1. Предложить и проверить гипотезы, почему данные об установках со второго
источника (Source 2) могут расходиться с внутренней атрибуцией

После объединения двух источников данных —**installs_main.csv (внутренняя атрибуция)** и **installs_s2.csv (данные рекламного канала Source 2)** — были рассчитаны сводные статистики по установкам.

Ниже представлено сравнение ключевых метрик:

	install_date	installs	installs_main
count	622	622	622
mean	2022-11-07 12:00:00	93.604502	64.972669
min	2022-01-01 00:00:00	26	15
25%	2022-06-05 06:00:00	52	37
50%	2022-11-07 12:00:00	71	50
75%	2023-04-11 18:00:00	110	84
max	2023-09-14 00:00:00	397	213
std	NaN	64.215678	38.882136

Ключевые наблюдения

1. Source 2 стабильно показывает больше установок, чем внутренняя атрибуция

Это видно как по среднему значению, так и по максимальным и медианным значениям.

- Среднее Source 2 выше на **~44%**
- Медиана выше на **~42%**
- Максимальные значения также значительно превышают внутреннюю метрику

Это говорит о **систематическом расхождении**, а не случайных всплесках.

2. Стандартное отклонение также выше у Source 2

- Source 2: **64.22**

- Internal: **38.88**

Это означает, что **данные Source 2 более волатильны**, что может быть признаком:

- повторных учётов установок,
- иных алгоритмов расчёта.

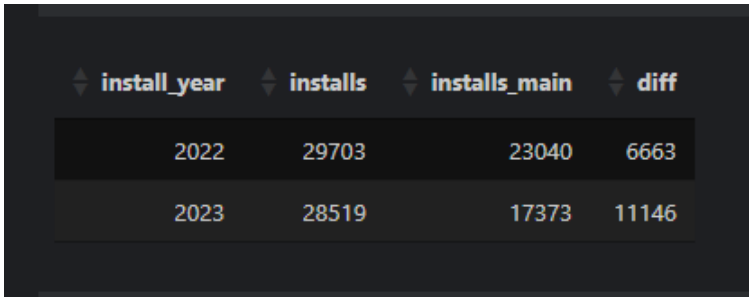
3. Период данных совпадает

Оба источника покрывают один и тот же период:

Значит, расхождение вызвано **не разным диапазоном дат**, а **разными методами атрибуции**.

Расхождения между Source 2 и внутренней атрибуцией действительно наблюдаются на уровне базовой статистики:
у Source 2 существенно больше установок по всем ключевым метрикам, а сами данные — более вариативные.

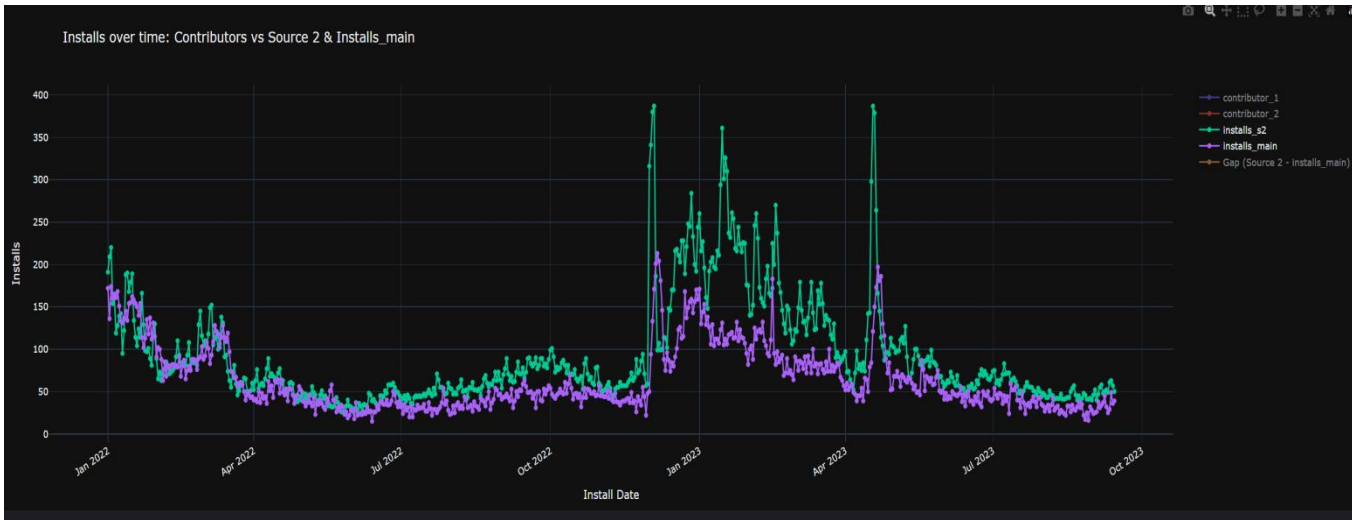
После первичной оценки общей статистики данные были агрегированы по годам.
Годовая группировка показывает, что разница между количеством установок Source 2 и внутренней атрибуцией **существенно увеличивается в 2023 году**.



