



# Аналитический отчет по результатам тестового задания в Inca Digital

## Содержание отчета

Содержание отчета

Этап 1 - Подготовка

Структура аналитического проекта

Исходные данные

Постановка задачи

Оценка критериев успеха

Тип задачи

Инструменты

Методы

Библиотеки

Понимание данных

**Описание переменных в наборах.**

Подготовка данных

Оценка качества данных

**Bitcoin large transactions**

The FinCEN Files

Этап 2 - Разведочный анализ данных

Анализ задачи

Разведочный анализ данных

Описательные статистики

Распределение данных

Слияние данных

Дополнительный анализ результатов

## Этап 1 - Подготовка

### Структура аналитического проекта

- Исходные данные

- Постановка задачи
- Оценка критериев успеха реализации проекта
- Тип задачи решаемой в проекте
- Используемые инструменты
- Библиотеки
- Понимание данных
- Подготовка и оценка данных
- Анализ задачи
- Разведочный анализ данных
- Выводы
- Сохранение результатов

## Исходные данные

### Тестовое задание

Careers - Inca Digital

👉 <https://inca.digital/careers/?form=data-analyst-challenge>



### Исходные данные для анализа


*Large Bitcoin blockchain transactions*

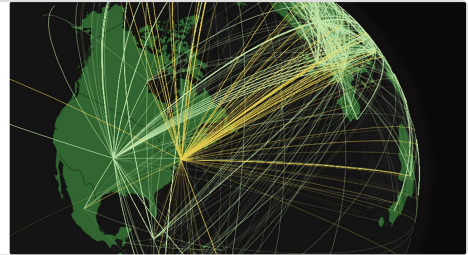
[https://inca.digital/files/Bitcoin-large-transactions-2015\\_2016\\_2017.zip](https://inca.digital/files/Bitcoin-large-transactions-2015_2016_2017.zip)

*FinCEN SAR dataset*

### Download FinCEN Files transaction data - ICIJ

The data in the FinCEN Files transactions map contains information on more than \$35 billion in transactions dated from 2000-2017 that were flagged by financial institutions as

 <https://www.icij.org/investigations/fincen-files/download-fincen-files-transaction-data/>



## Постановка задачи

- Match suspicious bank transactions to blockchain transaction hashes:

## Оценка критериев успеха

- Таблица в которой сопоставлены подозрительные банковские операции с хэшами блокчейн-транзакций

## Тип задачи

- Разведочный анализ данных

## Инструменты

- JupyterLab
- Python
- Notion
- SimpleMind Pro

## Методы

- |                               |  |
|-------------------------------|--|
| • <code>pd.read_csv</code>    | • <code>dtypes</code>                  |
| • <code>head().T</code>       | • <code>pd.concat</code>               |
| • <code>replace</code>        | • <code>pd.DataFrame.to_feather</code> |
| • <code>isnull()</code>       | • <code>pd.read_feather</code>         |
| • <code>duplicated()</code>   | • <code>info()</code>                  |
| • <code>shape</code>          | • <code>describe()</code>              |
| • <code>index</code>          | • <code>pd.merge</code>                |
| • <code>pd.to_datetime</code> | • <code>corr()</code>                  |
| • <code>.loc</code>           |  |

## Библиотеки

- pandas
- numpy
- seaborn
- matplotlib

## Понимание данных

Исходные данные представлены двумя наборами данных:

- Bitcoin large transactions
- The FinCEN Files

Набор Bitcoin large transactions содержит три файла о транзакциях за период с 2015 по 2017 годы.

Набор The FinCEN Files содержит два файла:

- Bank connections – характеристики банков участвующих в подозрительных операциях.
- Transactions map – характеристики переводов между банками.

## Описание переменных в наборах.

### Bitcoin large transactions

- **time** - Дата отправления
- **Hash** - Хэш
- **Sender** - Отправитель
- **Receiver** - Получатель
- **Transaction\_amount\_BTC** - Сумма транзакции BTC
- **Price** - Цена BTC в USD
- **Transaction\_amount\_USD** - Сумма транзакции USD

### The FinCEN Files

#### Bank connections

- **icij\_sar\_id** - id подозрительной транзакции
- **filer\_org\_name\_id** - Идентификатор регистрационного имени организации
- **filer\_org\_name** - Зарегистрированное имя организации
- **entity\_b\_id** - Идентификатор банка
- **entity\_b** - Название банка
- **entity\_b\_country** - Название страны
- **entity\_b\_iso\_code** - Международный код страны

### Transactions map

- **id** - Идентификатор
- **icij\_sar\_id** - Идентификатор подозрительной транзакции
- **filer\_org\_name\_id** - Идентификатор регистрационного имени организации
- **filer\_org\_name** - Зарегистрированное имя организации
- **begin\_date** - Дата начала транзакции
- **end\_date** - Дата завершения транзакции
- **originator\_bank\_id** - Идентификатор банка инициатора
- **originator\_bank** - Название банка инициатора
- **originator\_bank\_country** - Название страны банка инициатора
- **originator\_iso** - Международный код страны банка инициатора
- **beneficiary\_bank\_id** - Идентификатор банка получателя
- **beneficiary\_bank** - Название банка получателя
- **beneficiary\_bank\_country** - Название страны банка получателя
- **beneficiary\_iso** - Международный код страны банка получателя
- **number\_transactions** - Количество транзакций
- **amount\_transactions** - Сумма транзакций

### Подготовка данных

- По результатам предварительного анализа исходных данных выявлено, что данные достаточно хорошо подготовлены для дальнейшей обработки.
- В приведении названий к форме принятой в анализе нуждались только, данные набора Bitcoin large transactions.
- Выполнено преобразование передикторов time, begin\_date, end\_date к типу данных

datetime64[ns]

- Выполнено слияние таблиц набора Bitcoin large transactions в один набор по средствам конкатенации
- Результаты преобразований сохранены в формате feather для дальнейшего анализа

## Оценка качества данных

### Bitcoin large transactions

2015 год

Типы данных:

time	object
hash	object
sender	object
receiver	object
transaction_amount_btc	float64
price	float64
transaction_amount_usd	float64
dtype:	object

