

Исследовательский анализ данных с использованием MitoSheet

Разведочный анализ данных (англ. exploratory data analysis, EDA) — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.

Понятие введено математиком Джоном Тьюки, который сформулировал цели такого анализа следующим образом:

- максимальное «проникновение» в данные,
- выявление основных структур,
- выбор наиболее важных переменных,
- обнаружение отклонений и аномалий,
- проверка основных гипотез,
- разработка начальных моделей.

Основные средства разведочного анализа — изучение вероятностных распределений переменных, построение и анализ корреляционных матриц, факторный анализ, дискриминантный анализ

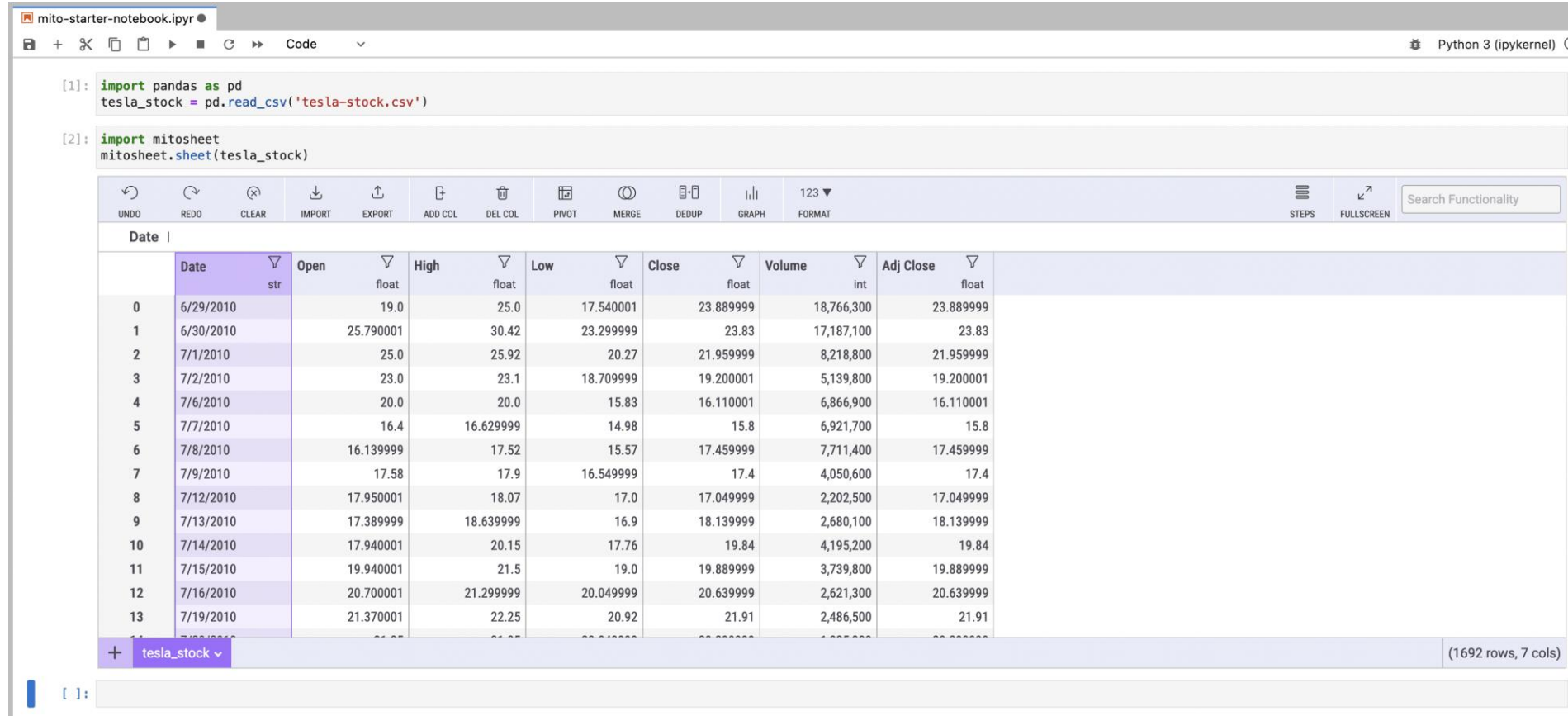
Исследовательский анализ данных - самый важный и решающий шаг в науке о данных. Он не только предоставляет нам информацию о точках данных и функциях, но также помогает обнаруживать различные шаблоны данных, ассоциации, зависимости данных и др.

Существует большое количество библиотек Python, которые могут помочь в процессе EDA, но либо они не удобны для пользователя, либо нам нужно работать с комбинацией различных библиотек для выполнения всех шагов EDA.

MitoSheet - это библиотека Python с открытым исходным кодом, которая используется для создания интерактивных информационных панелей, где мы можем анализировать данные, визуализировать различные функции и соответствующим образом манипулировать данными. Рассмотрим, как мы можем использовать различные функции MitoSheet.

MitoSheet в блокноте

- Mitosheet — это просто электронная таблица, которая существует в вашем блокноте JupyterLab. Это позволяет вам *очень быстро* переключаться между анализом данных в Python и изучением/манипулированием данными в Mitosheet.
- Поскольку Mitosheet оптимизирован для вашего анализа данных Python, каждая вкладка внутри вашего Mitosheet представляет собой dataframe Pandas. Давайте взглянем на отображение одного dataframe (цен акций Tesla) в Mitosheet:



The screenshot shows the MitoSheet interface within a JupyterLab notebook. The notebook has a tab titled 'mito-starter-notebook.ipynb'. The code cell contains two lines of Python code:

```
[1]: import pandas as pd
tesla_stock = pd.read_csv('tesla-stock.csv')

[2]: import mitosheet
mitosheet.sheet(tesla_stock)
```

Below the code cell, the Mitosheet interface is displayed. It features a toolbar with various data manipulation tools: UNDO, REDO, CLEAR, IMPORT, EXPORT, ADD COL, DEL COL, PIVOT, MERGE, DEDUP, GRAPH, and FORMAT. The table below the toolbar displays stock data for Tesla from 2010. The table has columns for Date, Open, High, Low, Close, Volume, and Adj Close. The data is loaded from a CSV file named 'tesla-stock.csv'.

	Date	Open	High	Low	Close	Volume	Adj Close
0	6/29/2010	19.0	25.0	17.540001	23.889999	18,766,300	23.889999
1	6/30/2010	25.790001	30.42	23.299999	23.83	17,187,100	23.83
2	7/1/2010	25.0	25.92	20.27	21.959999	8,218,800	21.959999
3	7/2/2010	23.0	23.1	18.709999	19.200001	5,139,800	19.200001
4	7/6/2010	20.0	20.0	15.83	16.110001	6,866,900	16.110001
5	7/7/2010	16.4	16.629999	14.98	15.8	6,921,700	15.8
6	7/8/2010	16.139999	17.52	15.57	17.459999	7,711,400	17.459999
7	7/9/2010	17.58	17.9	16.549999	17.4	4,050,600	17.4
8	7/12/2010	17.950001	18.07	17.0	17.049999	2,202,500	17.049999
9	7/13/2010	17.389999	18.639999	16.9	18.139999	2,680,100	18.139999
10	7/14/2010	17.940001	20.15	17.76	19.84	4,195,200	19.84
11	7/15/2010	19.940001	21.5	19.0	19.889999	3,739,800	19.889999
12	7/16/2010	20.700001	21.299999	20.049999	20.639999	2,621,300	20.639999
13	7/19/2010	21.370001	22.25	20.92	21.91	2,486,500	21.91

The table is displayed with a toolbar at the bottom showing the number of rows and columns: (1692 rows, 7 cols).

Mitosheet также может помочь вам легко импортировать файлы CSV и Excel в dataframe одним нажатием кнопки.

Deduplicate

Функция дедупликации Mito — это скрытно мощный инструмент для удаления нежелательных данных.

Чтобы использовать функцию дедупликации:

1. Нажмите кнопку дедупликации на панели инструментов Mito.
2. Выберите, какую запись дублированных данных вы хотите сохранить: первую, последнюю или ни одной.
3. Настройте, какие столбцы использовать для поиска повторяющихся данных. Две строки считаются дубликатами друг друга, если они имеют одинаковое значение во всех столбцах, выбранных в разделе Столбцы для дедупликации.

UNDO

REDO

CLEAR

IMPORT

EXPORT

ADD COL

DEL COL

PIVOT

MERGE

DEDUP

GRAPH

123

FORMAT

STEPS

FULLSCREEN

Search Functionality

PTS |

	Player	Tm	PTS
28	Cat Barber	ATL	0.0
551	M.J. Walker	PHO	0.0
98	Zylan Cheatham	UTA	0.0
235	Jaylen Hoard	OKC	0.0
41	Jordan Bell	CHI	0.0
266	DeJon Jarreau	IND	0.0
207	Tyler Hall	NYK	0.0
308	Arnoldas Kulboka	CHO	0.0
311	Anthony Lamb	SAS	0.0
388	Ade Murkey	SAC	0.0
128	Sam Dekker	TOR	0.0
584	Moses Wright	LAC	0.0
361	JaQuori McLaughlin	DAL	0.0
329	Didi Louzada	NOP	0.0

Drop Duplicates

Sheet to Deduplicate

Duplicates to Keep

Columns to Deduplicate On

Removed 0 rows

NBA_Player_Stats

first

Search

☒ Toggle All

☒ Player

☒ Tm

☒ PTS

(555 rows, 3 cols)

Изменение форматирования столбцов

Использовать кнопку форматирования на панели инструментов просто:

- 1. Выберите все столбцы, форматирование которых вы хотите обновить (вы можете выбрать сразу несколько столбцов, удерживая нажатой клавишу «Command» на Mac или «Control» на Windows).
- 2. Нажмите кнопку Формат на панели инструментов.
- 3. Выберите форматирование для применения к выбранным столбцам.