install_year	installs	installs_main	diff
2022	29703	23040	6663
2023	28519	17373	11146

Разница установок в 2023 году составляет **+4483** в пользу Source 2,
при том что данные охватывают период **только до сентября 2023 года**.

Следовательно можно предположить, что именно в 2023 году присутствуют аномалии и выбросы, приводящие к систематическому завышению метрик Source 2.



Далее был построен график динамики установок по датам.

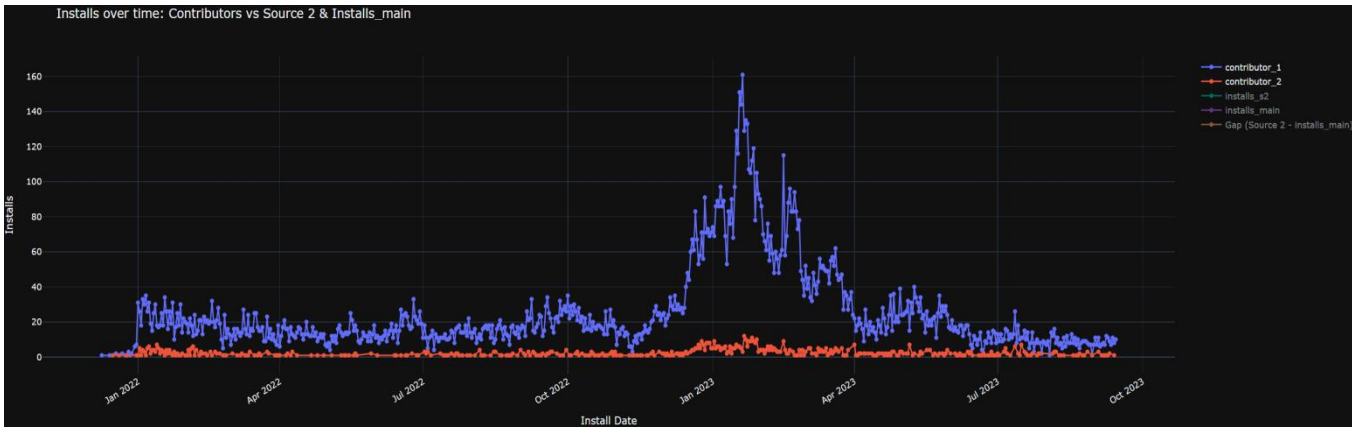
Как и предполагалось, наиболее значимые расхождения и всплески приходятся на период:

декабрь 2022 — май 2023

В этом диапазоне Source 2 показывает резкие скачки установок, которые не отражаются во внутренней атрибуции.

Это позволяет предположить:

- изменение политики атрибуции у Source 2,
- начало использования probabilistic matching (fingerprint),
- расширение окна атрибуции,
- включение view-through установок,
- изменение качества данных (например, ошибочные device match-и).



Чтобы понять природу расхождений, дополнительно были рассмотрены цепочки касаний:

- **contributor_1** — предпоследний источник перед установкой
- **contributor_2** — источник за два касания до установки

На графиках распределений по времени видно:

✓ Source 2 почти отсутствует в contributor_2

Это означает, что Source 2 **редко является первым касанием** и практически не открывает цепочки.

✓ Но Source 2 регулярно появляется в contributor_1

А это уже важный сигнал:
Source 2 может засчитывать установки, опираясь на присутствие в contributor_1.

✓ У contributor_1 также есть всплески в декабре 2022 — апреле 2023

А это совпадает с всплесками в самом installs_s2.