Общее количество пропущенных значений:  
0

Количество значений, отличных от NaN:  
59892

Дублирующих строк:  
77

Форма набора данных:  
(8556, 7)

Тип индекса набора данных:  
RangeIndex(start=0, stop=8556, step=1)

---

## 2016 год

Типы данных:

time	object
hash	object
sender	object
receiver	object
transaction_amount_btc	float64
price	float64
transaction_amount_usd	float64

dtype: object

---

Общее количество пропущенных значений:  
0

---

Количество значений, отличных от NaN:  
306719

---

Дублирующих строк:  
77

---

Форма набора данных:  
(43817, 7)

---

Тип индекса набора данных:  
RangeIndex(start=0, stop=43817, step=1)

---

## 2017 год

Типы данных:

time	object
hash	object
sender	object
receiver	object
transaction_amount_btc	float64
price	float64
transaction_amount_usd	float64

dtype: object

---

Общее количество пропущенных значений:  
0

---

Количество значений, отличных от NaN:  
69888

---

Дублирующих строк:  
310

---

Форма набора данных:  
(9984, 7)

---

Тип индекса набора данных:  
RangeIndex(start=0, stop=9984, step=1)

---

## The FinCEN Files

### Bank connections

Типы данных:

icij_sar_id	int64
filer_org_name_id	object
filer_org_name	object
entity_b_id	object
entity_b	object
entity_b_country	object
entity_b_iso_code	object
dtype:	object

---

Общее количество пропущенных значений:  
0

---

Количество значений, отличных от NaN:  
38486

---

Дублирующих строк:  
14

---

Форма набора данных:  
(5498, 7)

---

Тип индекса набора данных:  
RangeIndex(start=0, stop=5498, step=1)

---



## Transactions map

Типы данных:

id	int64
icij_sar_id	int64
filer_org_name_id	object
filer_org_name	object
begin_date	object
end_date	object
originator_bank_id	object
originator_bank	object
originator_bank_country	object
originator_iso	object
beneficiary_bank_id	object
beneficiary_bank	object
beneficiary_bank_country	object
beneficiary_iso	object
number_transactions	float64
amount_transactions	float64
dtype:	object

---

Общее количество пропущенных значений:  
123

---

Количество значений, отличных от NaN:  
71989

---

Дублирующих строк:  
0

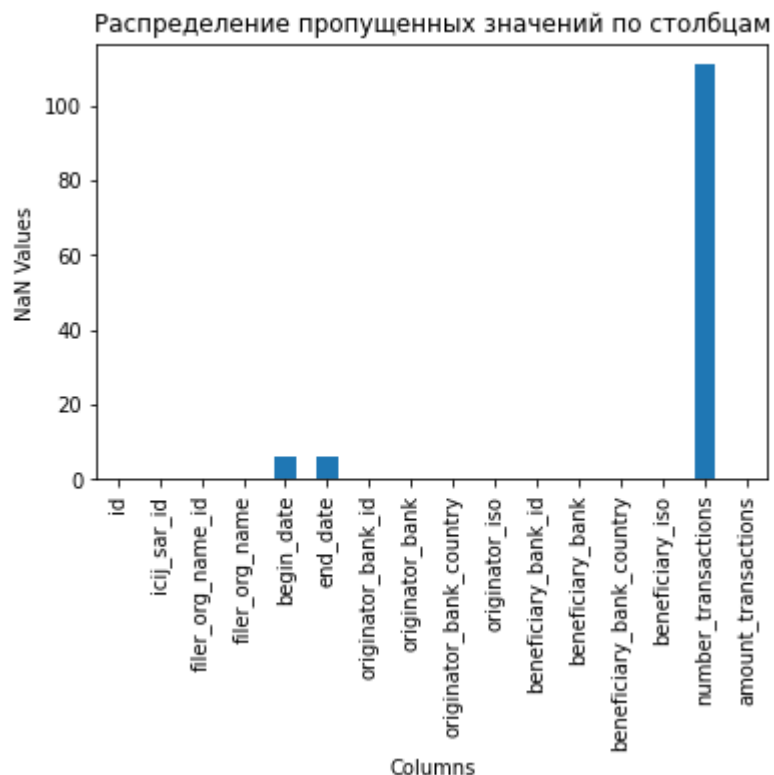
---

Форма набора данных:  
(4507, 16)

---

Тип индекса набора данных:  
RangeIndex(start=0, stop=4507, step=1)