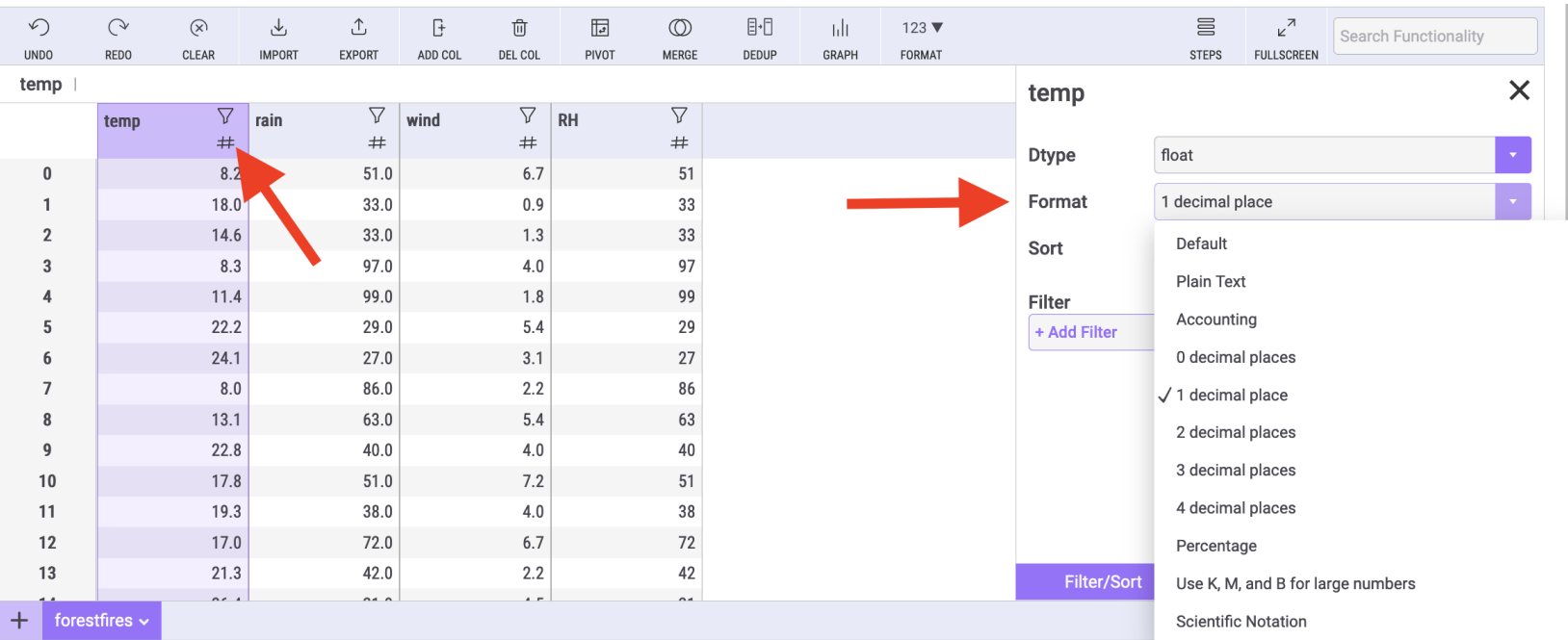
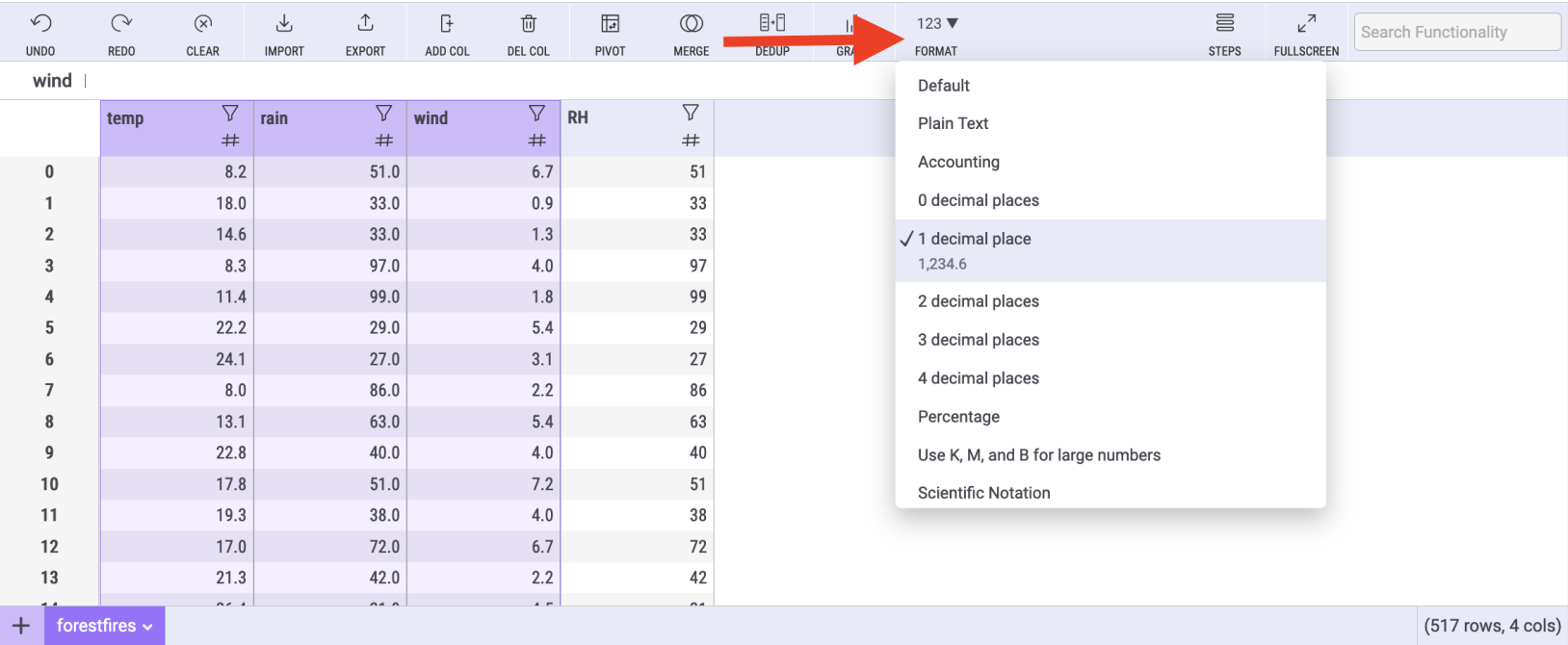
Если какой-либо из выбранных вами столбцов не является числовым, к этим столбцам не будет применено форматирование.

Использование панели управления колонкой

Другой способ изменить форматирование столбца — использовать панель управления столбцом.

Откройте панель управления колонкой, дважды щелкнув значок типа в заголовке колонки. В этом примере значок типа — это знак #.

Используйте выбор формата, чтобы выбрать форматирование для применения к столбцу.



Изучение данных в Mitosheet

- Как только у нас появится кадр данных, отображаемый в Mitosheet, мы можем сразу же начать исследовать наши данные, как в стандартной электронной таблице. Давайте выделим несколько способов использования Mitosheet для изучения набора данных.

- В отличие от простой распечатки фрейма данных, Mito сразу же позволяет увидеть весь набор данных. Как вы можете видеть на скриншоте выше, Mito позволяет невероятно легко:

- См. заголовки столбцов вашего dataframe.
- Прокрутите строки dataframe и найдите интересные точки данных.

Headers →

Column Dtypes

Dataframe Data →

Dataframes in this mitosheet

Data Shape → (1692 rows, 7 cols)

	Date	Open	High	Low	Close	Volume	Adj Close
0	6/29/2010	19.0	25.0	17.540001	23.889999	18,766,300	23.889999
1	6/30/2010	25.790001	30.42	23.299999	23.83	17,187,100	23.83
2	7/1/2010	25.0	25.92	20.27	21.959999	8,218,800	21.959999
3	7/2/2010	23.0	23.1	18.709999	19.200001	5,139,800	19.200001
4	7/6/2010	20.0	20.0	15.83	16.110001	6,866,900	16.110001
5	7/7/2010	16.4	16.629999	14.98	15.8	6,921,700	15.8
6	7/8/2010	16.139999	17.52	15.57	17.459999	7,711,400	17.459999
7	7/9/2010	17.58	17.9	16.549999	17.4	4,050,600	17.4
8	7/12/2010	17.950001	18.07	17.0	17.049999	2,202,500	17.049999
9	7/13/2010	17.389999	18.639999	16.9	18.139999	2,680,100	18.139999
10	7/14/2010	17.940001	20.15	17.76	19.84	4,195,200	19.84
11	7/15/2010	19.940001	21.5	19.0	19.889999	3,739,800	19.889999
12	7/16/2010	20.700001	21.299999	20.049999	20.639999	2,621,300	20.639999
13	7/19/2010	21.370001	22.25	20.92	21.91	2,486,500	21.91

3. Просмотрите типы каждого из столбцов в вашем dataframe.

4. Посмотрите количество строк и столбцов в dataframe.

- Существует множество более продвинутых способов изучения и понимания ваших данных, которые позволяют легко получить доступ к сводной статистике по столбцам , построить графики данных и многое другое

Редактирование данных в MitoSheet

Помимо того, что вы можете просматривать свои данные в электронной таблице, MitoSheet позволяет легко редактировать свои данные. Можно найти большинство основных действий на панели инструментов в верхней части MitoSheet.

Включенные действия позволяют редактировать столбцы в dataframe, создавать сводные таблицы, объединять несколько фреймов данных вместе и многое другое.