На предыдущем этапе было замечено, что **Source 2 часто появляется в поле contributor_1**,
что может указывать на то, что installs_s2 учитывает не только последний клик,
но и предпоследний (или считает его частично).

Чтобы проверить эту гипотезу, была проведена корректировка:

Новая метрика:

installs_main_pl_comt_1 = installs_main + contributor_1_count(Source 2)

То есть к внутренним установкам добавлялись случаи, когда Source 2 присутствовал как contributor_1.

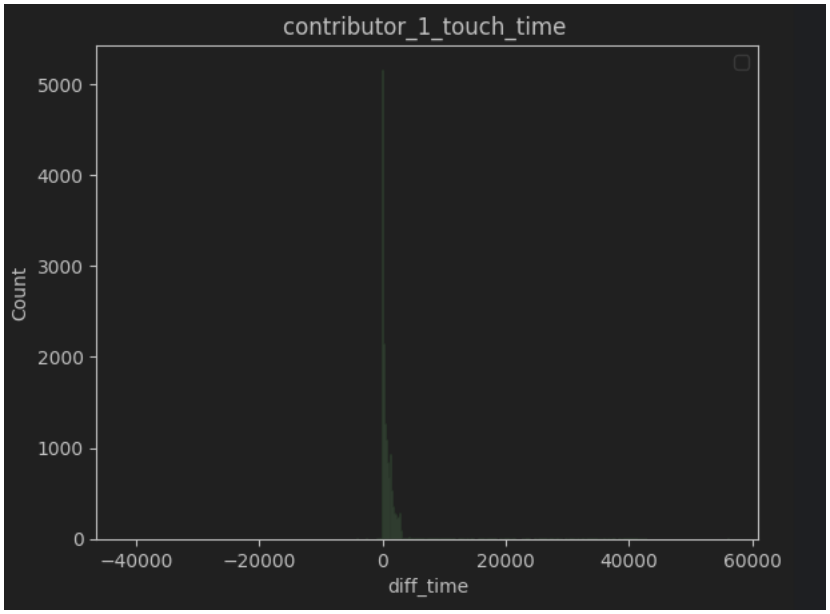
install_year	installs	installs_main	installs_main_pl_comt_1
2022	29703	23040	30072
2023	28519	17373	26229

После добавления contributor_1:

- линия внутренней атрибуции заметно приближается к installs_s2;
- общий уровень значений становится значительно ближе;
- пик декабрь 2022 — май 2023 также частично компенсируется.

Это подтверждает, что Source 2 действительно засчитывает часть установок, где он находится на позиции contributor_1.

В источнике так же есть данные contributor_1_touch_time и contributor_2_touch_time. Я так понимаю это дата совершения события в contributor_1 и contributor_2 соответсвенно.
Проверим какая разница между датами contributor_1_touch_time , contributor_2_touch_time и install_time из основного источника (main).
Для этого я от даты установки отнимал contributor_1_touch_time и получил разницу. Ниже предоставлена информация распределения этой разницы.



	dif_time
count	15916
mean	1449.214268
std	4699.575813
min	-41508.466667
25%	40.495833
50%	378.975
75%	1259.05
max	56170.533333

График и таблица статистик показывают, что:

- данные имеют **очень большой разброс** — присутствуют как сильно отрицательные, так и сильно положительные значения;
  это говорит о том, что касание могло происходить значительно раньше (или позже) времени установки;
- встречаются значения с лагом, измеряемым десятками часов или даже дней — что нетипично для классической last-click атрибуции.

Однако ключевое наблюдение:

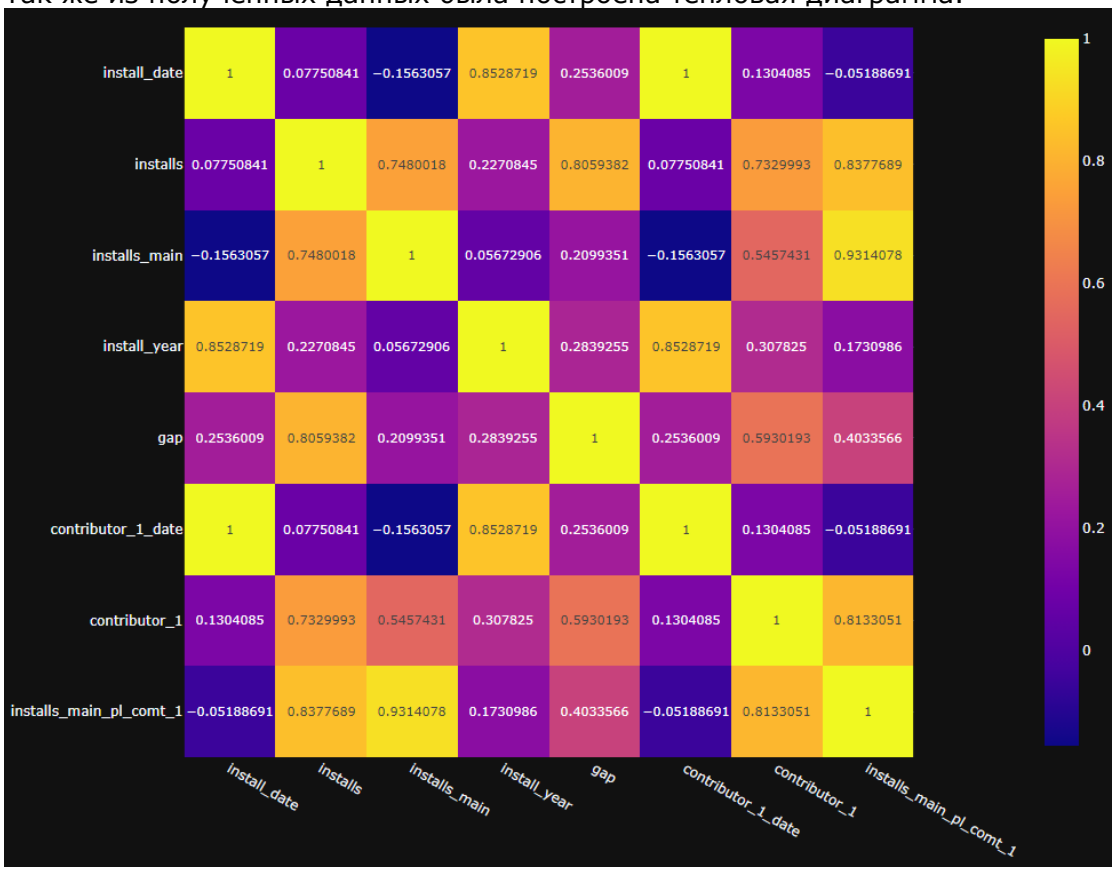
75% значений (для строк, где installs_main и contributor_1 == 'Source 2') находятся ниже 1259 минут

Это примерно **21 час**.

Что это означает?

- Для большинства установок Source 2 действительно находится во временном окне **до суток**, что соответствует стандартным практикам атрибуции.
- При этом оставшиеся 25% имеют **аномально большие лаги**, которые могут объяснять перерасход или разброс между внутренней атрибуцией и данными Source 2.

Так же из полученных данных была построена тепловая диаграмма.



Ключевые наблюдения:

- **Корреляция между installs_s2 и gap составляет ~0.8**
- **Корреляция между installs_main и gap — лишь ~0.2**

Что это означает:

1. Разрыв (gap) практически полностью определяется поведением installs_s2.
Если у Source 2 происходит всплеск, gap резко возрастает. Если спад — gap уменьшается.

2. installs_main остаётся стабильным и не демонстрирует синхронных скачков.
Это значит, что внутренний источник фиксирует установки в равномерном режиме,
без искусственных выбросов.

3. Такая диспропорция — прямой признак изменения или расширения алгоритма учёта Source 2.

Когда gap движется **следом за Source 2**, это говорит о том, что расхождение вызвано:

- дополнительными типами атрибуции (view-through),
- новым окном атрибуции,
- изменённым matching-алгоритмом,
- учётом повторных запусков (re-open) или reinstall,

Итоговые выводы анализа

1. Расхождение действительно существует и систематическое

Сравнение данных installs_s2 и installs_main показало устойчивую разницу,
которая не объясняется случайными колебаниями.

2. Аномалия во времени contributor_1 полностью совпадает со скачком installs_s2

Наибольшее различие наблюдается в декабре 2022 — мае 2023.
В этот же период зафиксирован резкий рост присутствия Source 2 в contributor_1.

Это ключевой индикатор: админка Source 2 начала учитывать предпоследний клик.