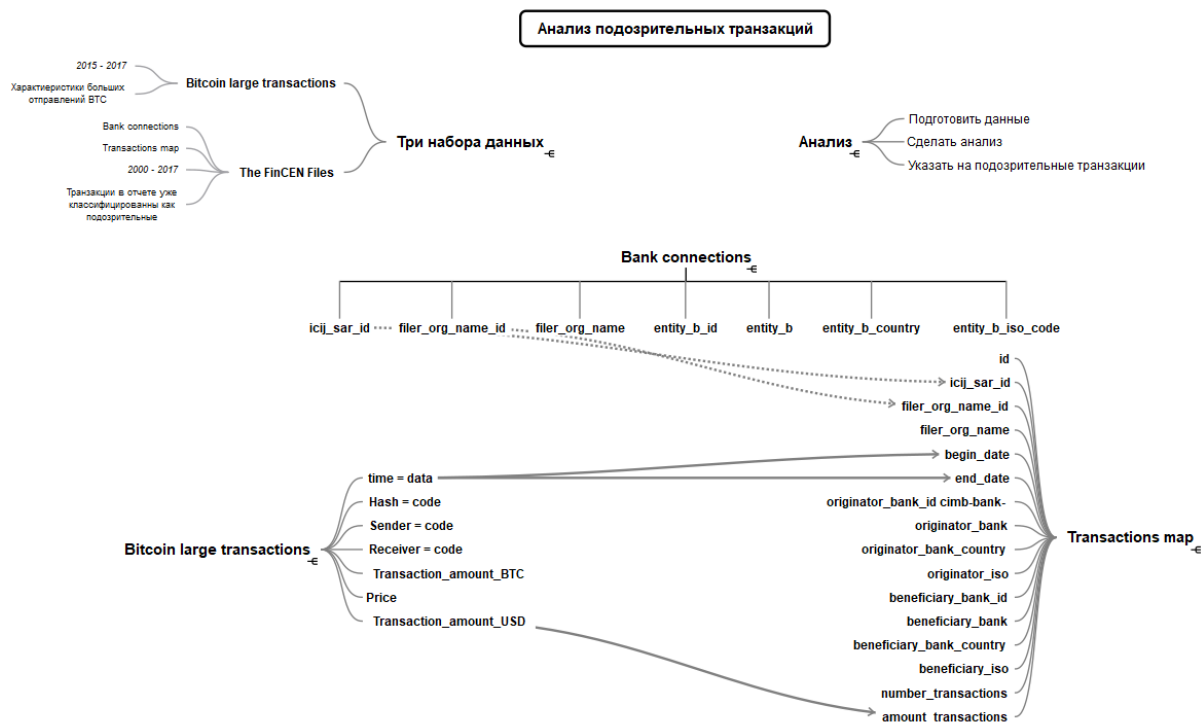
---



## Этап 2 - Разведочный анализ данных

### Анализ задачи

Используя диаграмму связей и инструмент для повышения эффективности мышления «Факт-карту» выполнен комплексный анализ задачи проекта.



На диаграмме стрелками указаны пересечения исходных наборов данных, по которым можно сделать слияние для достижения цели проекта.

## Разведочный анализ данных

### Описательные статистики

#### Bitcoin large transactions

	transaction_amount_btc	price	transaction_amount_usd
count	62357.000000	62357.000000	6.235700e+04
mean	4055.521536	579.962925	1.989889e+06
std	3471.662551	267.817411	1.887466e+06
min	729.399169	202.191840	1.000011e+06
25%	1940.666006	394.989312	1.248065e+06
50%	3516.977975	435.662526	1.575029e+06
75%	4967.371807	716.611568	2.155877e+06
max	172841.815707	1372.148194	6.140663e+07

