Тестовое задание на позицию Junior Traffic Analyst

Описание данных

visits.csv:

- id – id визита
- started_at – дата визита на сайт, то есть когда человек первый раз зашел к нам
- utm_medium – канал трафика
- device_type - указание девайса, с которого пользователь зашел на сайт
- utm_source - поисковая система пользователя
- browser - браузер, который использовал пользователь

views.csv:

- id – id просмотра экскурсии, которую посмотрел пользователь
- visit_id – id визита
- time – время просмотра страницы экскурсии

orders.csv:

- id – id заказа, который создал человек во время визита
- visit_id – id визита
- created_at – дата создания заказа
- state – статус заказа. Статус может быть 3-х типов:
  - pending - заказ еще не был проведен на момент выгрузки
  - held - заказ успешно проведен
  - canceled - заказ был отменен

Задание 1

Ссылка на Google Colab:
https://colab.research.google.com/drive/184iFDlfQL3ifE24vyEiJpjcX9Nyvt0z_?usp=sharing

1.пункт

Необходимо вычистить ботные и некорректные данные из набора данных **visits.csv**

a. Нам известно, что на органический канал трафика приходят подозрительные посещения из браузера Chrome Mobile, при utm_source яндекс и мобильном девайсе. Необходимо вычистить посещения, которые отвечают описанным условиям.
b. ***(не обязательно, но круто)***. *Можешь попробовать найти другие аномалии в данных и предложить, как их можно почистить.*

1.a. Для вычистки ботных и некорректных данных из visits.csv было необходимо выгрузить их в SQL-среду с использованием Python и SQLite. При преобразовании Data Frame в SQL таблицу по ключу id возникла ошибка (обнаружились дубликаты),

поэтому их я убрала на этапе подготовки. Далее были вычищены посещения из условий пункта 1.a.

1.b. Рассмотрим основные поля и их возможные аномалии в таблице visits:

- ботные/очень редкие браузеры (=browser),
- ошибки в каналах трафика(=utm_medium) и с поисковой системой пользователя (=utm_source),
- подозрительные девайсы (=device_type).

Также в таблице могут встречаться неполные данные. Если их количество очень маленькое и они не несут за собой важных инсайтов (возможно, просто ошибки выгрузки) – их можно убрать.
В нашем случае были подчищены браузеры с кол-вом визитов < 10, а также строки с пустыми значениями. С каналами трафика аномалий не выявлено, с поисковой системой пользователя возникли сомнения, но так как я не обладаю полной информацией – лучше не убирать. Также из таблицы были убраны пустые значения девайсов.

В итоге из изначальной таблицы visits с 850 912 строками **были удалены 48 078 строк**. Подробный ход анализа описан в Google Colab:)

2. пункт

Оцени эффективность работы органического канала трафика (=organic).

a. Предложи метрики, по которым мы сможем оценить эффективность органического канала трафика.
b. Объясни, как именно они помогают оценить эффективность.
c. Какой месяц был самый удачный для этого канала, а какой самый неудачный?

2.a. Органический трафик – это посетители, которые переходят на сайт из обычных поисковых систем. Его эффективность можно оценить по следующим доступным нам метрикам:

- кол-во визитов,
- кол-во просмотренных экскурсий,
- кол-во просмотренных экскурсий за визит,
- кол-во заказов,
- кол-во успешных(state=held) заказов,
- отношение успешных заказов к визитам (конверсия в успешный заказ).

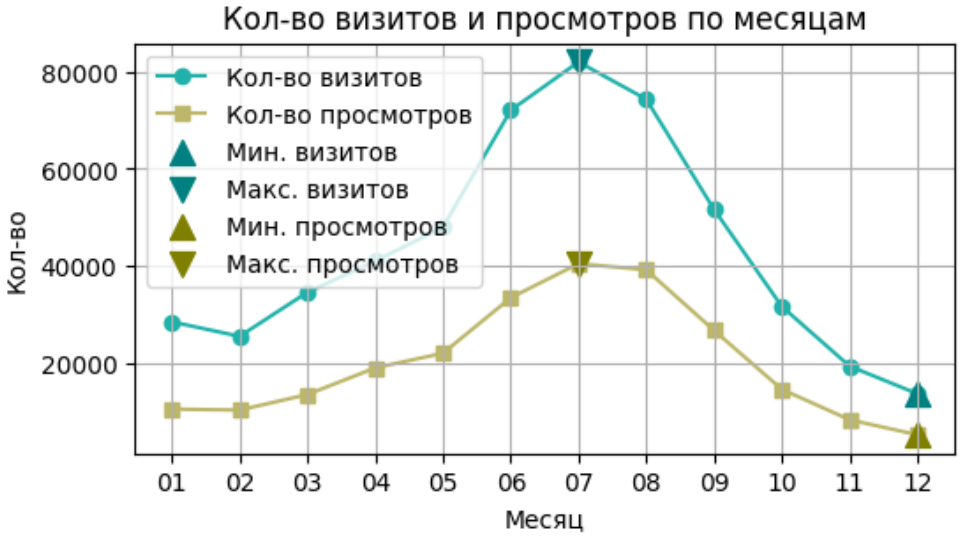
2.b. Опишем важность каждой из метрик:

- Визиты - базовая характеристика о работе органического трафика, т. е. работает ли он вообще, его масштаб.