The screenshot displays the MitoSheet interface within a Jupyter Notebook. The top section shows two code cells: the first imports pandas and reads a CSV file named 'tesla-stock.csv'; the second imports MitoSheet and creates a sheet from the loaded data. Below the code is the MitoSheet toolbar, which includes icons for Undo, Redo, Clear, Import, Export, Add Column, Delete Column, Pivot, Merge, Dedup, Graph, and Format. A red box highlights the toolbar area. To the right of the toolbar is a search bar labeled 'Search Functionality'. Below the toolbar is a data table with columns: Date, Open, High, Low, Close, Volume, and Adj Close. The table contains 14 rows of data. At the bottom of the table, it indicates '(1692 rows, 7 cols)'. A red arrow points from the text 'Mito Toolbar' to the toolbar area.

	Date	Open	High	Low	Close	Volume	Adj Close
0	6/29/2010	19.0	25.0	17.540001	23.889999	18,766,300	23.889999
1	6/30/2010	25.790001	30.42	23.299999	23.83	17,187,100	23.83
2	7/1/2010	25.0	25.92	20.27	21.959999	8,218,800	21.959999
3	7/2/2010	23.0	23.1	18.709999	19.200001	5,139,800	19.200001
4	7/6/2010	20.0	20.0	15.83	16.110001	6,866,900	16.110001
5	7/7/2010	16.4	16.629999	14.98	15.8	6,921,700	15.8
6	7/8/2010	16.139999	17.52	15.57	17.459999	7,711,400	17.459999
7	7/9/2010	17.58	17.9	16.549999	17.4	4,050,600	17.4
8	7/12/2010	17.950001	18.07	17.0	17.049999	2,202,500	17.049999
9	7/13/2010	17.389999	18.639999	16.9	18.139999	2,680,100	18.139999
10	7/14/2010	17.940001	20.15	17.76	19.84	4,195,200	19.84
11	7/15/2010	19.940001	21.5	19.0	19.889999	3,739,800	19.889999
12	7/16/2010	20.700001	21.299999	20.049999	20.639999	2,621,300	20.639999
13	7/19/2010	21.370001	22.25	20.92	21.91	2,486,500	21.91

Кроме того, вы можете использовать панель поиска действий для поиска любых функций, включенных в MitoSheet. Если вы не знаете, как найти некоторые функции в MitoSheet, легко их найти

Редактирование столбцов на листе

Одной из самых основных операций, которые вы можете выполнять в любой электронной таблице, является добавление столбца и написание формулы электронной таблицы. Давайте воспользуемся MitoSheet для расчета движения цены акций Tesla в течение каждого торгового дня. Во-первых, используйте кнопку добавления столбца, чтобы добавить столбец в набор данных. Затем дважды щелкните заголовок столбца, чтобы отредактировать его напрямую и переименовать

The screenshot shows the MitoSheet interface within a Jupyter Notebook. The code cell [1] imports pandas and reads 'tesla-stock.csv'. Cell [2] imports MitoSheet and creates a sheet from the data. The table below has columns: Date, Open, High, Low, Close, Movement, Volume, and Adj Close. A new column 'new-column-67qc' has been added at index 5. Red arrows highlight the 'ADD COL' button and the 'Movement' column header.

	Date	Open	High	Low	Close	Movement	Volume	Adj Close
0	6/29/2010	19.0	25.0	17.540001	23.889999	0	18,766,300	23.889999
1	6/30/2010	25.790001	30.42	23.299999	23.83	0	17,187,100	23.83
2	7/1/2010	25.0	25.92	20.27	21.959999	0	8,218,800	21.959999
3	7/2/2010	23.0	23.1	18.709999	19.200001	0	5,139,800	19.200001
4	7/6/2010	20.0	20.0	15.83	16.110001	0	6,866,900	16.110001
5	7/7/2010	16.4	16.629999	14.98	15.8	0	6,921,700	15.8
6	7/8/2010	16.139999	17.52	15.57	17.459999	0	7,711,400	17.459999
7	7/9/2010	17.58	17.9	16.549999	17.4	0	4,050,600	17.4
8	7/12/2010	17.950001	18.07	17.0	17.049999	0	2,202,500	17.049999
9	7/13/2010	17.389999	18.639999	16.9	18.139999	0	2,680,100	18.139999
10	7/14/2010	17.940001	20.15	17.76	19.84	0	4,195,200	19.84
11	7/15/2010	19.940001	21.5	19.0	19.889999	0	3,739,800	19.889999
12	7/16/2010	20.700001	21.299999	20.049999	20.639999	0	2,621,300	20.639999
13	7/19/2010	21.370001	22.25	20.92	21.91	0	2,486,500	21.91

tesla_stock (1692 rows, 8 cols)

```
[ ]: from mitosheet import *; register_analysis('UUID-53286588-678a-4d7d-bfd8-14d77bb757dc')

# Added column new-column-67qc to tesla_stock
tesla_stock.insert(5, 'new-column-67qc', 0)
```

- Теперь, когда у нас есть новый столбец, мы можем написать внутри него формулу! Просто:

1. Двойной щелчок по ячейке в столбце
2. Установите формулу, ссылаясь на заголовки других столбцов по их имени.

После нажатия клавиши ввода мы видим, что легко вычислили новый столбец в нашем фрейме данных, точно так же, как в электронной таблице.

Для каждого редактирования, которое вы вносите в MitoSheet, Mito генерирует приведенный ниже код, соответствующий этому редактированию. Как вы можете видеть на скриншоте ниже, мы добавили столбец и задали его формулу, и сгенерированный нами код делает то же самое. Теперь, когда вы сгенерировали код, вы можете его запустить! Запустив этот сгенерированный код, вы можете зафиксировать изменения, внесенные вами в ваши фреймы данных в MitoSheet (изменения, которые вы вносите, по умолчанию применяются только к копии фрейма данных). Чтобы использовать отредактированные кадры данных в остальной части вашего анализа, просто запустите сгенерированный код и продолжайте писать Python ниже, как обычно!

The image displays two screenshots of the Mito notebook interface, illustrating the process of adding a new column with a formula.

Top Screenshot: The notebook shows a data table with columns: Date, Open, High, Low, Close, Movement, Volume, and Adj Close. The 'Movement' column is highlighted, and a formula `=Close - Open` is being entered into the cell for the first row (6/29/2010). A red arrow points to the formula bar with the text "Writing a column formula".

Bottom Screenshot: The notebook shows the same data table, but now the 'Movement' column has been added with the formula `=Close - Open`. The code generated by Mito is displayed in the notebook cells:

```
[1]: import pandas as pd
tesla_stock = pd.read_csv('tesla-stock.csv')

[2]: import mitosheet
mitosheet.sheet(tesla_stock)
```

The generated code includes comments explaining the changes:

```
# Added column new-column-67qc to tesla_stock
tesla_stock.insert(5, 'new-column-67qc', 0)

# Renamed new-column-67qc to Movement in tesla_stock
tesla_stock.rename(columns={'new-column-67qc': 'Movement'})

# Set new-column-67qc in tesla_stock to =Close - Open
tesla_stock['Movement'] = tesla_stock['Close'] - tesla_stock['Open']
```

Создание сводной таблицы

- Чтобы создать сводную таблицу, щелкните значок сводной таблицы на панели инструментов Mito, а затем используйте панель задач для ее настройки.

Чтобы настроить сводную таблицу:

- Во-первых, используйте раздел строк, чтобы выбрать создание ключа для группировки фрейма данных по
- Затем, если вы хотите дополнительно разделить группы на отдельные ячейки, используйте раздел столбцов
- Наконец, используйте раздел значений, чтобы решить, как агрегировать данные в этих корзинах. Вы можете переключать методы агрегации, используя раскрывающийся список, который по умолчанию использует агрегацию «подсчет».

Чтобы отредактировать сводную таблицу, которую вы настроили, снова нажмите кнопку сводной таблицы, и снова появится меню.

Поиск функциональности

Mito содержит панель поиска, которая позволяет вам найти и получить доступ ко всем функциям, которые может предложить Mito. Просто нажмите на строку поиска в правом верхнем углу и начните вводить текст, чтобы найти нужную функцию.

	Brand	Brand count
0	1 To 3 Noodles	1
1	7 Select	2
2	7 Select/Nissin	1
3	A-One	4
4	A-Sha Dry Noodle	26
5	A1	3
6	ABC	12
7	Acecook	15
8	Adabi	4
9	Ah Lai	2
10	Ajinatori	2
11	Amianda	10
12	Amino	3
13	Annie Chun's	12

	First Name	Last Name	Gender	Country	Age	Date	Id
0	Dulce	Abril	Female	United States	32	15/10/2017	1562
1	Mara	Hashimoto	Female	Great Britain	25	16/08/2016	1582
2	Philip	Gent	Male	France	36	21/05/2015	2587
3	Kathleen	Hanner	Female	United States	25	15/10/2017	3549
4	Nereida	Magwood	Female	United States	58	16/08/2016	2468
5	Gaston	Brumm	Male	United States	24	21/05/2015	2554
6	Etta	Hurn	Female	Great Britain	56	15/10/2017	3598
7	Earlean	Melgar	Female	United States	27	16/08/2016	2456
8	Vincenza	Weiland	Female	United States	40	21/05/2015	6548

Фильтровать по условию

- Mito обеспечивает мощную фильтрацию за счет комбинации фильтров и групп фильтров.
1. Фильтры — это одно условие, которое оценивается как истинное или ложное для каждой ячейки в столбце.
 2. Группы фильтров — это совокупность фильтров, объединенных оператором И/ИЛИ
- Чтобы получить доступ к фильтрам столбцов Mito, нажмите кнопку фильтра в заголовке столбца, а затем щелкните раскрывающийся список «Добавить фильтр».
 - Фильтры Mito являются приемлемыми, то есть в наборе данных останутся только те ячейки, для которых выполнены условия фильтра.