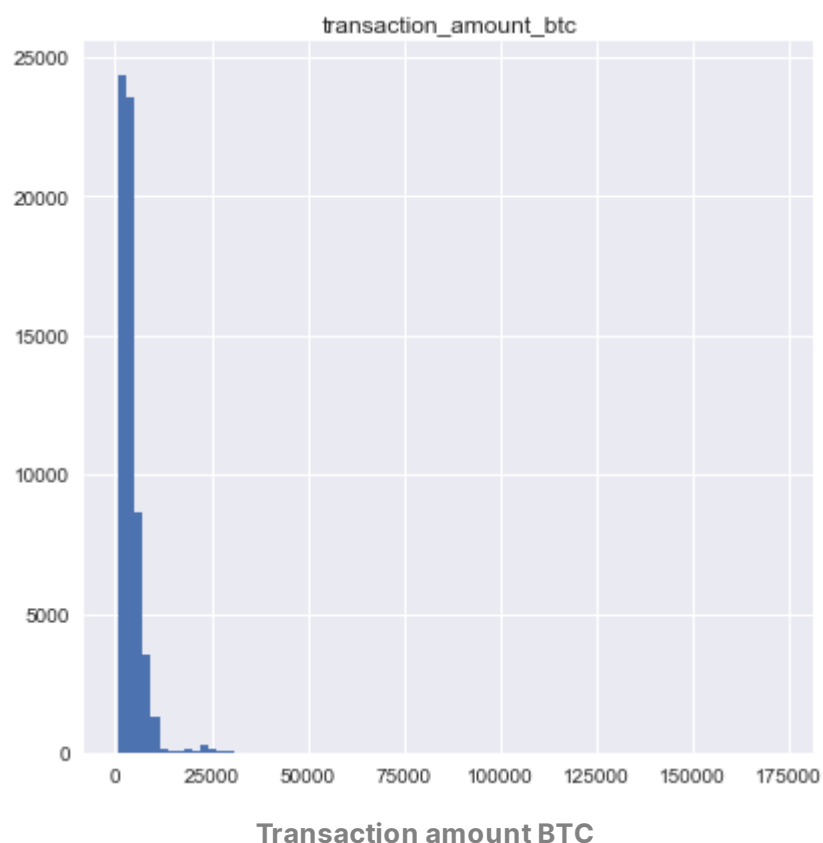
#### Transactions map

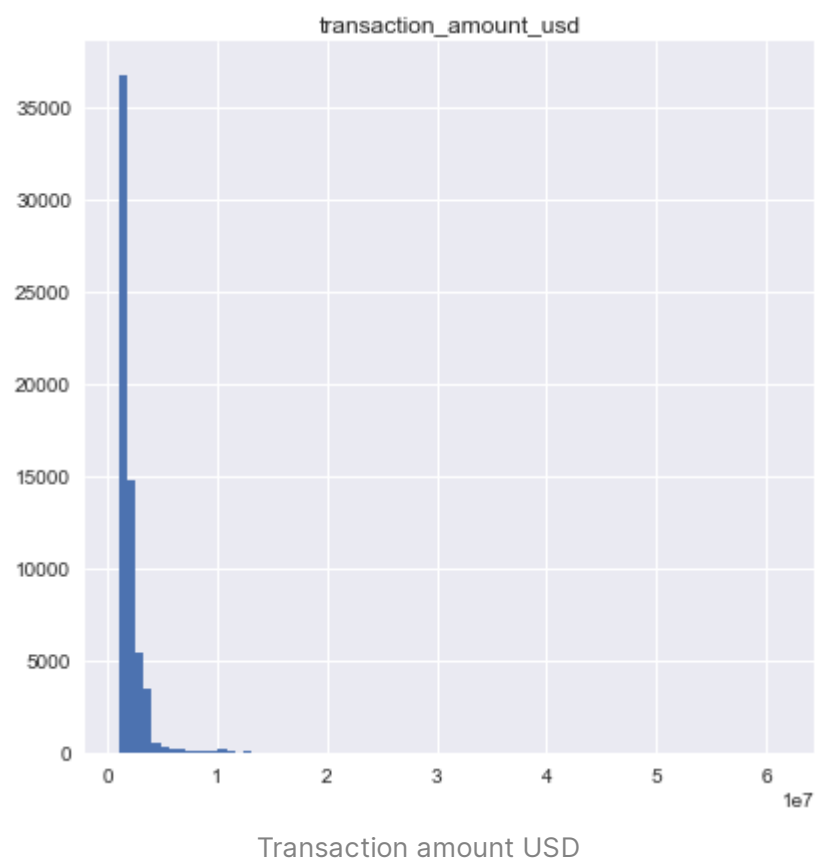
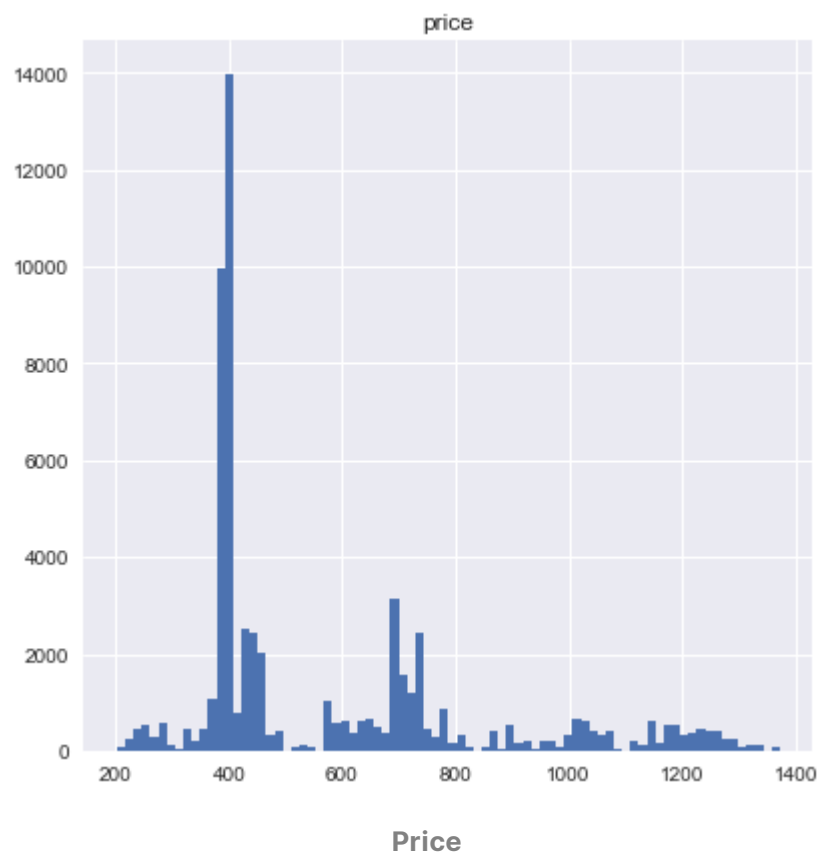
	id	icij_sar_id	number_transactions	amount_transactions
count	4507.000000	4507.000000	4396.000000	4.507000e+03
mean	233598.417351	3046.542933	4.129436	7.917073e+06
std	5836.150684	645.354556	9.892107	5.312478e+07
min	223254.000000	2208.000000	1.000000	1.180000e+00
25%	228068.500000	2441.000000	1.000000	6.704167e+04
50%	234944.000000	2905.000000	1.000000	4.950000e+05
75%	238380.000000	3461.000000	3.000000	2.813811e+06
max	243960.000000	4411.000000	174.000000	2.721000e+09

