

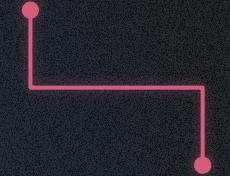
WEB scrapping and API

Подготовка данных для решения
задачи прогнозирования цены на
криптовалюты

Старицын Владислав 225
Исанбирдин Тимур 225
Анохин Андрей 225
Гуртовой Игнат 225



Ознакомиться с git можно здесь



План реализации проекта

01

API

Извлечение
данных путем
работы с API



02

Selenium

Извлечение
данных путем
работы скрапинга



03

EDA

Обработка данных
и финальные
выводы



API

Используя API сервиса Coingecko, мы получили базовую информацию о рынке криптовалют:

- 1) Капитализацию всего рынка
- 2) Топ 10 криптовалют по капитализации
- 3) Топ 10 криптовалют по объему торгов

Как видно, Bitcoin по капитализации составляет больше половины всего рынка, а на втором месте идет Ethereum

Top cryptocurrencies by market capitalization:

Bitcoin (Symbol: btc) - Market Cap: \$1.735.817.347.692
Ethereum (Symbol: eth) - Market Cap: \$264.415.706.312
Tether (Symbol: usdt) - Market Cap: \$142.461.570.678
XRP (Symbol: xrp) - Market Cap: \$141.748.421.596
BNB (Symbol: bnb) - Market Cap: \$86.443.054.792
Solana (Symbol: sol) - Market Cap: \$73.099.428.696
USDC (Symbol: usdc) - Market Cap: \$56.439.611.751
Cardano (Symbol: ada) - Market Cap: \$33.681.724.530
Dogecoin (Symbol: doge) - Market Cap: \$29.523.372.800
TRON (Symbol: trx) - Market Cap: \$20.837.466.641

Top cryptocurrencies by trading volume in the past 24 hours:

Tether (Symbol: usdt) - Volume: \$98.348.563.501
Bitcoin (Symbol: btc) - Volume: \$58.690.542.728
Ethereum (Symbol: eth) - Volume: \$29.613.558.541
USDC (Symbol: usdc) - Volume: \$12.274.448.905
First Digital USD (Symbol: fdusd) - Volume: \$9.791.838.236
XRP (Symbol: xrp) - Volume: \$8.173.700.456
Solana (Symbol: sol) - Volume: \$7.636.872.661
Cardano (Symbol: ada) - Volume: \$4.381.090.253
Dogecoin (Symbol: doge) - Volume: \$1.944.637.061
Wrapped SOL (Symbol: sol) - Volume: \$1.492.070.664

Total cryptocurrency market cap worldwide: \$2.982.810.747.849.9185

```
2025-03-05 10:07:40,143 - INFO - Starting to fetch top cryptocurrencies ranked by market capitalization.  
2025-03-05 10:07:40,567 - INFO - Successfully fetched data on top market cap cryptocurrencies.  
2025-03-05 10:07:40,567 - INFO - Starting to fetch the most traded cryptocurrencies over the past 24 hours.  
2025-03-05 10:07:41,137 - INFO - Successfully fetched data on top volume cryptocurrencies.  
2025-03-05 10:07:41,138 - INFO - Fetching the total market capitalization for all cryptocurrencies.  
2025-03-05 10:07:41,456 - INFO - Successfully fetched global market cap data.
```

API

Следующий шаг – получение базовой информации о токенах (BTC, ETH) за последние 2 года: ценах открытия/закрытия/мин/макс и торговом объеме

На сервисе Coingecko эта информация доступна только по платной подписке. Поэтому мы нашли другой сервис, через API которого можно получить эту информацию –

[Cryptocompare.com](#).