UNDOREDOCLEARIMPORTEXPORTADD COLDEL COLPIVOTMERGEDEDUPGRAPH123▼FORMATSTEPFULLSCREENSearch Functiona

Brand |

	Review # #	Brand Abc	Variety Abc	Style Abc
0	2,580	New Touch	T's Restaurant Tantan	Cup
1	2,579	Just Way	Noodles Spicy Hot Se	Pack
2	2,578	Nissin	Cup Noodles Chicken	Cup
3	2,577	Wei Lih	GGE Ramen Snack To	Pack
4	2,576	Ching's Secret	Singapore Curry	Pack
5	2,575	Samyang Foods	Kimchi song Song Ra	Pack
6	2,574	Acecook	Spice Deli Tantan Me	Cup
7	2,573	Ikeda Shoku	Nabeyaki Kitsune Ude	Tray
8	2,572	Ripe'n'Dry	Hokkaido Soy Sauce	Pack
9	2,571	KOKA	The Original Spicy Sti	Pack
10	2,570	Tao Kae Noi	Creamy tom Yum Kun	Pack
11	2,569	Yamachan	Yokohama Tonkotsu	Pack
12	2,568	Nongshim	Mr. Bibim Stir-Fried K	Pack
13	2,567	Nissin	Deka Buto Kimchi Po	Bowl
14	2,566	Nissin	Deka Buto Ramen Pack	Pack

+

ramen_ratings ▼

Brand

Dtypestr

FormatDefault

Mito only supports formatting columns with dtype int or float.

SortAscendingDescending

Filter

Wherecontains

+ Add Filter

Filter/Sort

Values

Summary Stats

(2580 rows, 7 cols)

Фильтровать по значению

- Фильтрация Mito по значению предназначена для того, чтобы помочь вам определить все уникальные значения в столбце и легко удалить ненужные.
- Получить доступ к фильтрации по уникальным значениям на вкладке «Значения» панели управления столбцами. Для каждого уникального значения в столбце на вкладке отображаются: само значение, количество строк, содержащих это значение, и процент строк, содержащих это значение.
- Чтобы отфильтровать значение из столбца, просто установите флажок рядом со значением, которое вы хотите удалить.
- Чтобы помочь вам определить значения, которые вы, возможно, захотите удалить, вы можете отсортировать и найти уникальное значение

UNDOREDOCLEARIMPORTEXPORTADD COLDEL COLPIVOTMERGEGRAPHSAVEREPLAYSTEPSDOCSFULLSCREEN

State |

	Zip	City	State	Division	Parking	Pets
1	21001	Aberdeen	MD	Mid-Atlantic	N	N
2	97321	Albany	OR	Pacific Northwest	Y	Y
3	49224	Albion	MI	Central	N	Y
4	87102	Albuquerque	NM	Southwest	Y	Y
5	24910	Alderson	WV	Mid-Atlantic	N	N
6	22301	Alexandria	VA	Mid-Atlantic	Y	Y
7	44601	Alliance	OH	Central	N	N
8	49707	Alpena	MI	Central	N	N
9	79830	Alpine	TX	Southwest	N	N
10	62002	Alton	IL	Central	Y	N
11	16601	Altoona	PA	Mid-Atlantic	Y	N
12	1002	Amherst	MA	Northeast	N	Y
13	12010	Amsterdam	NY	Northeast	N	N
14	92806	Anaheim	CA	Southwest	Y	Y
15	48104	Ann Arbor	MI	Central	Y	N

+ Airport_Pets_csv

State

Unique Values (Toggle to Filter)

search values

Ascending Value

☒ AL

3 (0.35%)

☒ AR

5 (0.59%)

☒ AZ

18 (2.11%)

☒ CA

195 (22.83%)

☒ CO

14 (1.64%)

☒ CT

7 (0.82%)

☒ DC

1 (0.12%)

☒ DE

3 (0.35%)

☒ FL

32 (3.75%)

☒ GA

5 (0.59%)

☒ IA

7 (0.82%)

☒ ID

15 (1.76%)

☒ IL

36 (4.22%)

☒ IN

11 (1.29%)

☒ KS

6 (0.70%)

Filter/Sort

Values

Summary Stats

(854, 8)

Изменения типа

- Изменения типа «укажи и щелкни» в Mito упрощают определение типа данных каждого столбца в вашем фрейме данных и приведение этих столбцов к другому типу.

Определение типов данных

- Для каждого столбца в вашем фрейме данных тип данных отображается в заголовке столбца. Значок поможет вам различать целые числа, числа с плавающей запятой, строки, логические значения и дельты дат/времени.

UNDO

REDO

CLEAR

IMPORT

EXPORT

ADD COL

DEL COL

PIVOT

MERGE

DEDUP

GRAPH

123

FORMAT

STEPS

FULLSCREEN

Search Functiona

Stars |

	Review # ##	Brand Abc	Variety Abc	Style Abc
0	2,580	New Touch	T's Restaurant Tantan	Cup
1	2,579	Just Way	Noodles Spicy Hot Se	Pack
2	2,578	Nissin	Cup Noodles Chicken	Cup
3	2,577	Wei Lih	GGE Ramen Snack To	Pack
4	2,576	Ching's Secret	Singapore Curry	Pack
5	2,575	Samyang Foods	Kimchi song Song Ra	Pack
6	2,574	Acecook	Spice Deli Tantan Me	Cup
7	2,573	Ikeda Shoku	Nabeyaki Kitsune Udo	Tray
8	2,572	Ripe'n'Dry	Hokkaido Soy Sauce	Pack
9	2,571	KOKA	The Original Spicy Sti	Pack
10	2,570	Tao Kae Noi	Creamy tom Yum Kun	Pack
11	2,569	Yamachan	Yokohama Tonkotsu	Pack
12	2,568	Nongshim	Mr. Bibim Stir-Fried K	Pack
13	2,567	Nissin	Deka Buto Kimchi Pot	Bowl
14	2,566	Nissin	Deka Buto Kimchi Pot	Bowl

+

ramen_ratings

Stars

Dtype

Format

Mito only supports

Sort

Filter

+ Add Filter

str

bool

int

float

str

datetime

timedelta

Filter/Sort

Values

Summary Stats

(2580 rows, 7 cols)

Изменение типов данных

Есть два способа изменить тип данных столбцов внутри Mito.

- Во-первых, для столбцов данных (любой столбец, который не был создан формулой электронной таблицы Mito) вы можете использовать раскрывающийся список выбора типа столбца на панели управления столбцом. Просто выберите dtype, к которому вы хотите привести серию столбцов.
- Для столбцов формул можно использовать формулы электронной таблицы приведения типов. Для этого добавьте в формулу функцию VALUE, TEXT, BOOL или DATEVALUE, чтобы преобразовать столбец в число, строку, логическое значение или дату и время соответственно.
- Например, чтобы преобразовать результат оператора IF в логическое значение, заключите оператор if в функцию BOOL.
- =BOOL(IF(A > 100, 1, 0))

Графики

Графики в Mito предназначены для того, чтобы помочь вам получить интуитивное представление о ваших данных и создать готовые к презентации графики для обмена идеями. Mito создает интерактивные и настраиваемые графики, используя графическую библиотеку с открытым исходным кодом Plotly Express.



Настройка графика

Data Source: лист, содержащий данные, которые вы хотите отобразить в виде графика.

Тип диаграммы: тип диаграммы, которую вы хотите создать. Mito поддерживает: Scatter plots, Line charts, Bar charts, Histograms, Box plots, Violin plots, Strip plots, Density heatmaps, Density Countour maps

Ось X: данные для построения графика по оси X.

Вы можете выбрать несколько рядов либо вдоль оси x, либо по оси y (но не по обеим), если ряды имеют схожие dtypes.

Ось Y: данные для построения графика по оси Y.

Вы можете выбрать несколько рядов либо вдоль оси x, либо по оси y (но не по обеим), если ряды имеют схожие dtypes.

Color by Column: дополнительный столбец для дальнейшей разбивки графических данных с использованием атрибута цвета: Для достижения наилучших результатов выберите столбец с несколькими уникальными значениями. Этот параметр можно выбрать для всех графиков, за исключением тепловой карты плотности.

Filter to safe size: по умолчанию Mito отображает только первые 1000 строк данных, чтобы гарантировать, что вкладка браузера не выйдет из строя при попытке загрузить слишком много данных в график. Отключение фильтра для сохранения размера графов всего фрейма данных может замедлить или привести к сбою вкладки браузера.

Слияние данных

Функцию слияния Mito можно использовать для объединения наборов данных по горизонтали. Слияние ищет совпадения между ключевым столбцом первого листа и ключевым столбцом второго листа. Mito поддерживает все наиболее распространенные слияния.

Left Merge: включает все строки с первого листа и только совпадающие строки со второго листа. Включает все матчи.

Right Merge: включает все строки со второго листа и только совпадающие строки с первого листа. Включает все матчи.

Inner Merge: включает только те строки, которые совпадают на обоих листах.

Outer Merge: включает все строки с обоих листов, независимо от того, есть ли совпадения на другом листе.

Lookup Merge (слияние поиска): левое соединение, но включает только первое совпадение со второго листа, если их несколько. Так же, как Vlookup в Excel.

