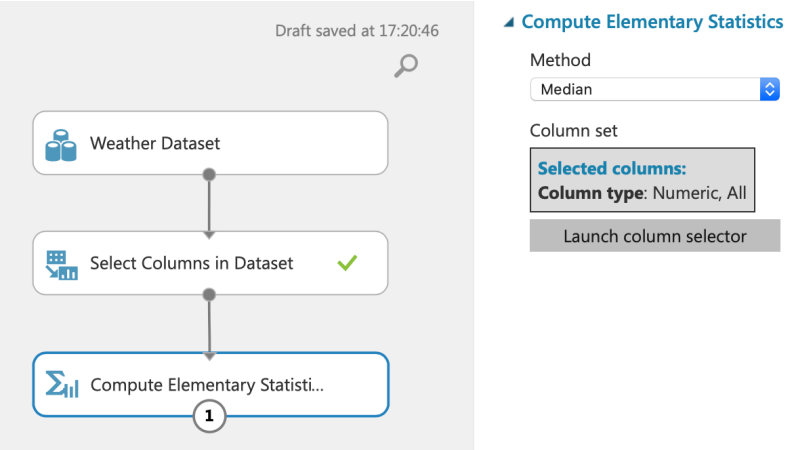


Рис. 15: Вычисление медианы.

Для того чтобы запустить эксперимент, нужно нажать кнопку `Run` на нижней панели, после чего станет доступен просмотр вычисленных значений.

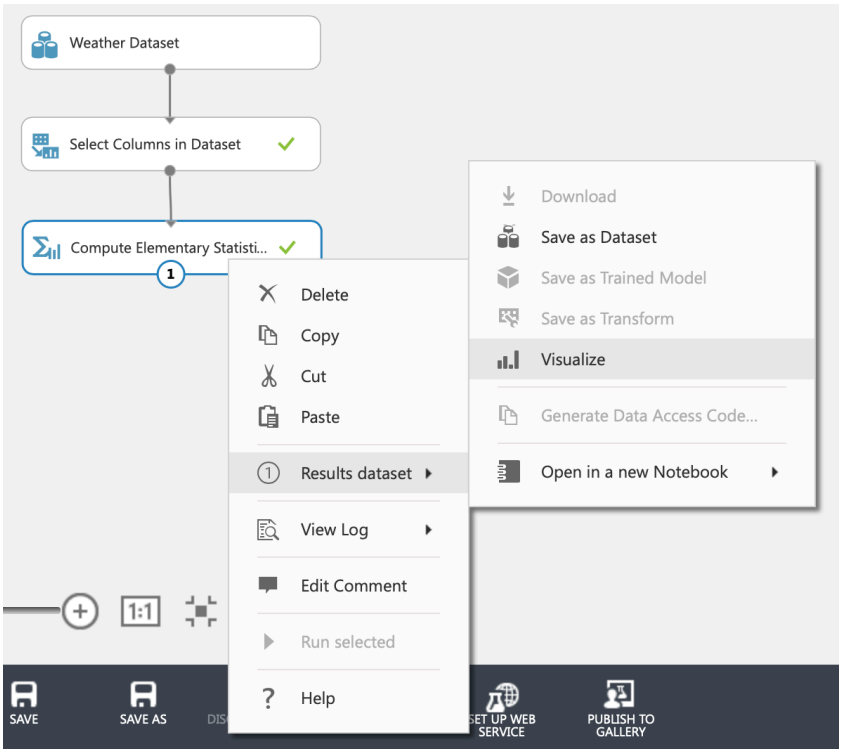


Рис. 16: Просмотр результатов после запуска эксперимента.