Заметим, что сервис позволял получать данные лишь по 2000 строк одним запросом. Так как нам были необходимы данные за последние два года по двум криптовалютам (суммарно ~35к строк), мы реализовали для этого специальную функцию

```
2025-03-05 10:11:51,792 - INFO - Retrieved 16008 hours of data for BTC.  
2025-03-05 10:11:51,792 - INFO - Starting to fetch hourly crypto date.  
2025-03-05 10:11:56,134 - INFO - Successfully fetched data on hourly crypto date.  
2025-03-05 10:11:56,135 - INFO - Retrieved 17521 hours of data for BTC.  
2025-03-05 10:11:56,135 - INFO - Successfully searched all data for our period  
2025-03-05 10:11:56,135 - INFO - Starting search all data for our period  
2025-03-05 10:11:56,135 - INFO - Starting to fetch hourly crypto date.  
2025-03-05 10:11:56,649 - INFO - Successfully fetched data on hourly crypto date.  
2025-03-05 10:11:56,650 - INFO - Retrieved 2001 hours of data for ETH.  
2025-03-05 10:11:56,650 - INFO - Starting to fetch hourly crypto date.  
2025-03-05 10:11:57,319 - INFO - Successfully fetched data on hourly crypto date.  
2025-03-05 10:11:57,320 - INFO - Retrieved 4002 hours of data for ETH.  
2025-03-05 10:11:57,320 - INFO - Starting to fetch hourly crypto date.  
2025-03-05 10:11:57,861 - INFO - Successfully fetched data on hourly crypto date.  
2025-03-05 10:11:57,862 - INFO - Retrieved 6003 hours of data for ETH.
```

	datetime	open	high	low	close	volume	name
0	2023-03-06 07:00:00	22408.75	22431.59	22396.28	22398.78	383.37	BTC
1	2023-03-06 08:00:00	22398.78	22451.00	22367.84	22418.25	826.90	BTC
2	2023-03-06 09:00:00	22418.25	22424.13	22374.98	22405.67	894.94	BTC
3	2023-03-06 10:00:00	22405.67	22414.12	22378.55	22404.66	724.60	BTC
4	2023-03-06 11:00:00	22404.66	22405.66	22367.11	22383.84	428.49	BTC

Selenium

Этапы

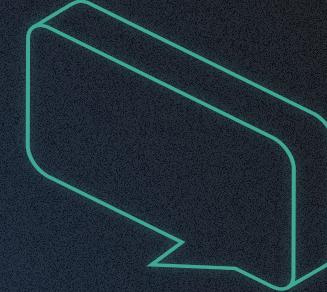


Инициализация и настройка браузера

Обработка всплывающих окон (баннеров с куками)

Навигация по странице и выполнение кликов

Скачивание и сохранение данных



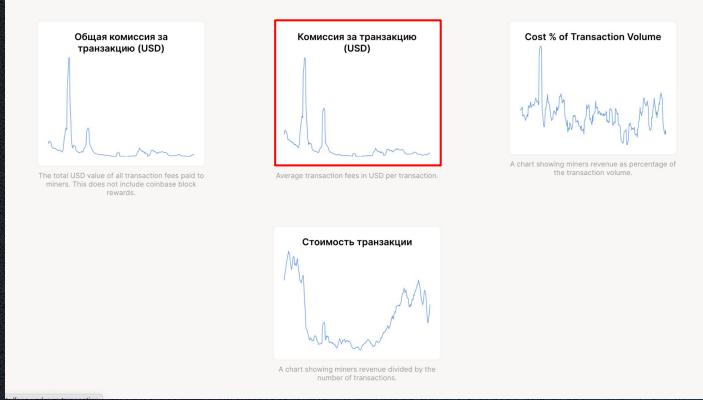
Код создаёт экземпляр браузера Chrome с заданными настройками. Это обеспечивает стабильную работу скрипта и автоматизированное сохранение скачиваемых JSON-файлов.

Скрипт ждёт появления кнопки с текстом «Принять» или «Принять все». Это гарантирует, что баннер не будет мешать дальнейшим действиям по клику на элементы страницы.