3. 75% установок Source 2 совершаются в течение ~21 часа после касания contributor_1

Это укладывается в нормальный атрибуционный период,
но оставшиеся 25% имеют большие лаги — до многих суток.

Это свидетельствует об **расширенном окне атрибуции** Source 2.

4. У внутренней атрибуции окно жёстко ограничено — она учитывает только факт установки

Поэтому installs_main остаётся стабильным, в то время как installs_s2 реагирует на дополнительные клики/показы.

5. Для окончательной валидации необходимо изучить логи

Чтобы исключить технические артефакты (дубли, ложные матч-комбинации),
требуется анализ логов обоих источников:

- postback-и Source 2
- install event внутренних данных
- уникальные device_id / advertising_id
- временные последовательности кликов

Финальный вывод

Все проведённые проверки подтверждают: Source 2 изменил или расширил алгоритм атрибуции.
Он начал учитывать клики или просмотры гораздо сильнее, чем внутренняя модель,
и допускает более широкие временные интервалы.

Внутрення атрибуция работает стабильно и учитывает только реальные установки,
в то время как Source 2 “дотягивает” себе дополнительные события,
основанные на логике касаний и вероятностных методов.

2) Проверить данные на наличие аномалий и найти инсайты, которые могут
помочь принять решения об атрибуции установок в некоторых каналах

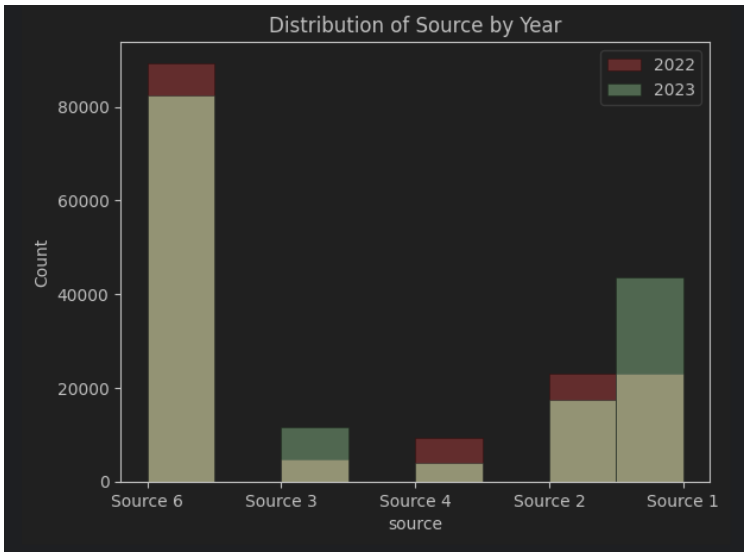
Детальней рассмотрим источник installs_main.csv.

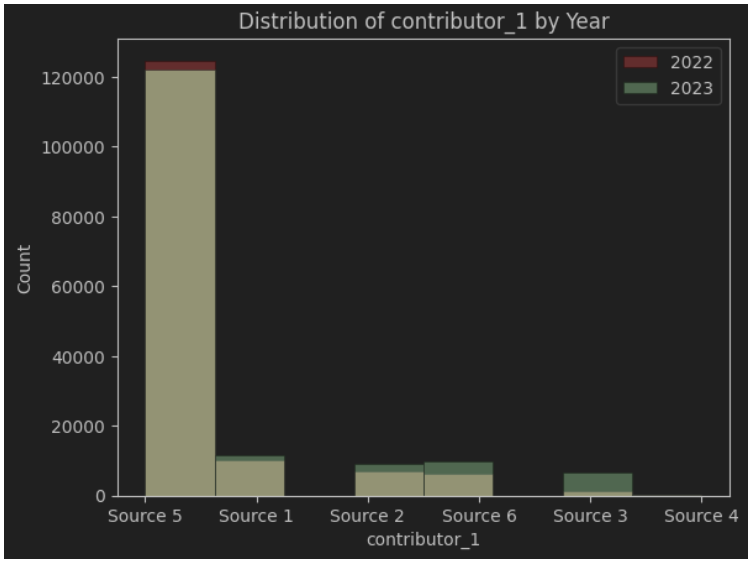
```
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   Unnamed: 0               308568 non-null  int64
 1   install_time             308568 non-null  datetime64[ns]
 2   source                   308568 non-null  object
 3   contributor_1            308568 non-null  object
 4   contributor_2            308568 non-null  object
 5   contributor_1_touch_time  61883 non-null   datetime64[ns]
 6   contributor_2_touch_time  15021 non-null   datetime64[ns]
 7   contributor_1_date        61883 non-null   datetime64[ns]
 8   contributor_2_date        15021 non-null   datetime64[ns]
 9   install_date             308568 non-null  datetime64[ns]
 10  install_year             308568 non-null  int32
 11  install_hour             308568 non-null  int32
 12  contributor_1_hour        61883 non-null   float64
 13  contributor_2_hour        15021 non-null   float64
 14  install_day              308568 non-null  int32
dtypes: datetime64[ns](6), float64(2), int32(3), int64(1), object(3)
```