## Распределение данных

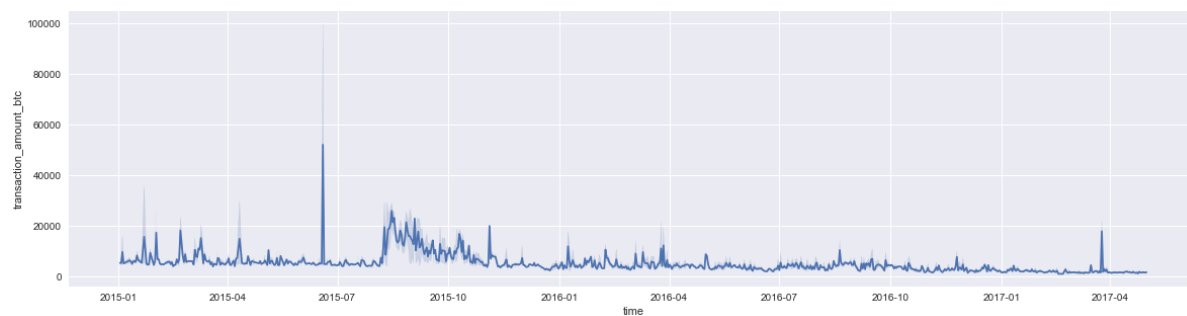
### Bitcoin large transactions

*Гистограммы распределения*

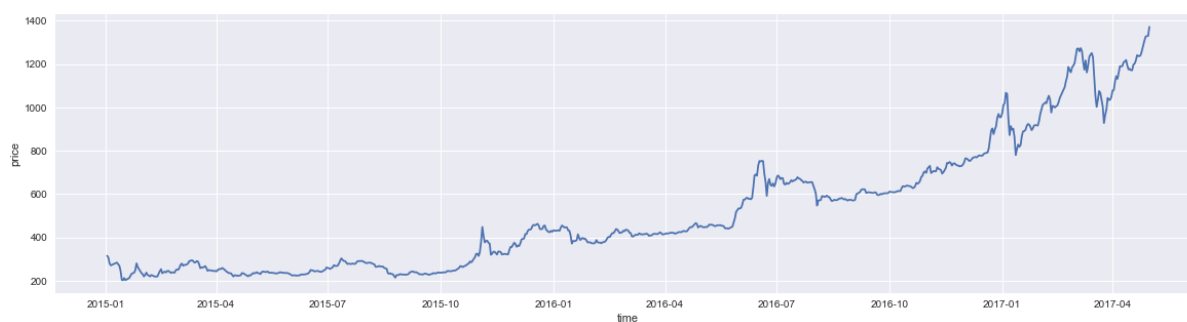




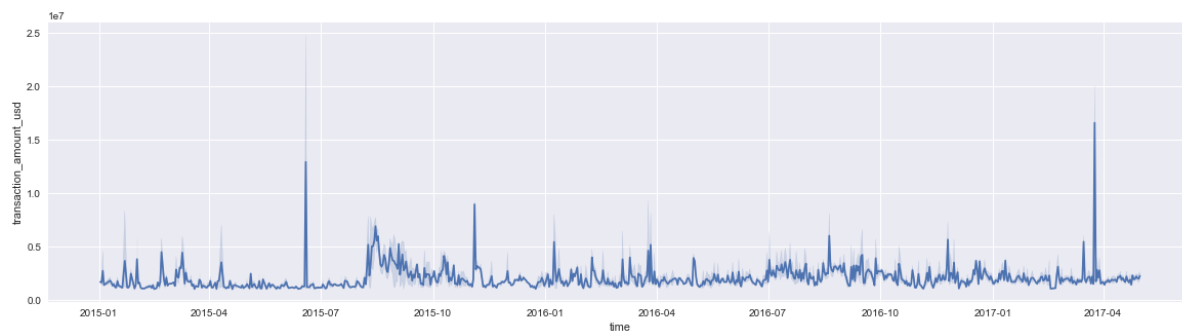
## Распределение по времени



Transaction amount BTC

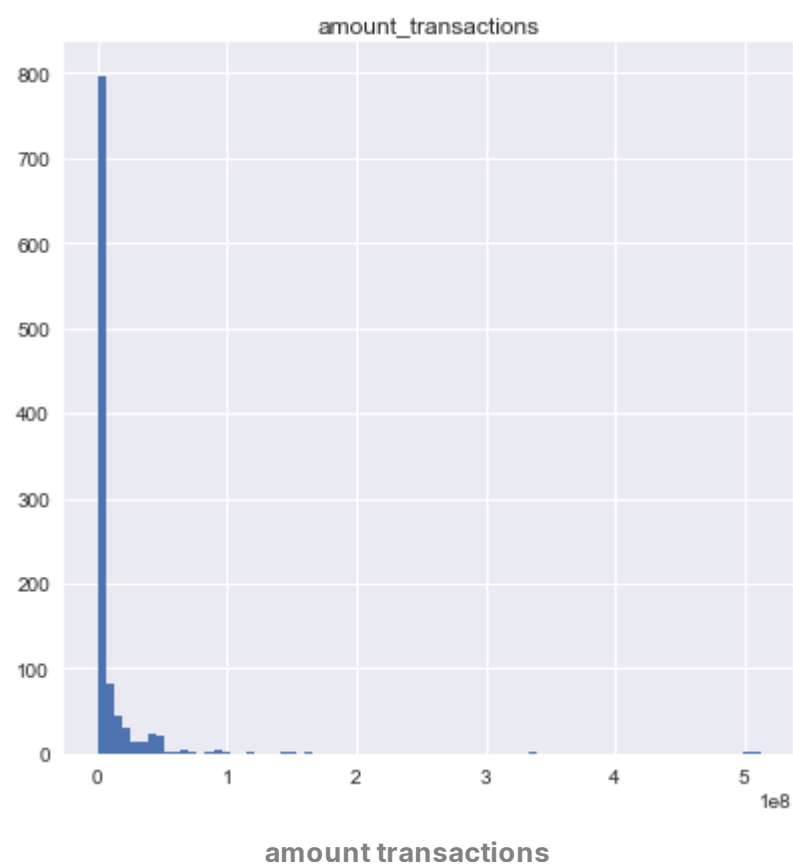
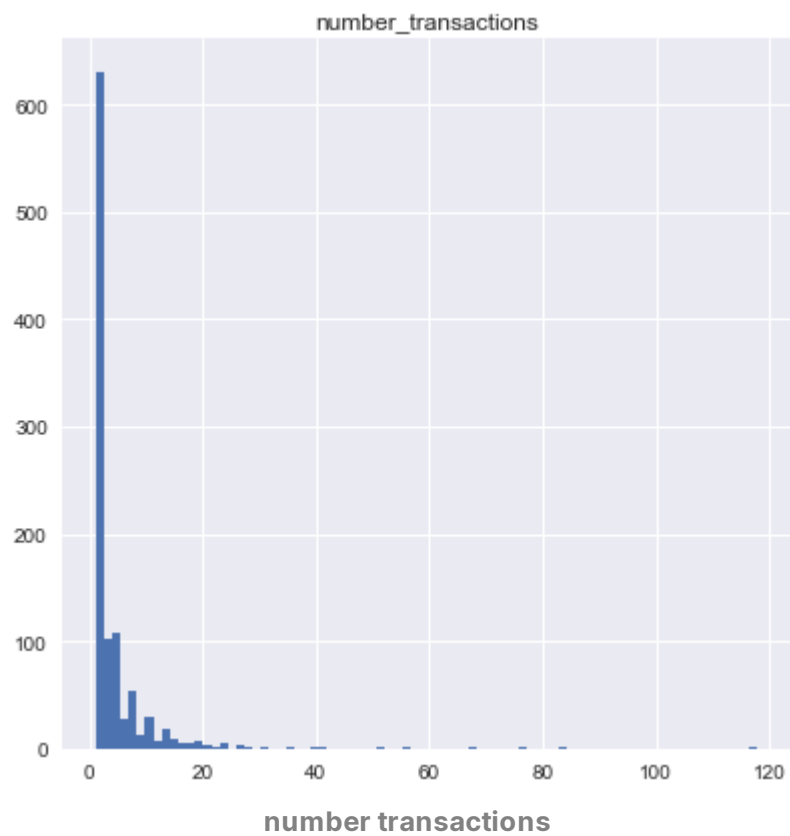