UNDOREDOCLEARIMPORTEXPORTADD COLDEL COLPIVOTMERGEDEDUPGRAPH123FORMATSTEPSFULLSCREENSearch Functiona

Brand |

	Brand	Brand count	Review #	Variety
	Abc	#	#	Abc
0	1 To 3 Noodles	1	2,351	Chatpat Masala
1	7 Select	2	2,124	Nissin Instant Noodle
2	7 Select/Nissin	1	1,703	Super Tom Yum Shrin
3	A-One	4	1,669	Mi Ly Instant Noodles
4	A-Sha Dry Noodle	26	2,216	Chow Mein Spicy Flav
5	A1	3	2,038	Vegetarian Herbal No
6	ABC	12	1,267	Mi Cup Rasa Soto Aya
7	Acecook	15	2,574	Spice Deli Tantan Me
8	Adabi	4	1,439	Mi Segera Mi Goreng
9	Ah Lai	2	1,671	Penang Favourite Wh
10	Ajinatori	2	2,427	HALAL Chicken Shoy
11	Amianda	10	1,097	Homely Dried Noodle
12	Amino	3	1,006	Barszcz czerwony (B
13	Annie Chun's	12	1,407	Ramen House Thai Co
14	Arai	2	1,749	Curry Mel Bawang Wk

Merge Sheets Together

Merge Type

lookup

First Sheet

ramen_ratings_pivot

Second Sheet

ramen_ratings

Merge Key

Brand

Columns to Keep

Search

☒ Toggle All

☒ Brand countint

Columns to Keep

Search

☒ Toggle All

☒ Review #int

☒ Varietystr

☒ Stylestr

+ ramen_ratings ramen_ratings_pivot Merged_DF

(355 rows, 8 cols)

Редактирование отдельных ячеек

Mito позволяет напрямую редактировать значение любой ячейки, если столбец не создан с помощью формулы. Просто дважды щелкните ячейку и отредактируйте значение в редакторе ячеек. Интеллектуальная обработка типов гарантирует, что ваши изменения имеют смысл в контексте ваших данных.

Поиск различий между наборами данных

Функция различия dataframe Mito позволяет находить уникальные записи в двух разных наборах данных.

Вы можете получить доступ к этой функции через панель задач слияния, а затем выбрать тип слияния «уникальный слева» или «уникальный справа».

Уникальный слева: Включает каждую строку из первого листа, которая не совпадает со вторым листом.

Уникальный справа: Включает каждую строку со второго листа, которая не имеет совпадений на первом листе.

Как и другие типы слияний, Mito использует ключ слияния, который вы выбираете для каждого листа, для обнаружения совпадений.

Если вы хотите определить уникальность, просматривая несколько столбцов в каждом фрейме данных, попробуйте создать новый столбец и с помощью функции CONCAT объединить все столбцы в один столбец. Затем выберите этот новый столбец в качестве ключа слияния!

↶

↷

✕

↓

↑

+

✕

📊

🔄

📄

📊

123 ▼

☰

↗

Search Functionality

temp |

	temp	rain	wind	RH
	##	##	##	##
0	8.2	51.0	6.7	51
1	18.0	33.0	0.9	33
2	14.6	33.0	1.3	33
3	8.3	97.0	4.0	97
4	11.4	99.0	1.8	99
5	22.2	29.0	5.4	29
6	24.1	27.0	3.1	27
7	8.0	86.0	2.2	86
8	13.1	63.0	5.4	63
9	22.8	40.0	4.0	40
10	17.8	51.0	7.2	51
11	19.3	38.0	4.0	38
12	17.0	72.0	6.7	72
13	21.3	42.0	2.2	42

Merge Sheets Together

Merge Type

lookup

First Sheet

forestfires

Columns

Search

☒ Toggle

☒ rain

☒ wind

☒ RH

inner

Only includes rows that have matches in both sheets.

outer

Includes all rows from both sheets, regardless of whether there is a match in the other sheet.

unique in left

Includes each row from the first sheet that doesn't have a match in the second sheet.

unique in right

Includes each row from second sheet that doesn't have a match in the first sheet.

Second Sheet

forestfires_new

Columns

Search

☒ Toggle

☒ rain

+

forestfires_old ▼

forestfires_new ▼

df3 ▼

(517 rows, 4 cols)

Объединение (concatenating) dataframes

Объединение фреймов данных позволяет вам объединять два или более фреймов данных, размещая строки друг над другом. Чтобы объединить несколько фреймов данных вместе в новый фрейм данных, просто нажмите кнопку Concat на панели инструментов, а затем настройте объединение.

Затем выберите тип соединения, который вы хотите использовать в объединении. При объединении фреймов данных возможны два разных типа соединений:

Внутренний: включает только те столбцы, которые совпадают на всех листах.

Внешний: включает все столбцы со всех листов, независимо от того, есть ли совпадения на других листах. Автоматически заполняет несоответствующие строки значениями NaN.

Далее у вас есть возможность игнорировать или сохранить исходные индексы в исходных кадрах данных. По умолчанию Mito сбросит индексы во вновь объединенном фрейме данных.

Наконец, установите кадры данных, которые вы хотите объединить в один новый кадр данных. Mito автоматически создаст новый лист, представляющий это соединение.

Undo

Redo

Clear

Import

Export

Add Col

Del Col

Pivot

Concat

Dedup

Graph

Format

123

Steps

Fullscreen

Search Functionality

0 |

1

	0	First Name	Last Name	Gender	Country	Age	Date	Id
	int	str	str	str	str	int	str	int
0	1	Dulce	Abril	Female	United States	32	15/10/2017	1,562
1	2	Mara	Hashimoto	Female	Great Britain	25	16/08/2016	1,582
2	3	Philip	Gent	Male	France	36	21/05/2015	2,587
3	4	Kathleen	Hanner	Female	United States	25	15/10/2017	3,549
4	5	Nereida	Magwood	Female	United States	58	16/08/2016	2,468
5	6	Gaston	Brumm	Male	United States	24	21/05/2015	2,554
6	7	Etta	Hurn	Female	Great Britain	56	15/10/2017	3,598
7	8	Earlean	Melgar	Female	United States	27	16/08/2016	2,456
8	9	Vincenza	Weiland	Female	United States	40	21/05/2015	6,548
9	1	Dulce	Abril	Female	United States	32	15/10/2017	1,562
10	2	Mara	Hashimoto	Female	Great Britain	25	16/08/2016	1,582
11	3	Philip	Gent	Male	France	36	21/05/2015	2,587
12	4	Kathleen	Hanner	Female	United States	25	15/10/2017	3,549
13	5	Nereida	Magwood	Female	United States	58	16/08/2016	2,468

Concatenate dataframes

Join Typeinner

Ignore Existing IndexesOn

Dataframes to Concatenate

+ Add

Sheet1

Sheet2

+ Sheet1 Sheet2 df3

(18 rows, 8 cols)

Экспорт в CSV и Excel

Для файлов Excel и CSV есть время и место во многих рабочих процессах обработки данных, поэтому Mito упрощает экспорт ваших данных.

Экспорт в csv — это самый быстрый способ загрузки больших наборов данных из Mito. Он поддерживает экспорт только одного листа за раз.

С другой стороны, экспорт в файлы Excel позволяет экспортировать несколько листов в одну книгу. Если у вас есть Mito Pro, вы также можете сохранить форматирование, примененное к вашим листам, с помощью форматирования Mito в экспортированном файле Excel.

Чтобы экспортировать данные, нажмите кнопку экспорта на панели инструментов Mito, а затем используйте селектор «Тип экспорта» на панели задач, чтобы выбрать между экспортом в CSV и Excel.

Поскольку Mito **генерирует код python pandas** для каждого редактирования, которое вы делаете в электронной таблице Mito, измененные кадры данных легко использовать в остальной части вашего анализа. Mito не принуждает вас выполнять весь анализ в Mito — он побуждает вас использовать остальную часть мощной экосистемы аналитики данных Python.

Чтобы использовать измененные кадры данных в остальной части вашего анализа, запустите код, который генерирует Mito, щелкнув ячейку, содержащую код, и нажмите кнопку воспроизведения на панели инструментов Jupyter (или используйте сочетание клавиш Shift + Enter)

После того, как вы запустили сгенерированный код, вы можете использовать измененные фреймы данных в своем анализе, как обычно, просто используя имена фреймов данных.

UNDO

REDO

CLEAR

IMPORT

EXPORT

ADD COL

DEL COL

PIVOT

MERGE

DEDUP

GRAPH

123

FORMAT

STEPS

FULLSCREEN

Search Functionality

Rk	Rk	Player	Pos	Age	Tm	G	GS
	#	Abc	Abc	#	Abc	#	
0	1	Precious Achiuwa	C	22	TOR		34
1	2	Steven Adams	C	28	MEM		44
2	3	Bam Adebayo	C	24	MIA		21
3	4	Santi Aldama	PF	21	MEM		25
4	5	LaMarcus Aldridge	C	36	BRK		32
5	6	Nickeil Alexander-Wa	SG	23	NOP		42
6	7	Grayson Allen	SG	26	MIL		41
7	8	Jarrett Allen	C	23	CLE		39
8	9	Jose Alvarado	PG	23	NOP		18
9	10	Justin Anderson	SF	28	IND		3
10	11	Kyle Anderson	PF	28	MEM		38
11	12	Giannis Antetokounm	PF	27	MIL		39
12	13	Thanasis Antetokoun	SF	29	MIL		28
13	14	Carmelo Anthony	PF	37	LAL		43

df_2021_2022_NBA_Player_Stats

Download

Export Type

Sheet to Export

CSV exports will not reflect any formatting changes made in Mito.

Download CSV file

Mito позволяет вам выполнять вычисления Excel в Python и генерирует для вас код Pandas . По сути, это как **руководство по написанию кода на Python**.

Bamboolib – графический
интерфейс (GUI) для pandas

Библиотека **bamboolib** — это **графический интерфейс для pandas DataFrames**, который позволяет любому работать с Python в Jupyter Notebook или JupyterLab.