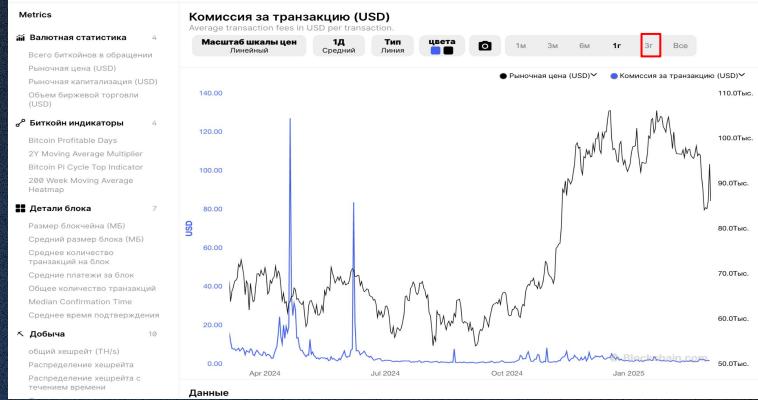
Скрипт переходит к выбору нужного временного периода и затем нажимает на кнопку «Скачать JSON». Это последовательное выполнение действий позволяет корректно взаимодействовать с динамически загружаемым содержимым страницы.

Скаченные JSON-файлы автоматически сохраняются в указанную папку загрузки, что позволяет в дальнейшем использовать их для анализа и построения EDA.

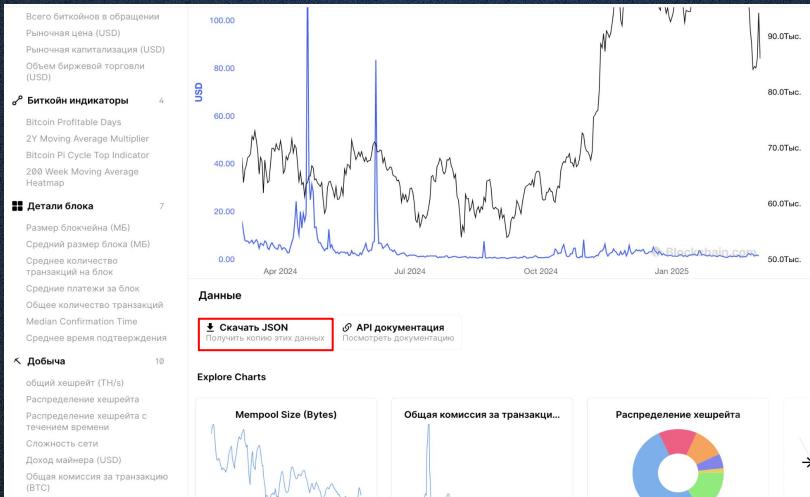
01



02



03



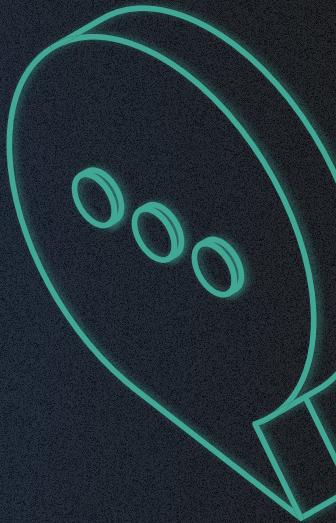
17521

Общая доля пропущенных значений 0.09%

Алгоритм борьбы с пропусками(метод локальной интерполяции):