- Просмотренные экскурсии – общая заинтересованность посетителей в экскурсиях, интерес пользователей.
- Просмотренные экскурсии за визит (среднее(кол-во просмотров) на визит) – усредненная метрика, показывает медианную активность пользователя, вовлеченность, заинтересованность в выборе экскурсии.
- Заказы – реальные покупатели, оформившие заказ; прямая ценность.
- Успешные заказы – заказы, приносящие деньги; основная показательная метрика для анализа канала органического трафика (можно также посмотреть отношение успешных заказов ко всем заказам, не возникает ли каких-то проблем на этапе оформления заказа).
- Успешные заказы к визитам – конверсия в заказы, показывает сколько пользователей приобрели экскурсию за визит, также напрямую связана с прибылью. Наверное, самая важная метрика, связанная с выручкой.

2.c. Чтобы понять какой месяц был самый удачный для этого канала, а какой самый неудачный нам нужно рассмотреть несколько метрик, визуализировать результаты и сделать выводы. Пока что можем предположить, что это будет какой-то из летних месяцев!

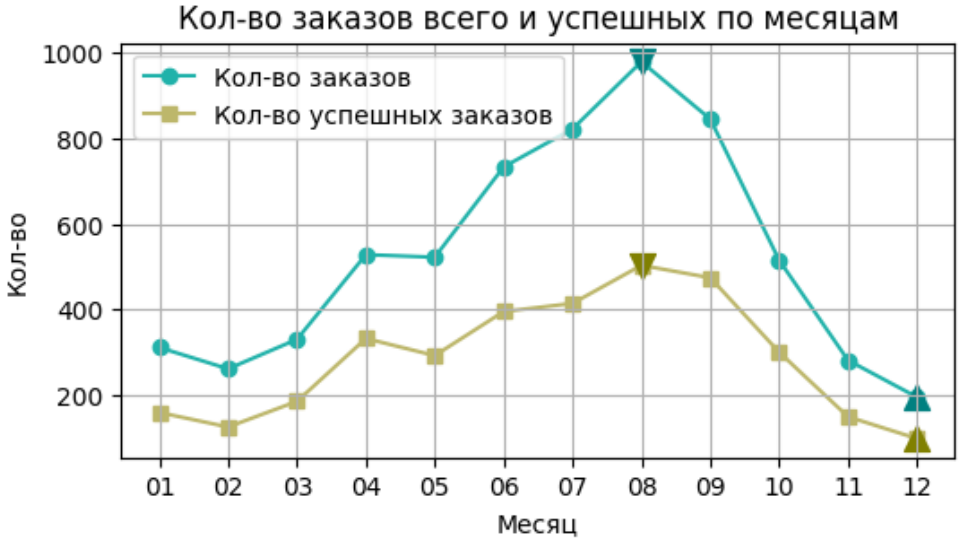
Рассмотрим две самые базовые метрики – количество визитов и количество просмотренных экскурсий по месяцам.



По графику видим, что месяцем с наибольшим кол-вом визитов и просмотров является *июль*, когда как наименьшее кол-во в этих двух характеристиках у *декабря*.
Также посмотрим на среднее количество просмотренных экскурсий за визит (отношение предыдущих двух метрик).

Видим, что ситуация отличается, и уже другой месяц лидирует в этой метрике – наибольшее количество просмотров на один визит в *августе*, наименьшее – в *январе*.

Перейдем к метрикам, связанным с заказами. Посмотрим на общее количество заказов и часть успешных заказов на графике ниже.



Наилучший месяц по кол-ву заказов (как успешных, так и общего количества) – *август*, наихудший – *декабрь*.

На последок проанализируем последнюю рассмотренную мною метрику – конверсию визитов в заказы.

Для этой метрики самым эффективным месяцем является *октябрь*, при этом самым неэффективным оказался *февраль*, на втором месте – *июль*. Неожиданные результаты! То есть самый масштабный месяц по трафику оказался одним из самых неэффективных по конверсии в успешные заказы.