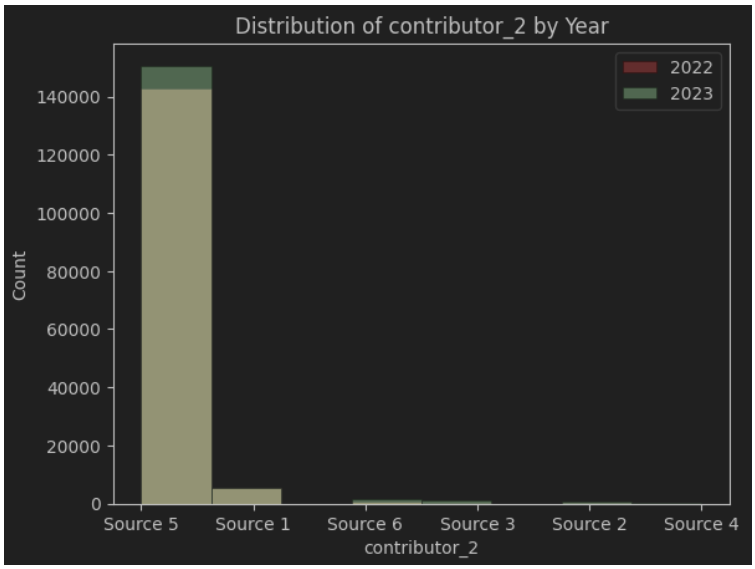
Анализ трёх основных полей:

- **source** — фактический источник установки
- **contributor_1** — предпоследний источник
- **contributor_2** — пред-предпоследний источник

позволяет понять общий состав трафика и его динамику.







Анализ source*:*

- **Source 6** является доминирующим источником установок в исследуемом периоде.
- **Source 2** резко увеличивает свою долю в 2023 году.

Это может указывать на:

- рост рекламной активности Source 2,
- либо изменение механизма атрибуции (например, расширение окна атрибуции).

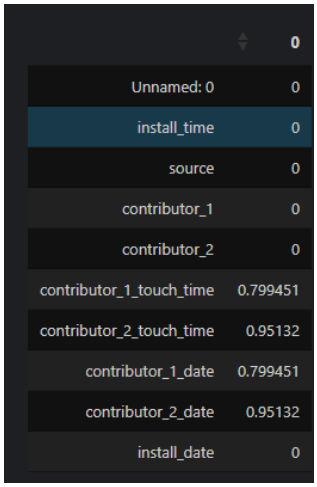
Анализ contributor_1 и contributor_2

- В обоих полях **contributor_1** и **contributor_2** доминирует **Source 5**.
- Это означает, что пользователи часто взаимодействуют с рекламой Source 5 **до того, как произведут установку**, однако финальная атрибуция уходит другим каналам.

Возможное объяснение:

- Source 5 может работать как **канал привлечения**.
- Или же Source 5 проигрывает борьбу за атрибуцию последнего касания.

Видно, что в полях contributor_1_touch_time и contributor_2_touch_time есть пропуски.



	0
Unnamed: 0	0
install_time	0
source	0
contributor_1	0
contributor_2	0
contributor_1_touch_time	0.799451
contributor_2_touch_time	0.95132
contributor_1_date	0.799451
contributor_2_date	0.95132
install_date	0

contributor_1_touch_time и contributor_2_touch_time так как нет описаний данных полей я предположу это время некого события на источнике (клик). Так как источники указаны , а времени нет, можно предположить, что это был просто просмотр.
Однако можно проверить сколько соответствуют условию contributor_1=source и наоборот, при условии что contributor_1_touch не пустой. contributor_2 не будем рассматривать, так как много пропусков.