Price

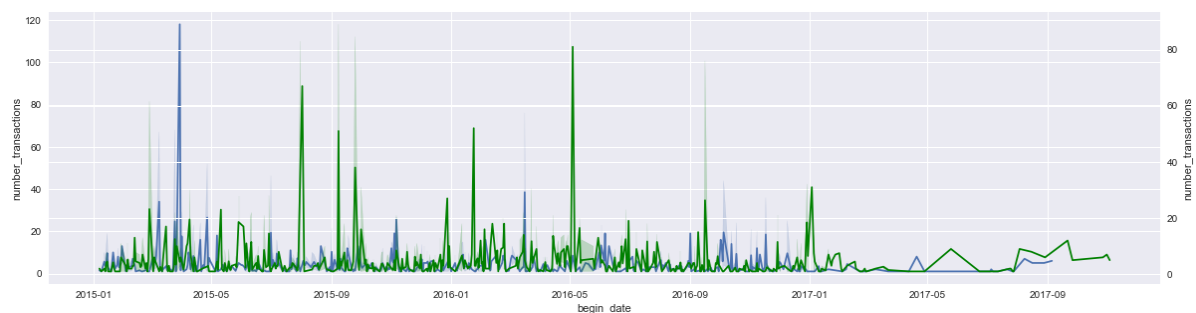


Transaction amount USD

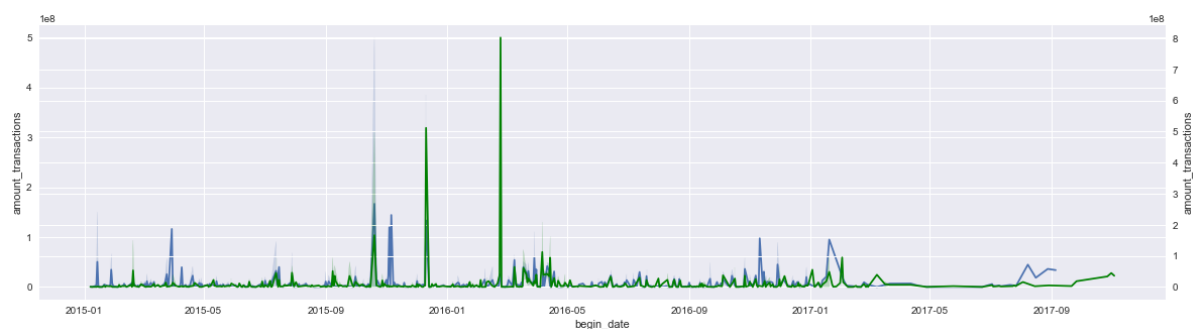
## Transactions map



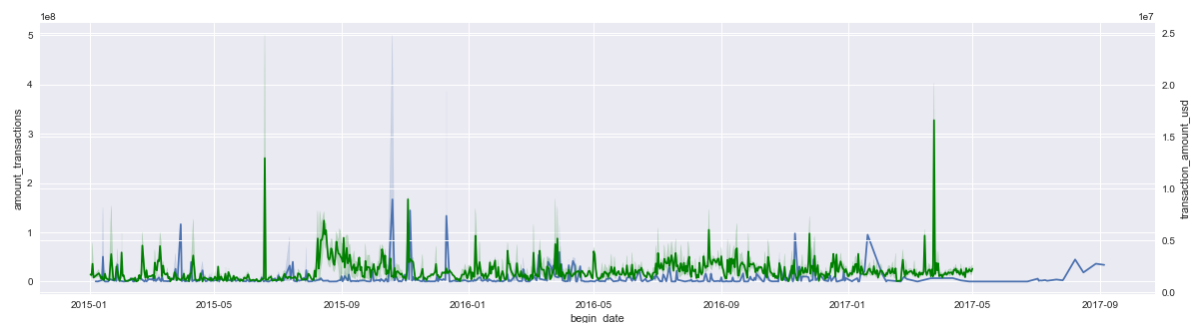
*Сопоставление по времени begin\_date и end\_date с number\_transactions*



Сопоставление по времени *begin\_date* и *end\_date* с *amount\_transactions*



Сопоставление по времени *begin\_date* и *time* с *transaction\_amount\_usd* и *amount\_transactions*



## Слияние данных

Таблица Bank connections, набора FinCEN и Transactions map имеют общие предикторы *icij\_sar\_id*, *filer\_org\_name\_id*, *filer\_org\_name*.

Выполнено внутреннее слияние по данным предикторам и фильтрация по периоду времени с 2015 года, чтобы итоговую таблицу можно было анализировать совместно с набором Bitcoin large transactions.

Пример итогового набора FinCEN за период с 2015 по 2017 годы



id	223254	223254	223254	223254	223254
icij_sar_id	3297	3297	3297	3297	3297
filer_org_name_id	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp
filer_org_name	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.
begin_date	2015-03-25 00:00:00	2015-03-25 00:00:00	2015-03-25 00:00:00	2015-03-25 00:00:00	2015-03-25 00:00:00
end_date	2015-09-25 00:00:00	2015-09-25 00:00:00	2015-09-25 00:00:00	2015-09-25 00:00:00	2015-09-25 00:00:00
originator_bank_id	cimb-bank-berhad	cimb-bank-berhad	cimb-bank-berhad	cimb-bank-berhad	cimb-bank-berhad
originator_bank	CIMB Bank Berhad	CIMB Bank Berhad	CIMB Bank Berhad	CIMB Bank Berhad	CIMB Bank Berhad
originator_bank_country	Singapore	Singapore	Singapore	Singapore	Singapore
originator_iso	SGP	SGP	SGP	SGP	SGP
beneficiary_bank_id	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr
beneficiary_bank	Barclays Bank Plc	Barclays Bank Plc	Barclays Bank Plc	Barclays Bank Plc	Barclays Bank Plc
beneficiary_bank_country	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom
beneficiary_iso	GBR	GBR	GBR	GBR	GBR
number_transactions	68	68	68	68	68
amount_transactions	5.68985e+07	5.68985e+07	5.68985e+07	5.68985e+07	5.68985e+07
entity_b_id	asb-bank-limited-auckland-new-zealand-nzl	china-citic-bank-international-ltd-hong-kong-hkg	commonwealth-bank-of-australia-sydney-australi...	national-australia-bank-limited-melbourne-vict...	investec-bank-switzerland-ag-zurich-switzerlan...
entity_b	Asb Bank Limited	China Citic Bank International Ltd	Commonwealth Bank of Australia	National Australia Bank Limited	Investec Bank
entity_b_country	New Zealand	Hong Kong	Australia	Australia	Switzerland
entity_b_iso_code	NZL	HKG	AUS	AUS	CHE