Выводы:

- Если за “удачливость” месяца для органического канала мы понимаем масштаб и популярность сайта (вкладки с экскурсиями в Анапе), то безусловными лидерами будут летние месяцы, как и предполагалось, а именно *июль*. А самым непопулярным сайт был в *декабре*.
- Если же мы хотим сфокусироваться на выручке, и под “удачливостью” иметь в виду коммерческую выгоду, то самым прибыльным по количеству купленных заказов стал *август*, а наименее прибыльным – *декабрь*.
- Рассмотрев конверсию визитов в заказы и связав “удачливость” с эффективностью сервиса, самым удачным стал *октябрь*, а наименее удачным – *февраль*.

Подводя итоги анализа различных метрик, я бы всё-таки **отдала первенство самого удачного месяца для органического канала *августу***, так как он находится в ТОП-3 по метрикам масштаба (популярности), а также оказался самым выгодным по заказам. **Самым неудачным для органического канала по большинству показателей оказался *декабрь***.

3. пункт

Построй прогноз визитов для органического канала трафика для июля-сентября 2024 года по месяцам.

a. Как изменится прогноз, если органический трафик будет работать эффективнее на 5%?

Построим прогноз для органического канала трафика для июля-сентября 2024 года по месяцам используя модель для прогноза временных рядов – SARIMA.
Для начала визуализируем временной ряд для визитов с 2023-03-01 по 2024-06-30.

Применяем модель SARIMA и получаем прогнозные значения. Визуализированные результаты представлены на графике ниже.

Как и ожидалось, новый пик будет в июле, снижение в августе, и стремительное снижение в сентябре. Также на графике виден доверительный интервал прогноза (розовая полупрозрачная полоса), которая показывает возможные колебания значений.

В пункте а. было необходимо посмотреть на изменения в прогнозе, если трафик будет работать на 5% эффективнее. Добавим на предыдущий график новую линию – обновленный прогноз с повышенной эффективностью:

Теперь посмотрим на численные значения визитов в таблице.

Дата	Прогноз визитов	Прогноз визитов с трафиком + 5%
2024-07-01	94331	99047
2024-08-01	85516	89791
2024-09-01	59470	62443

4. пункт

Выбери канал с самым высоким потенциалом. Объясни, почему именно этот канал обладает высоким потенциалом.

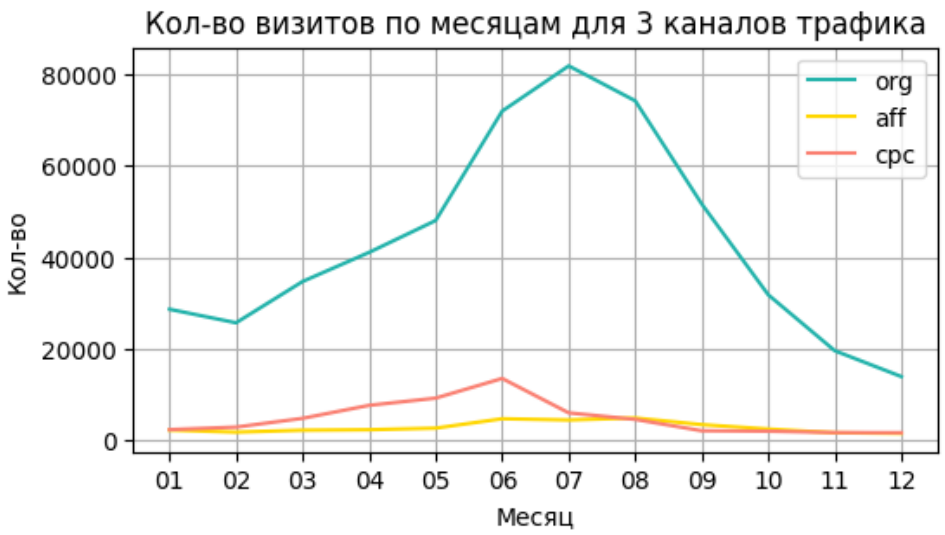
Для выбора канала с самым высоким потенциалом **необходимо определится – что характеризует высокий потенциал**. Перечислим основные идеи:

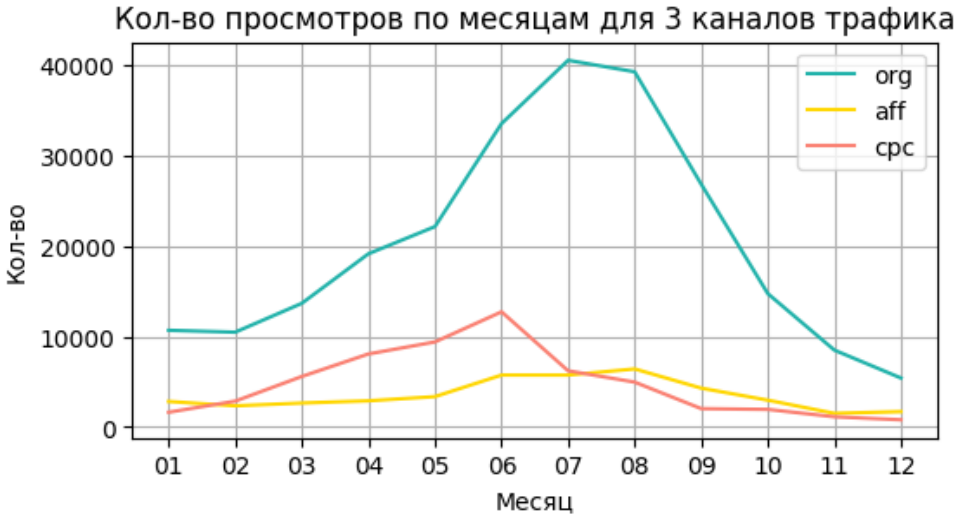
- у этого канала стабильные метрики,
- наблюдается рост в доходе канала, высокая конверсия,
- высокие прогнозные значения.