Ключевые особенности:

- Интуитивно понятный графический интерфейс, который **экспортирует код Python**
- Поддерживает все распространенные преобразования и визуализации
- Преобразования идут с **полным управлением клавиатуры**
- Предоставляет передовые методы анализа для исследования данных

В платной версии:

- Добавляйте пользовательские преобразования, визуализации и загрузчики данных с помощью простых плагинов Python.
- Интегрируйте внутренние библиотеки Python вашей компании

Подготовка данных

Изменение строки в DateTime




Вы загрузили данные и увидели, что столбец даты является строкой. Теперь нажмите на тип столбца (маленькая буква сбоку от названия столбца), выберите новый тип данных, формат и имя, а затем нажмите execute (выполнить).

Как видите, в ячейку был добавлен дополнительный код.

```
import pandas as pd
df = pd.read_csv(r'/Users/ismaeltrevi/Desktop/all_games.csv', sep=',', decimal='.')
df
```

executed in 352ms, finished 14:09:39 2021-10-30

Show static HTML



History

Export

☒ Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows x 6 columns - preview

[All columns]

Update

	o name	o platform	o release_date	o summary	i meta_score	o user
0	The Legend of Z...	Nintendo 64	November 23, 19...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	September 20, 2...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	April 29, 2008	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	September 8, 1999	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	April 29, 2008	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	November 12, 20...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	May 23, 2010	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	October 26, 2018	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	November 18, 20...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	September 17, 2...	Los Santos is a v...	97	8.3

Удаление столбцов

Если какой-то столбец вам больше не нужен, просто введите drop (удалить) в поле поиска, выберите drop, укажите столбец, который хотите убрать, и нажмите execute.

Разделение строки

Предположим, вам нужно разделить столбец со списком людей на два столбца.

Просто введите split (разделить) в поле поиска, выберите столбец, который нужно разделить, separator (разделитель) и максимальное количество необходимых столбцов - всё готово!

```
import pandas as pd
df = pd.read_csv(r'/Users/ismaeltrevi/Desktop/all_games.csv', sep=',', decimal='.')
df['release_date'] = pd.to_datetime(df['release_date'], infer_datetime_format=True)
df['score'] = df['meta_score'].astype('string')
df['score'] = pd.to_numeric(df['meta_score'], downcast='float', errors='coerce')
df = df.drop(columns=['score'])
df = df.rename(columns={'meta_score': 'score'})
df['release_date'] = pd.to_datetime(df['release_date'], format='%Y-%m-%d')
df['release_date'] = pd.to_datetime(df['release_date'], format='%Y-%m-%d')
df['release_date'] = pd.to_datetime(df['release_date'], format='%YY-%m-%dd')
df
```

executed in 488ms, finished 14:24:44 2021-10-30

Show static HTML



History

Export



Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows x 6 columns - preview

[All columns]

Update

	o name	o platform	d release_date	o summary	i score	o user
0	The Legend of Z...	Nintendo 64	1998-11-23 00:0...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	2000-09-20 00:0...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	2008-04-29 00:0...	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	1999-09-08 00:0...	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	2008-04-29 00:0...	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	2007-11-12 00:0...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	2010-05-08 00:0...	Super Mario Gal...	97	8.4

Создание нового столбца с другим типом данных и именем

Если вам понадобится новый столбец с другим типом данных и именем, создавать заново ничего не придется. Просто кликните на тип данных столбца, выберите новый формат и имя, а затем нажмите execute. Новый столбец сразу появится в наборе данных.

На изображении видно, что в выбранном столбце meta_score тип данных изменен на float и указано новое имя. В результате был создан новый столбец.

```
import pandas as pd
df = pd.read_csv(r'/Users/ismaeltrevi/Desktop/all_games.csv', sep=',', decimal='.')
df['release_date'] = pd.to_datetime(df['release_date'], infer_datetime_format=True)
df['score'] = df['meta_score'].astype('string')
df
```

executed in 352ms, finished 14:09:39 2021-10-30

Show static HTML



History

Export

☒ Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows x 7 columns - preview

[All columns]

Update

	o name	o platform	d release_date	o summary	i meta_score	o user
0	The Legend of Z...	Nintendo 64	1998-11-23 00:0...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	2000-09-20 00:0...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	2008-04-29 00:0...	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	1999-09-08 00:0...	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	2008-04-29 00:0...	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	2007-11-12 00:0...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	2010-05-23 00:0...	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	2018-10-26 00:0...	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	2014-11-18 00:0...	Grand Theft Auto...	97	7.9

- Если в наборе данных будет тип DateTime, Vamboolib также создаст графики, показывающие, как изменялись данные в течение времени.

Выбор столбцов




Теперь визуализируем несколько столбцов. Для этого выберем название игры, платформу и балл. Затем введем select (выбрать) в строке поиска, выберем нужные столбцы и нажмем execute.

В завершение всех этих шагов Vamboolib создаст код, который может использовать даже тот, у кого не установлена эта библиотека.

```
import pandas as pd
df = pd.read_csv(r'/Users/ismaeltrevi/Desktop/all_games.csv', sep=',', decimal='.')
df['release_date'] = pd.to_datetime(df['release_date'], infer_datetime_format=True)
df['score'] = df['meta_score'].astype('string')
df['score'] = pd.to_numeric(df['meta_score'], downcast='float', errors='coerce')
df = df.drop(columns=['score'])
df = df.rename(columns={'meta_score': 'score'})
df['release_date'] = pd.to_datetime(df['release_date'], format='%Y-%m-%d')
df['release_date'] = pd.to_datetime(df['release_date'], format='%Y-%m-%d')
df['release_date'] = pd.to_datetime(df['release_date'], format='%YY-%m-%dd')
split_df = df['name'].str.split('\ ', n=3, expand=True)
split_df.columns = ['name' + f"_{id}" for id in range(len(split_df.columns))]
df = pd.merge(df, split_df, how="left", left_index=True, right_index=True)
df = df.drop(columns=['name_0', 'name_1', 'name_2', 'name_3'])
df
```

executed in 488ms, finished 14:24:44 2021-10-30

Show static HTML



HistoryExport

☒ Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows x 6 columns - preview

[All columns]

Update

	o name	o platform	d release_date	o summary	i score	o user_review
	The Legend of Z...	Nintendo 64	1998-11-23 00:0...	As a young boy, ...	99	9.1
	Tony Hawk's Pro...	PlayStation	2000-09-20 00:0...	As most major p...	98	7.4
	Grand Theft Auto...	PlayStation 3	2008-04-29 00:0...	[Metacritic's 200...	98	7.7
	SoulCalibur	Dreamcast	1999-09-08 00:0...	This is a tale of s...	98	8.4
	Grand Theft Auto...	Xbox 360	2008-04-29 00:0...	[Metacritic's 200...	98	7.9
	Super Mario Gal...	Wii	2007-11-12 00:0...	[Metacritic's 200...	97	9.1
	Super Mario Gal...	Wii	2010-05-23 00:0...	Super Mario Gal...	97	9.1

Фильтрация данных

Если вам потребуется отфильтровать набор данных или создать новый с отфильтрованной информацией, найдите filter в строке поиска, выберите то, что нужно отфильтровать. Если нужно создать новый набор данных, укажите это, и нажмите execute.

df

executed in 224ms, finished 15:27:18 2021-10-30

Show static HTML

History

Export

☒ Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows x 6 columns - preview

[All columns]

Update

	o name	o platform	o release_date	o summary	i meta_score	f user
0	The Legend of Z...	Nintendo 64	November 23, 19...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	September 20, 2...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	April 29, 2008	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	September 8, 1999	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	April 29, 2008	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	November 12, 20...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	May 23, 2010	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	October 26, 2018	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	November 18, 20...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	September 17, 2...	Los Santos is a v...	97	8.3
10	Disco Elysium: T...	PC	March 30, 2021	Disco Elysium - ...	97	8.3




Слияние данных

Чтобы соединить два набора данных, выполните поиск по слову merge, выберите два набора данных, укажите тип соединения, выберите ключевой столбец, который нужно использовать для объединения наборов данных, и нажмите execute. Таким способом можно как создать новый набор данных, так и отредактировать текущий.

```
below_9 = df.loc[df['user_review'] <= 9]
below_9
```

executed in 224ms, finished 15:27:18 2021-10-30

Show static HTML



HistoryExport

☒ Live Code Export

Filter: removed 1,459 rows (7%)

Search transformations

or

Create plot

or

Explore DataFrame

17,341 rows x 6 columns - preview

[All columns]

Update

	o name	o platform	o release_date	o summary	i meta_score	f user
1	Tony Hawk's Pro...	PlayStation	September 20, 2...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	April 29, 2008	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	September 8, 1999	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	April 29, 2008	[Metacritic's 200...	98	7.9
7	Red Dead Rede...	Xbox One	October 26, 2018	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	November 18, 20...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	September 17, 2...	Los Santos is a v...	97	8.3
10	Disco Elysium: T...	PC	March 30, 2021	Disco Elysium - ...	97	8.3
11	Grand Theft Auto V	Xbox 360	September 17, 2...	Los Santos is a s...	97	8.3
12	Tony Hawk's Pro...	Dreamcast	November 6, 2000	Hawk's back - wi...	97	6.2
13	The Legend of Z...	Switch	March 3, 2017	Forget everythin...	97	8.7