Первым рассмотрим где не совпадает contributor_1 и source.

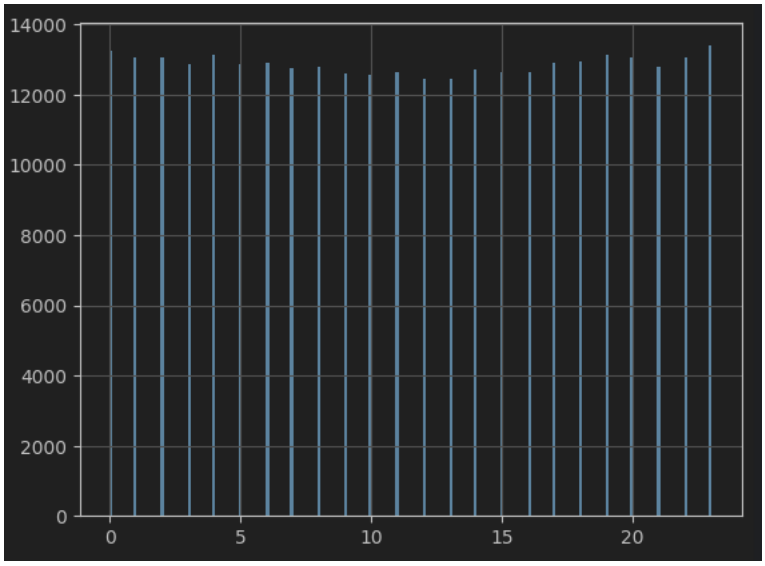
contributor_1	Unnamed: 0
Source 2	15916
Source 1	11312
Source 6	8930
Source 3	7020
Source 4	335
Source 5	28

И где совпадают contributor_1 и source

contributor_1	Unnamed: 0
Source 1	10497
Source 3	746
Source 6	7099

Из таблиц можно сделать вывод, что Source 2 кликают часто, но установка не происходит, а вот с Source 1 обратная картина.

Гипотеза о влиянии времени на установку.



Результат проверки

По графику:

- **распределение установок по часам суток практически равномерное**,
- нет выраженных всплесков ни утром, ни вечером,

Анализ самых популярных цепочек установок.
Для более глубокого понимания поведения пользователей и конкуренции источников за последнее касание был проведён анализ наиболее частых цепочек:

	install_year	source	contributor_1	contributor_2	count_install
79	2022	Source 6	Source 5	Source 5	81219
193	2023	Source 6	Source 5	Source 5	68973
106	2023	Source 1	Source 5	Source 5	32738
14	2022	Source 1	Source 5	Source 5	15701
29	2022	Source 2	Source 5	Source 5	15032
61	2022	Source 4	Source 5	Source 5	8875
151	2023	Source 3	Source 5	Source 5	8716
125	2023	Source 2	Source 5	Source 5	7936
71	2022	Source 6	Source 2	Source 5	4662
198	2023	Source 6	Source 6	Source 5	4564

Ключевые наблюдения из цепочек

1. Source 5 — доминирующий contributor

Практически во всех топ-10 цепочек:

- **contributor_1 = Source 5**,
- **contributor_2 = Source 5**.

Это означает:

- пользователи очень часто взаимодействуют с Source 5 перед установкой;
- однако **Source 5 редко получает финальную атрибуцию (source)**.

Это подтверждает выводы предыдущего анализа:

Source 5 работает как массовый рекламный канал, создаёт много касаний, но плохо выигрывает последнее касание.

2. Source 6 — системный победитель атрибуции

В **двух первых строках** цепочек:

- Source 6 → финальный источник,
- Source 5 → оба предыдущих источника.

Такая комбинация встречается **140+ тысяч раз**, что делает её самым распространённым паттерном поведения пользователей.

Самые популярные цепочки показывают, что **Source 5** является основным источником первых касаний, но **финальную атрибуцию почти всегда берут Source 6, Source 1 или Source 2**.

Это доказывает, что расхождение между installs_s2 и installs_main связано не с поведением пользователей, а с различиями в алгоритмах атрибуции.

Дополнительно можно посмотреть дашборд в папке Power_BI, а так же была попытка реализовать модель придикта поля source (predict_source.ipynb)