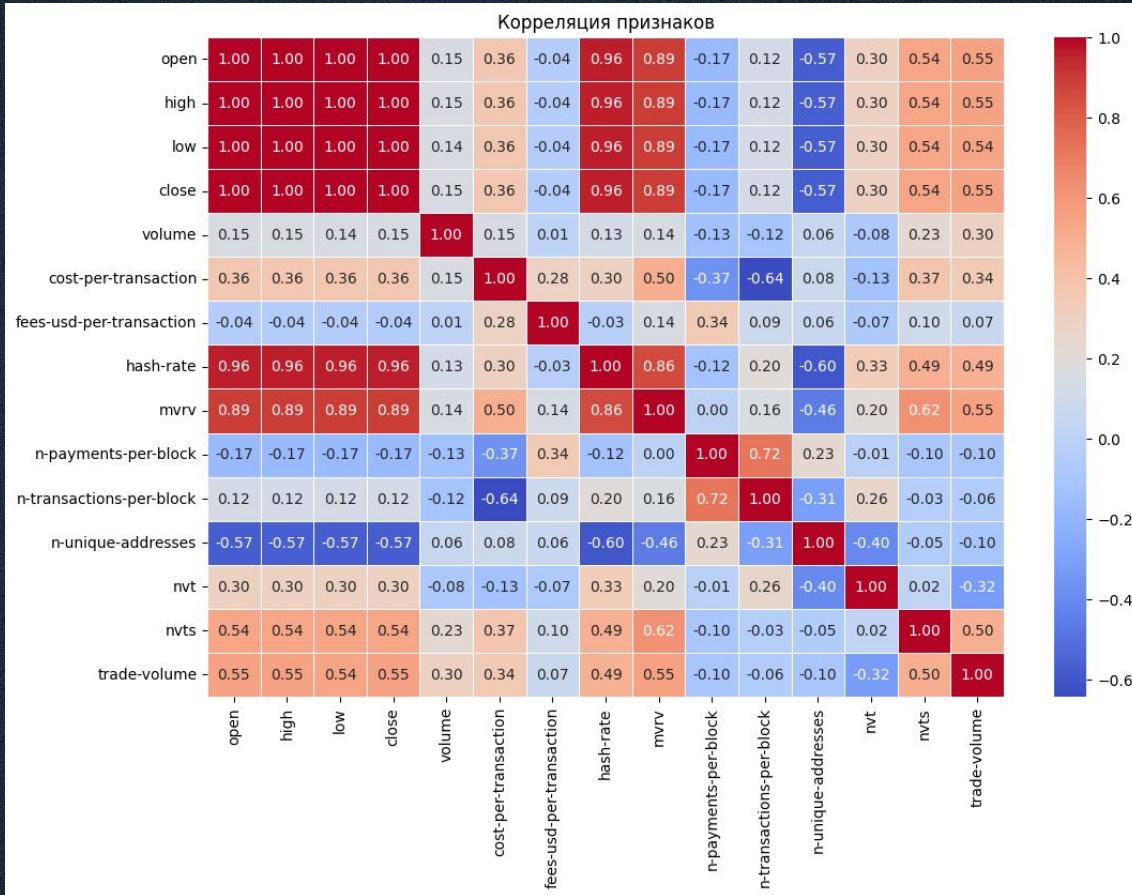
Заполняем блоки значением
(предыдущее значение + следующее значение) / 2

Для каждого непрерывного блока пропусков мы берем последнее известное значение перед блоком и первое известное значение после блока. Если вдруг один из краев отсутствует, то используется другое



EDA

Ценовые показатели сильно связаны друг с другом и имеют умеренную взаимосвязь с объёмами торгов, тогда как часть сетевых метрик движется в противофазе с ценой. Для построения прогностических моделей эти наблюдения помогут выбрать релевантные признаки, учесть мультиколлинеарность и, возможно, провести дополнительный анализ (например, взаимодействие сетевых метрик с ценой)