Извлечение datetime-атрибутов

А что если нужно извлечь строку, к примеру день недели и месяц, из столбца даты? Нужно ли для этого знать код или искать его в Google? С Vamboolib не понадобится ни то, ни другое. Просто найдите свойство `extract datetime`, выберите столбец `date` и укажите то, что хотите извлечь.

```
df['release_date'] = pd.to_datetime(df['release_date'], infer_datetime_format=True)
df
```

executed in 227ms, finished 15:30:38 2021-10-30

Show static HTML

   History Export ☒ Live Code Export

Search transformations or Create plot or Explore DataFrame

18,800 rows x 8 columns - preview [All columns] Update

	o name	o platform	d release_date	o summary	i meta_score	o user
0	The Legend of Z...	Nintendo 64	1998-11-23 00:0...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	2000-09-20 00:0...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	2008-04-29 00:0...	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	1999-09-08 00:0...	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	2008-04-29 00:0...	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	2007-11-12 00:0...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	2010-05-23 00:0...	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	2018-10-26 00:0...	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	2014-11-18 00:0...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	2013-09-17 00:0...	Los Santos is a v...	97	8.3
10	Disco Elysium: T...	PC	2021-03-30 00:0...	Disco Elysium - ...	97	8.3

Grouping By (группировка)

Group by — одна из самых полезных операций, которую можно выполнить с помощью Pandas. Однако иногда она может оказаться довольно сложной. К счастью, Vamboolib упрощает эту задачу. Найдите Group by в поле поиска, выберите столбцы, которые требуется сгруппировать, а затем вычисления, которые хотите произвести.

```
df['user_review'] = pd.to_numeric(df['user_review'], downcast='integer', errors='coerce')
df
```

executed in 231ms, finished 15:38:33 2021-10-30

Show static HTML



History

Export



Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows × 6 columns - preview

[All columns]

Update

	o name	o platform	o release_date	o summary	i meta_score	f user
0	The Legend of Z...	Nintendo 64	November 23, 19...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	September 20, 2...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	April 29, 2008	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	September 8, 1999	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	April 29, 2008	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	November 12, 20...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	May 23, 2010	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	October 26, 2018	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	November 18, 20...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	September 17, 2...	Los Santos is a v...	97	8.3
10	Disco Elysium: T...	PC	March 30, 2021	Disco Elysium - ...	97	8.3

Простое исследование данных

- Bamboolib отлично помогает для исследовательского анализа данных. Теперь исследование данных является неотъемлемой частью любого процесса обработки данных. А написание всего кода для исследования данных и создания всех диаграмм сложно и требует большого терпения и усилий, чтобы все получилось правильно.

```
bam.show(train)
```

executed in 318ms, finished 17:10:21 2019-12-14

Pro Tip: save the time writing `bam.show(df)` and just write `df`. Enable this feature via `bam.enable()` and disable via `bam.disable()`

Add Python Code Select/Drop columns Filter Sort New column formula Aggregate/Group by Join Replace value

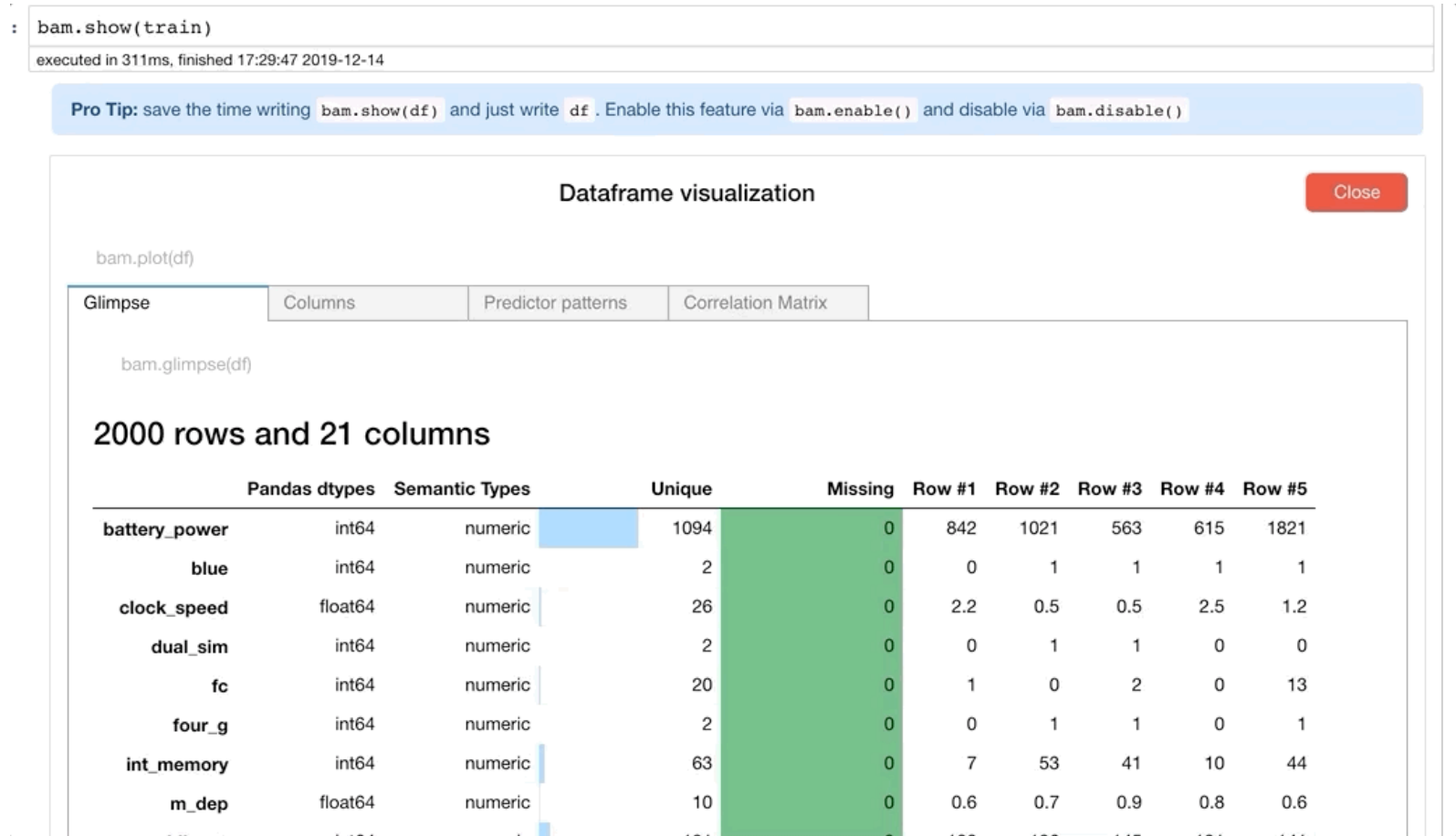
Set/Update values OneHotEncoder Pivot Unpivot/Melt **Visualize DataFrame**

2000 rows x 21 columns

	i ram	i sc_h	i sc_w	i talk_time	i three_g	i touch_screen	i wifi	i price_range
	3826.0	14.0	9.0	13.0	1.0	1.0	1.0	3.0
	1482.0	18.0	0.0	2.0	1.0	0.0	0.0	1.0
	2680.0	7.0	1.0	4.0	1.0	0.0	1.0	2.0
	373.0	14.0	9.0	3.0	1.0	0.0	1.0	0.0
	568.0	17.0	15.0	11.0	1.0	1.0	1.0	0.0
	3554.0	10.0	9.0	19.0	1.0	0.0	1.0	3.0
	3752.0	10.0	2.0	18.0	1.0	1.0	0.0	3.0
	1835.0	19.0	13.0	16.0	1.0	1.0	0.0	1.0
	2337.0	11.0	1.0	18.0	0.0	1.0	1.0	1.0
	2819.0	17.0	15.0	3.0	1.0	1.0	0.0	3.0
	3283.0	17.0	1.0	15.0	1.0	0.0	0.0	3.0

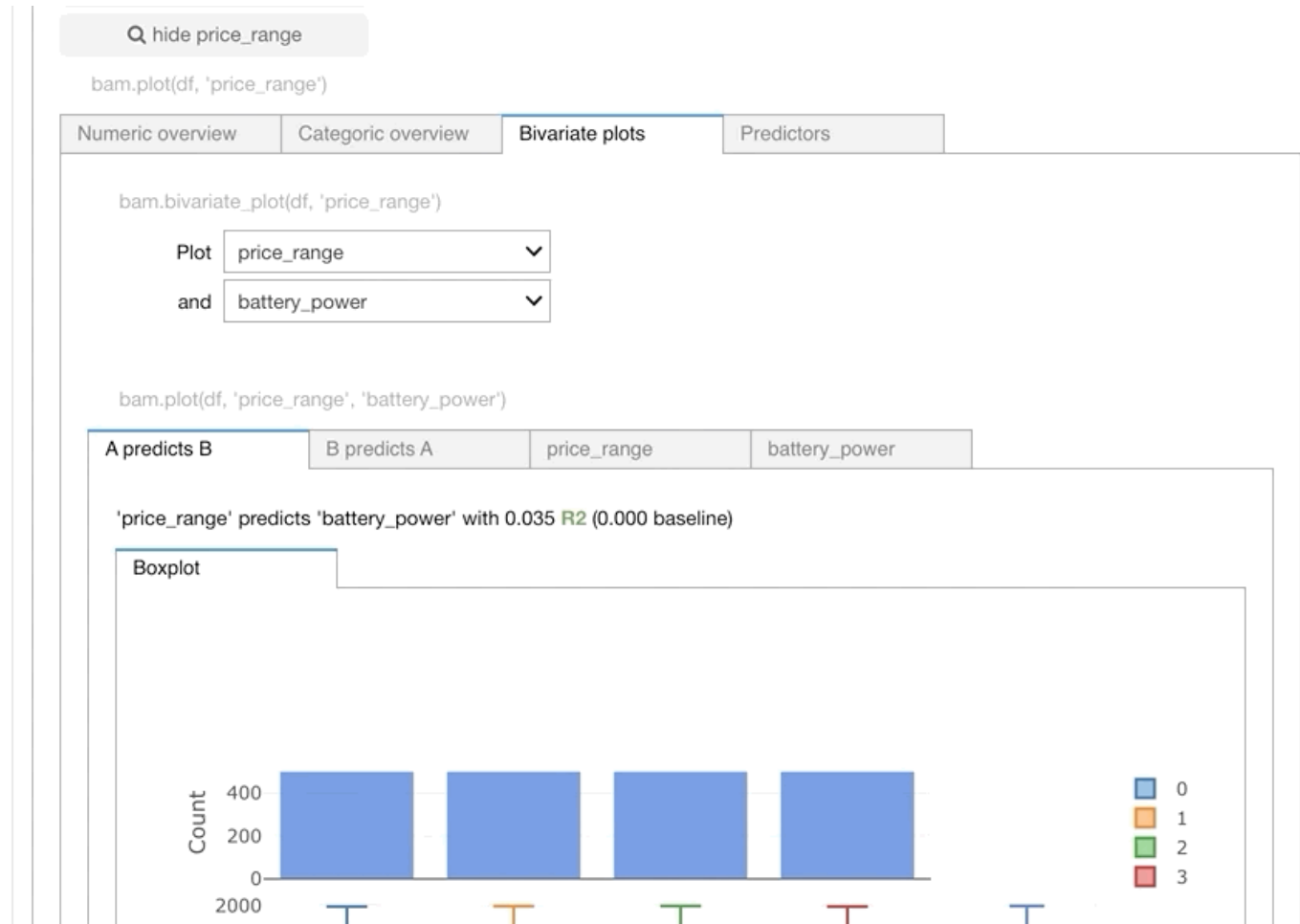
- Bamboolib делает всю работу по исследованию данных легкой.
- Например: **краткий обзор данных**, как только вы нажмете «Visualize Dataframe».