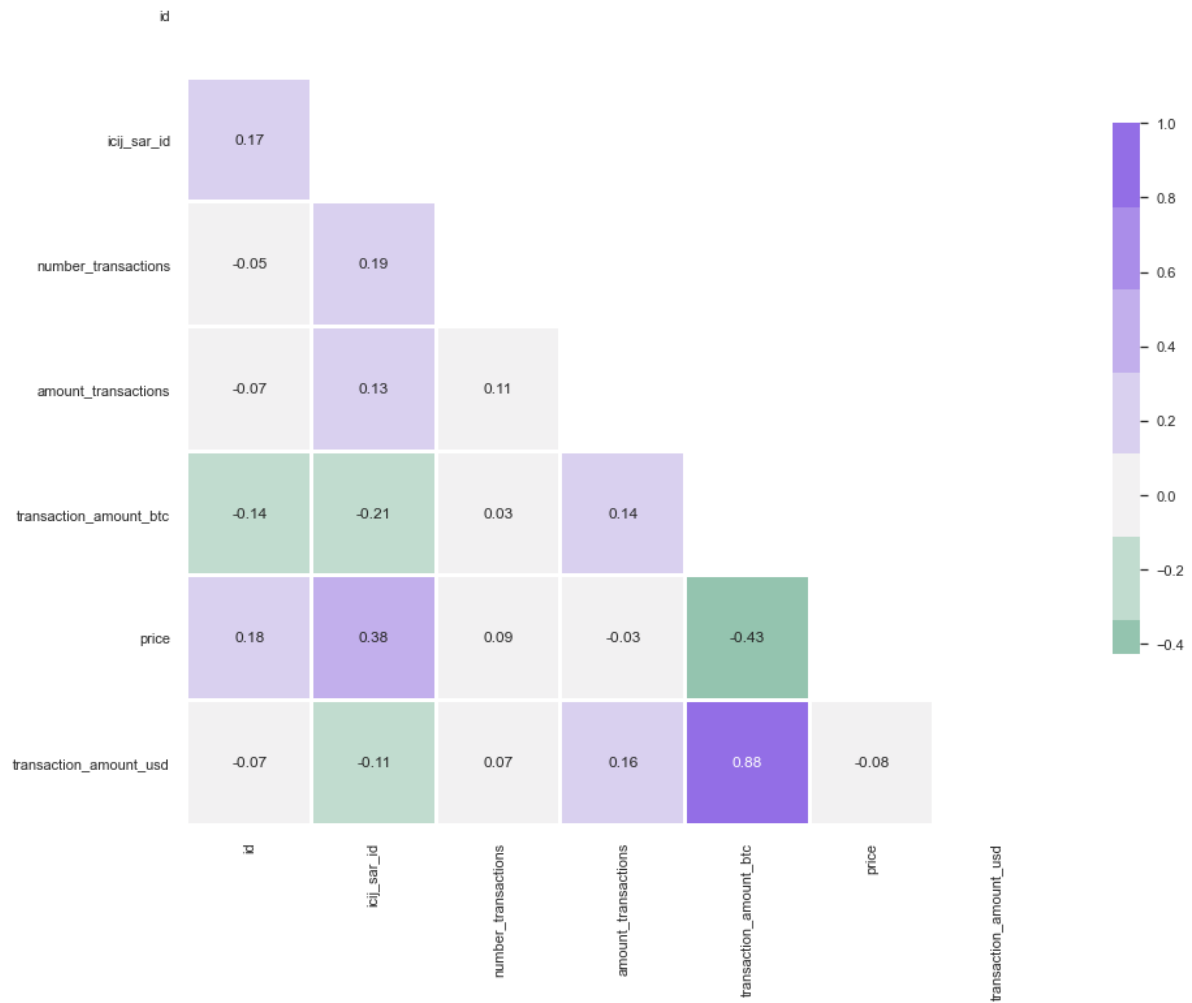
Анализ графика сопоставления по времени begin\_date и time с transaction\_amount\_usd и amount\_transactions показывает, что по этим переменным наборы не пересекаются в полной мере, поэтому выполним слияние наборов по времени оставив transaction\_amount\_usd и amount\_transactions без изменений.

Иллюстрация итоговой таблицы в которой сопоставлены подозрительные банковские операции с хэшами блокчейн-транзакций