EDA

01

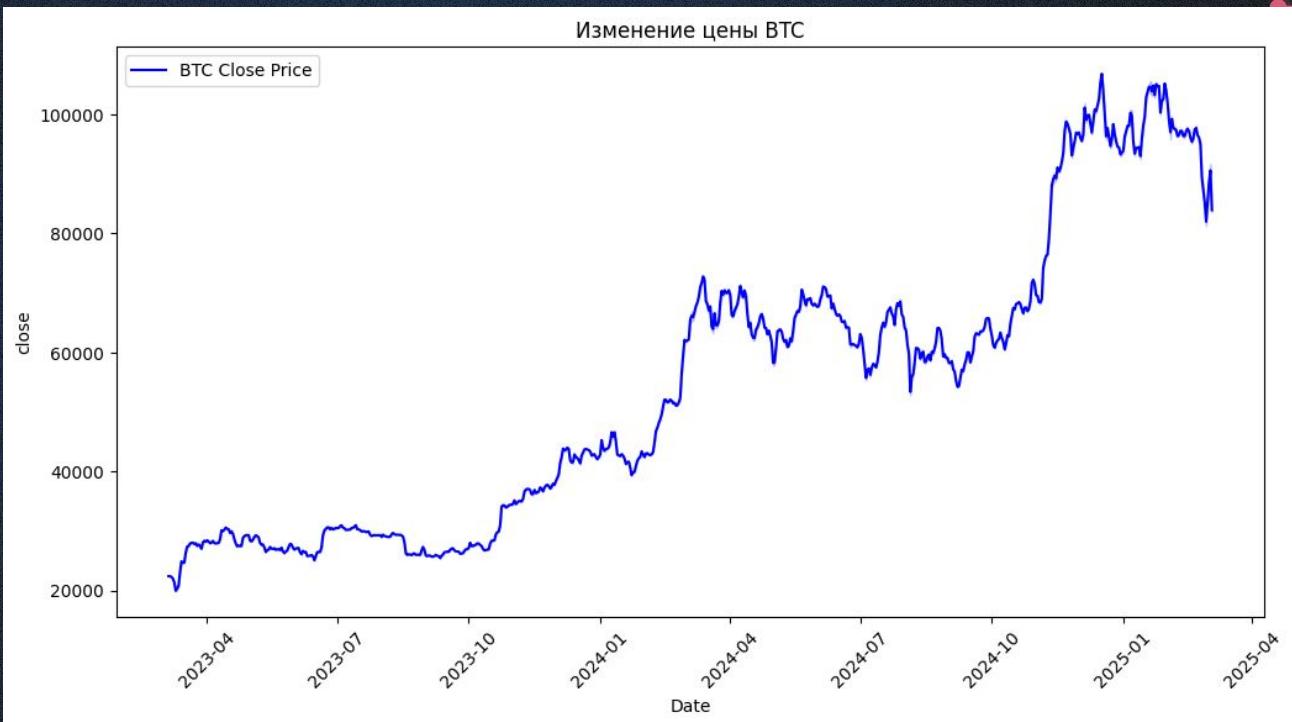
Рост с
коррекциями

02

Резкие
импульсные
движения

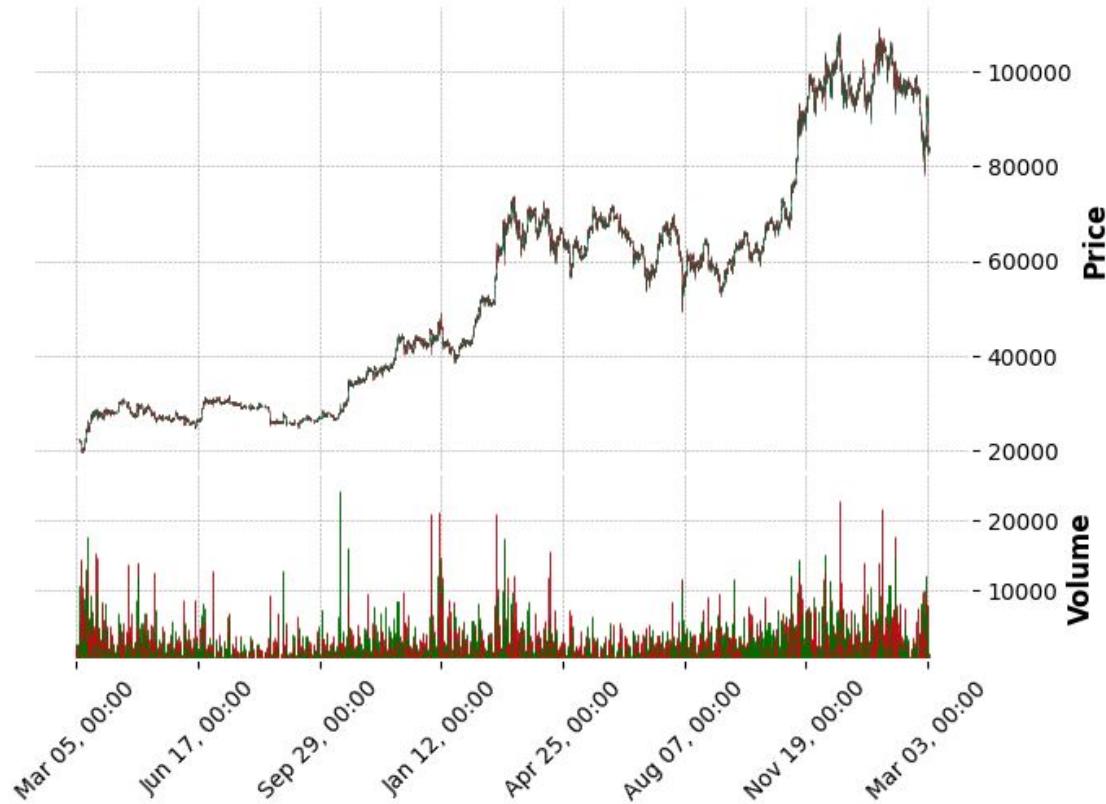
03

Долгосрочный
бычий тренд



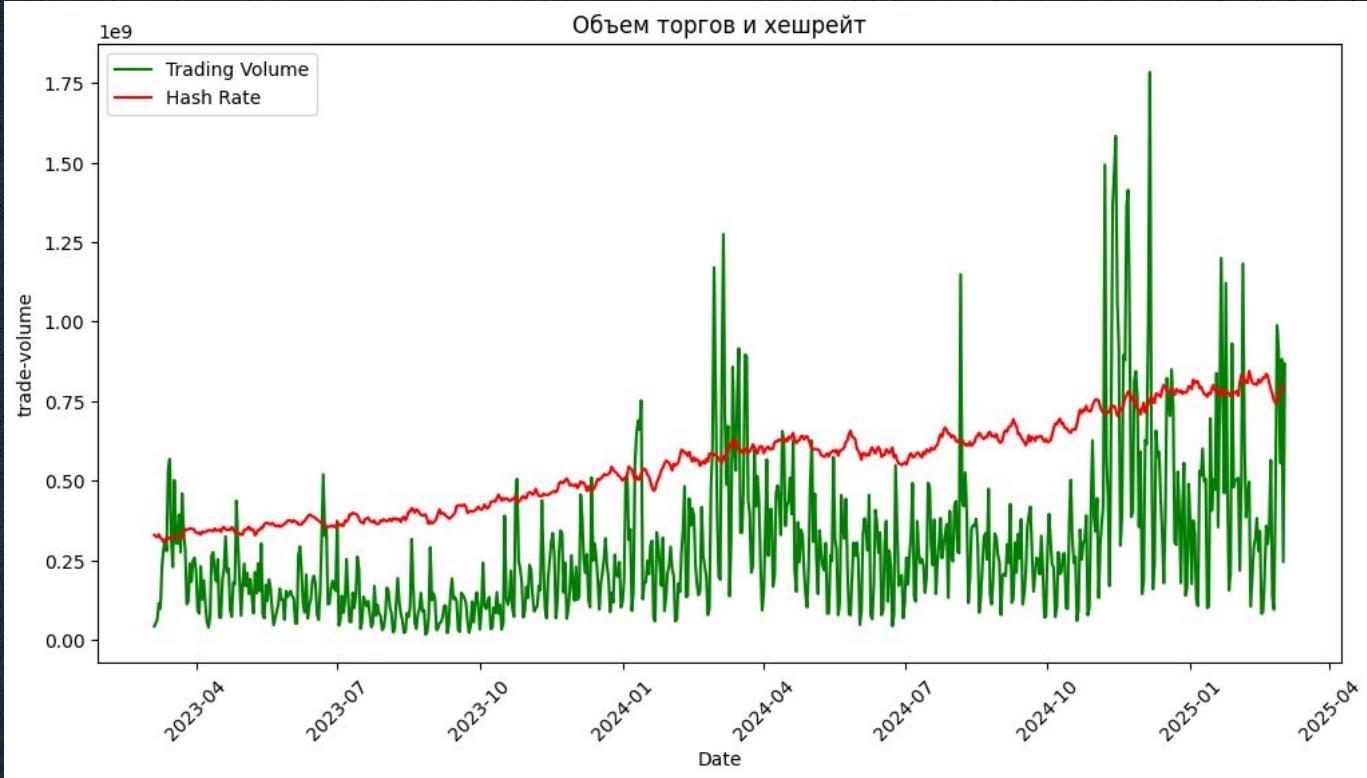
EDA

График японских свечей BTC



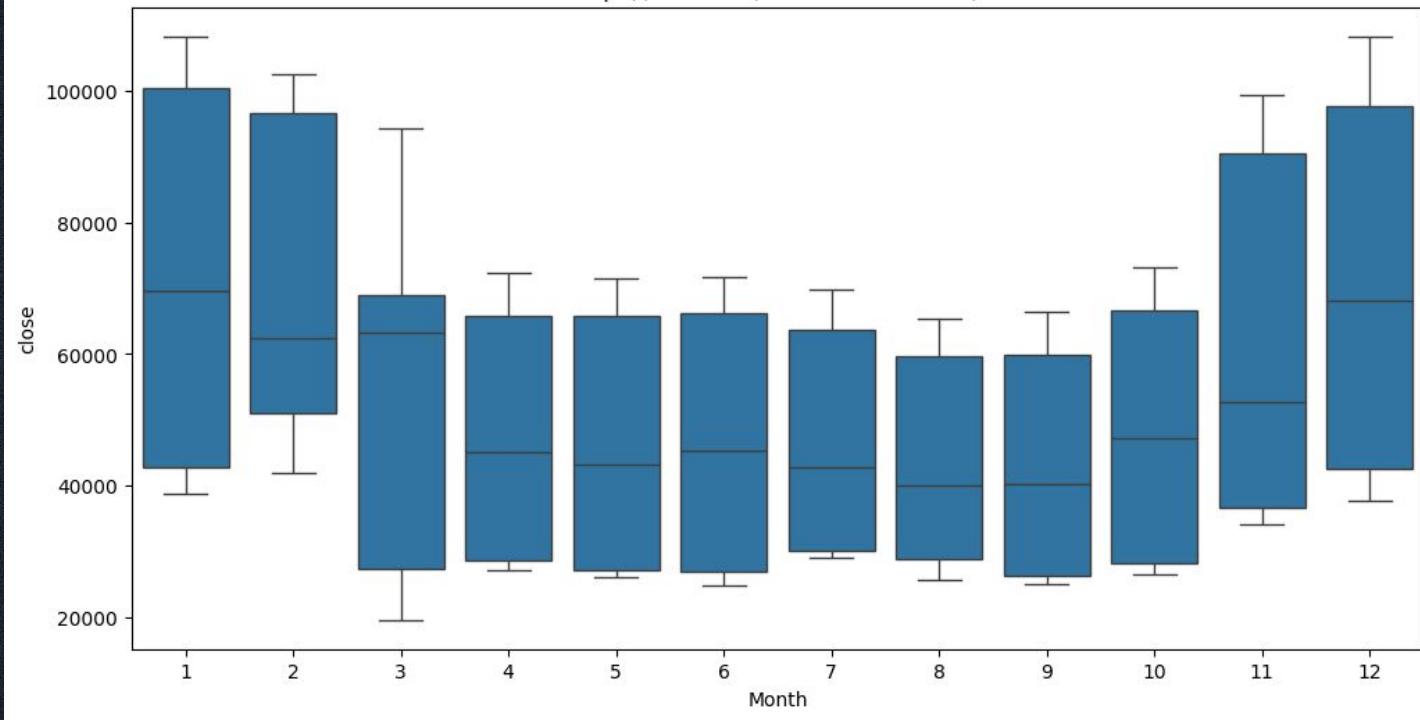
EDA

- Рост и волатильность объема торгов