- можно увидеть отсутствующие значения в каждом столбце, а также количество уникальных значений и несколько экземпляров.
- также можем получить одномерную статистику и информацию на уровне столбца.
- ***Здесь мы получаем информацию о нашей целевой переменной — Price Range (ценовой диапазон)***



- мы наблюдаем одномерную статистику столбца, а также наиболее важные предикторы для нашего целевого столбца.

- Получение таких красивых графиков с помощью стандартных библиотек Python, таких как seaborn или plotly, обычно требует некоторого количества кода. Хотя [plotly express](#) очень помогает в этом, предоставляя простые функции для большинства диаграмм, Bamboolib автоматически создает для нас много важных диаграмм.
- Вы можете сделать это для каждой переменной в вашем наборе данных и попытаться получить представление о ваших данных.



Визуализация данных

- Bamboolib — отличный инструмент для быстрой визуализации данных. Например, чтобы создать гистограмму, достаточно нажать на create plot (создать график), выбрать тип фигуры и ось x - диаграмма готова!

```
import pandas as pd
df = pd.read_csv(r'/Users/ismael/Downloads/all_games.csv', sep=',', decimal='.')
df
```

Show static HTML



History

Export

☒ Live Code Export

Search transformations

or

Create plot

or

Explore DataFrame

18,800 rows × 6 columns - preview

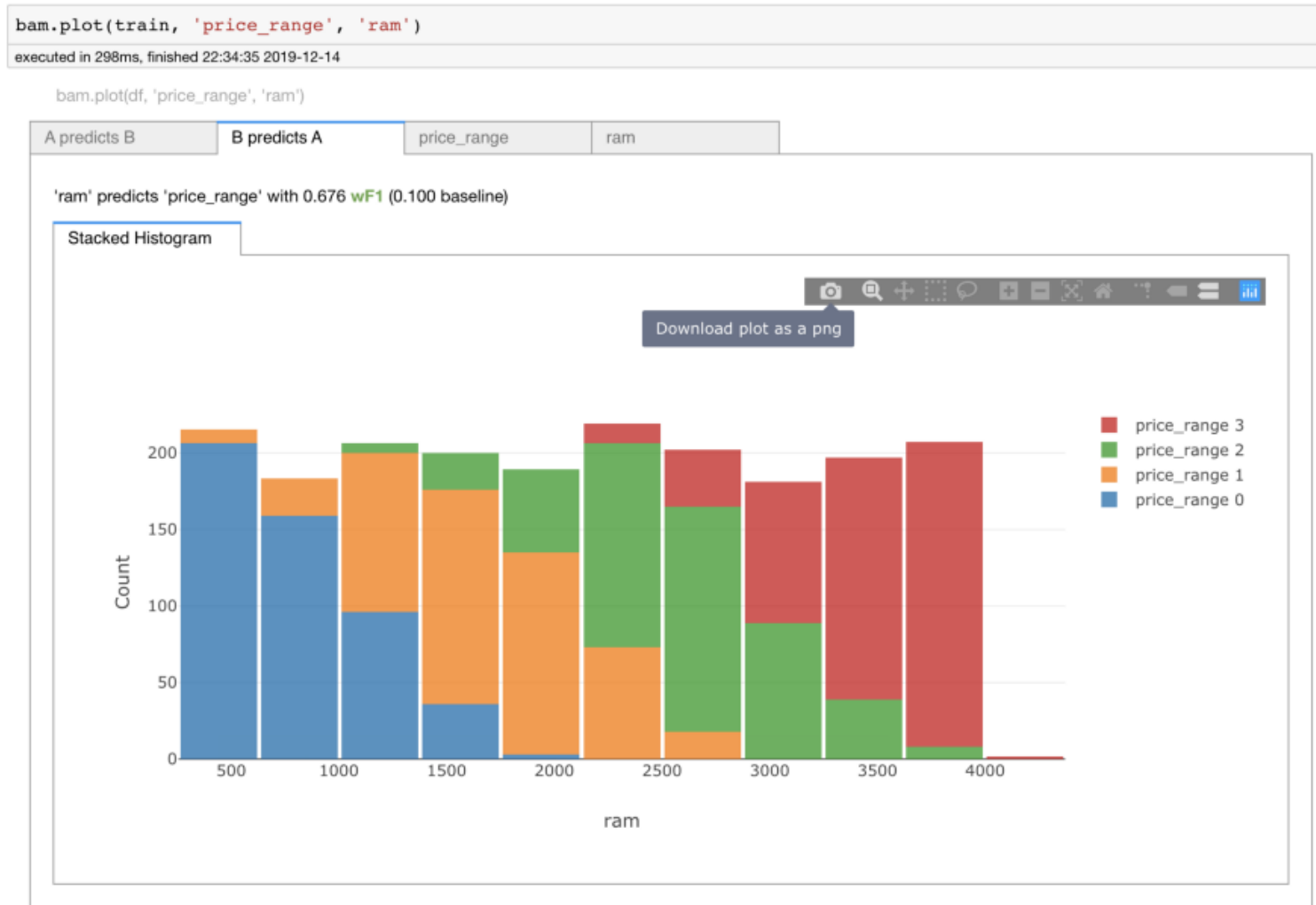
[All columns]

Update

	o name	o platform	o release_date	o summary	i meta_score	o user_
0	The Legend of Z...	Nintendo 64	November 23, 19...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	September 20, 2...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	April 29, 2008	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	September 8, 1999	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	April 29, 2008	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	November 12, 20...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	May 23, 2010	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	October 26, 2018	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	November 18, 20...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	September 17, 2...	Los Santos is a v...	97	8.3

Можно также экспортировать код этих диаграмм для использования в презентациях - экспортировать эти диаграммы в формате PNG.

Для этого просто скопируйте фрагмент кода, который показан над каждым графиком. Например, вы можете скопировать и запустить код, чтобы увидеть `price_range` vs `ram`, и вы увидите возможность загрузить эти графики в формате PNG. В бэкэнде все они представляют собой `plotly graphs`.






```
bam.plot(train, 'price_range', 'ram')
```

Исследование данных

С Vamboolib процесс исследования данных очень прост. Можно получить информацию о любом датасете одним щелчком мыши. Для этого нажмите на Explore DataFrame, после чего Vamboolib выдаст сводную статистику со средним значением, медианой, квартилями, стандартным отклонением, количеством наблюдений, пропущенными значениями, количеством положительных и отрицательных наблюдений и многое другое. Библиотека также создает графики, отражающие распределение данных.

df

Show static HTML

   History Export ☒ Live Code Export

Search transformations or Create plot or Explore DataFrame

18,800 rows x 6 columns - preview [All columns] Update

	o name	o platform	d release_date	o summary	i meta_score	f user
0	The Legend of Z...	Nintendo 64	1998-11-23 00:0...	As a young boy, ...	99	9.1
1	Tony Hawk's Pro...	PlayStation	2000-09-20 00:0...	As most major p...	98	7.4
2	Grand Theft Auto...	PlayStation 3	2008-04-29 00:0...	[Metacritic's 200...	98	7.7
3	SoulCalibur	Dreamcast	1999-09-08 00:0...	This is a tale of s...	98	8.4
4	Grand Theft Auto...	Xbox 360	2008-04-29 00:0...	[Metacritic's 200...	98	7.9
5	Super Mario Gal...	Wii	2007-11-12 00:0...	[Metacritic's 200...	97	9.1
6	Super Mario Gal...	Wii	2010-05-23 00:0...	Super Mario Gal...	97	9.1
7	Red Dead Rede...	Xbox One	2018-10-26 00:0...	Developed by th...	97	8.0
8	Grand Theft Auto V	Xbox One	2014-11-18 00:0...	Grand Theft Auto...	97	7.9
9	Grand Theft Auto V	PlayStation 3	2013-09-17 00:0...	Los Santos is a v...	97	8.3
10	Disco Elysium: T...	PC	2021-03-30 00:0...	Disco Elysium - ...	97	8.3

- Если в наборе данных будет тип DateTime, Vamboolib также создаст графики, показывающие, как изменялись данные в течение времени.