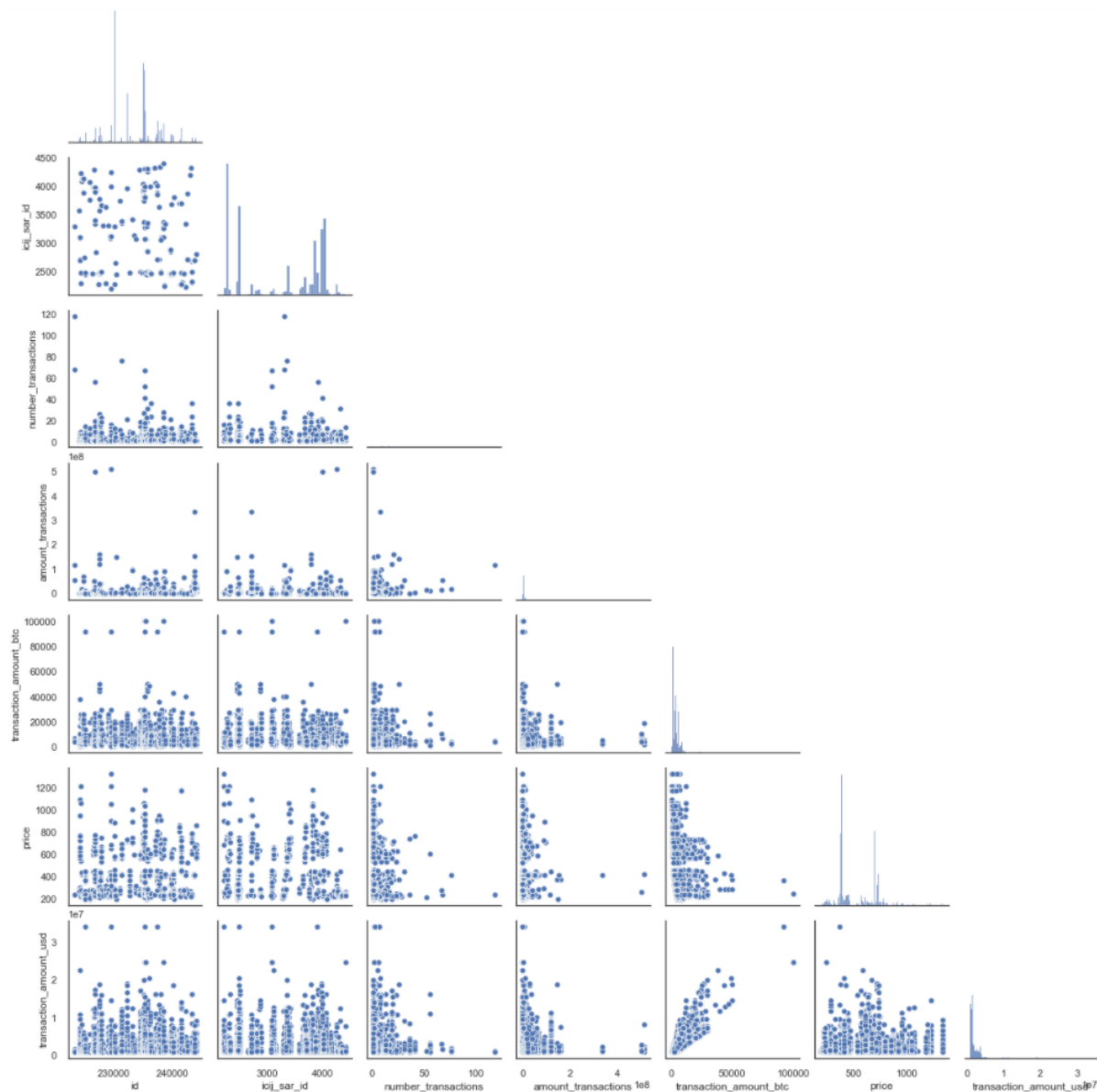
	0	1	2	3	4
id	223254	223254	223254	223254	223254
icij_sar_id	3297	3297	3297	3297	3297
filer_org_name_id	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp	the-bank-of-new-york-mellon-corp
filer_org_name	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.	The Bank of New York Mellon Corp.
begin_date	2015-03-25 00:00:00	2015-03-25 00:00:00	2015-03-25 00:00:00	2015-03-25 00:00:00	2015-03-25 00:00:00
end_date	2015-09-25 00:00:00	2015-09-25 00:00:00	2015-09-25 00:00:00	2015-09-25 00:00:00	2015-09-25 00:00:00
originator_bank_id	cimb-bank-berhad	cimb-bank-berhad	cimb-bank-berhad	cimb-bank-berhad	cimb-bank-berhad
originator_bank	CIMB Bank Berhad	CIMB Bank Berhad	CIMB Bank Berhad	CIMB Bank Berhad	CIMB Bank Berhad
originator_bank_country	Singapore	Singapore	Singapore	Singapore	Singapore
originator_iso	SGP	SGP	SGP	SGP	SGP
beneficiary_bank_id	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr	barclays-bank-plc-london-england-gbr
beneficiary_bank	Barclays Bank Plc	Barclays Bank Plc	Barclays Bank Plc	Barclays Bank Plc	Barclays Bank Plc
beneficiary_bank_country	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom
beneficiary_iso	GBR	GBR	GBR	GBR	GBR
number_transactions	68	68	68	68	68
amount_transactions	5.68985e+07	5.68985e+07	5.68985e+07	5.68985e+07	5.68985e+07
entity_b_id	asb-bank-limited-auckland-new-zealand-nzl	china-citic-bank-international-ltd-hong-kong-hkg	commonwealth-bank-of-australia-sydney-australi...	national-australia-bank-limited-melbourne-vict...	investec-bank-switzerland-ag-zurich-switzerlan...
entity_b	Asb Bank Limited	China Citic Bank International Ltd	Commonwealth Bank of Australia	National Australia Bank Limited	Investec Bank
entity_b_country	New Zealand	Hong Kong	Australia	Australia	Switzerland
entity_b_iso_code	NZL	HKG	AUS	AUS	CHE
hash	e27f7544ef6d370234e9df8fc5775077b8d7d0bc6f8898...	e27f7544ef6d370234e9df8fc5775077b8d7d0bc6f8898...	e27f7544ef6d370234e9df8fc5775077b8d7d0bc6f8898...	e27f7544ef6d370234e9df8fc5775077b8d7d0bc6f8898...	e27f7544ef6d370234e9df8fc5775077b8d7d0bc6f8898...
sender	3KBUuGko4H5ke7EVsq987PLK1c5Askdd7y	3KBUuGko4H5ke7EVsq987PLK1c5Askdd7y	3KBUuGko4H5ke7EVsq987PLK1c5Askdd7y	3KBUuGko4H5ke7EVsq987PLK1c5Askdd7y	3KBUuGko4H5ke7EVsq987PLK1c5Askdd7y
receiver	3KgtbGgaX2ngstNpyv7Lwph5WeVeqGbpM	3KgtbGgaX2ngstNpyv7Lwph5WeVeqGbpM	3KgtbGgaX2ngstNpyv7Lwph5WeVeqGbpM	3KgtbGgaX2ngstNpyv7Lwph5WeVeqGbpM	3KgtbGgaX2ngstNpyv7Lwph5WeVeqGbpM
transaction_amount_btc	7000	7000	7000	7000	7000
price	245.49	245.49	245.49	245.49	245.49
transaction_amount_usd	1.71843e+06	1.71843e+06	1.71843e+06	1.71843e+06	1.71843e+06

## Дополнительный анализ результатов

### Корреляционный анализ



## Взаимоотношения между переменными



## Оценка наличия выбросов в данных