- Рост хешрейта с небольшими скачками



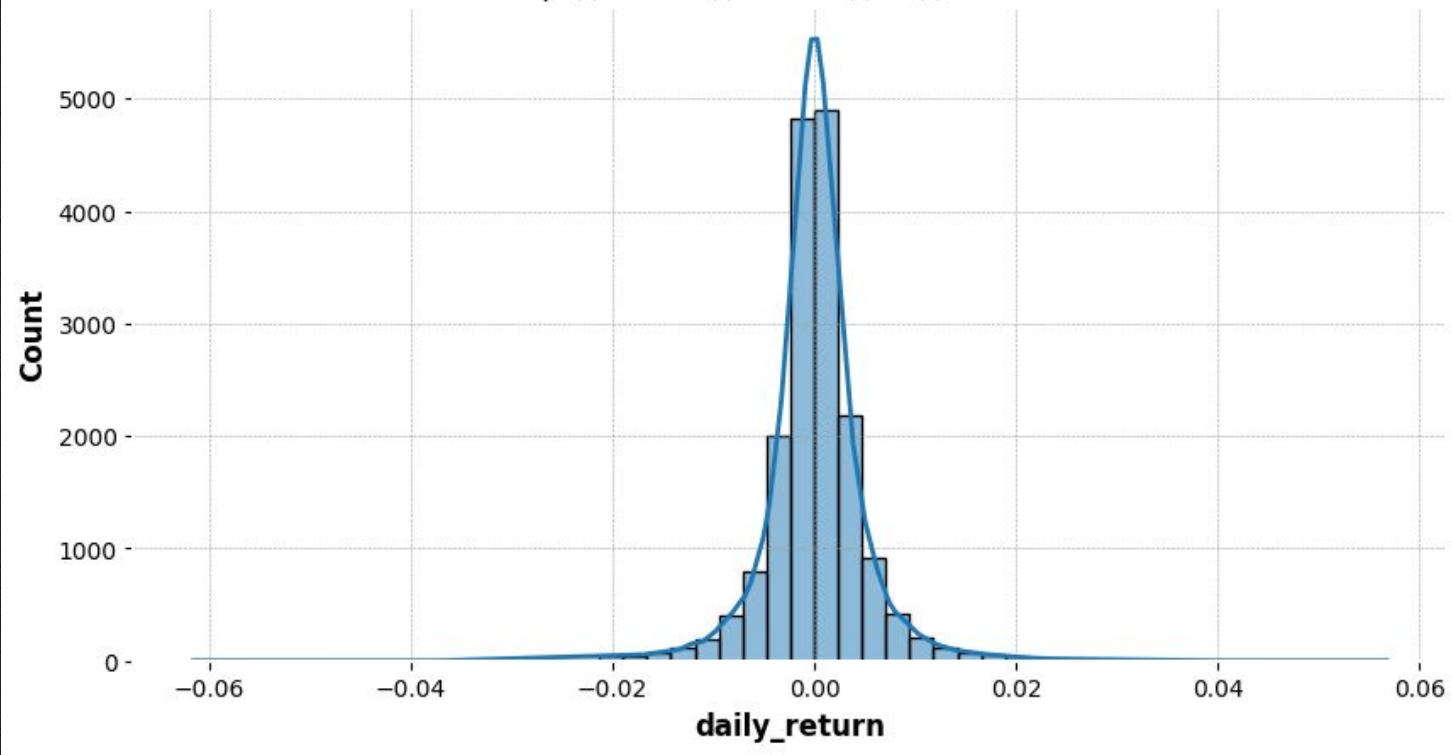
EDA

Распределение цены BTC по месяцам



EDA

Распределение дневных доходностей BTC



Выводы

① Автоматизированный сбор данных

Наш проект реализует механизм сбора данных, который автоматически собирает актуальные данные с ведущих источников

③ Работа с пропусками

Так как наш датасет представляет собой временной ряд, важно сохранить динамику изменения показателей. Для числовых признаков мы не заполняем пропуски глобальными значениями, а применяем метод локальной интерполяции

②

Качество и полнота данных

Мы предусмотрели обработку всплывающих окон, динамическое ожидание элементов и корректное сохранение скачанных JSON-файлов

④

Анализ взаимосвязей и трендов

С помощью собранных данных можно проводить глубокий анализ взаимосвязей между различными метриками блокчейна

⑤

Датасет для прогнозирования цены

В результате мы получили датасет, который в дальнейшем может быть использован для прогнозирования цен на криптовалюту

Спасибо за внимание!