Соответственно, такой канал будет выгодно развивать и масштабировать.

Для начала сравним метрики рассмотренные в пункте 2. В графиках буду использовать сокращения для каналов трафика: org – органический канал, aff – affiliate (партнерский канал), cpc – cost-per-click (поисковая реклама).

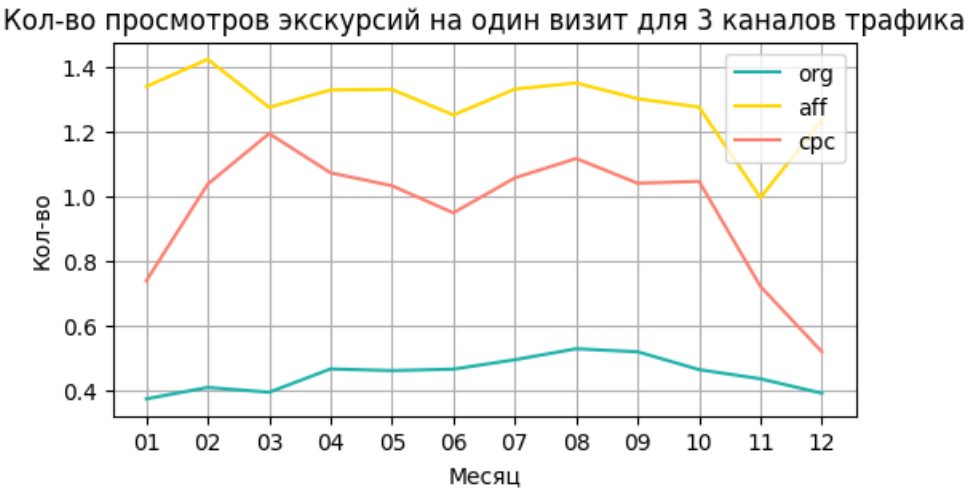
Ниже демонстрируются 2 графика с количеством визитов и просмотров экскурсий по каналам.





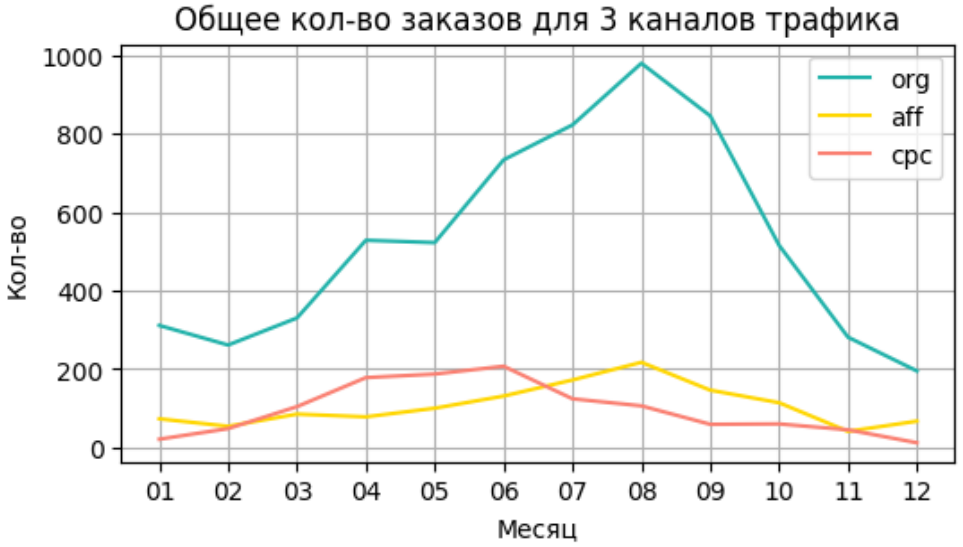
Можем заметить, что лидером остается *органический канал*, *партнерский* и *cpc каналы* достаточно близки по этим метрикам, но *cpc канал* имеет более высокие показатели с февраля по июль.

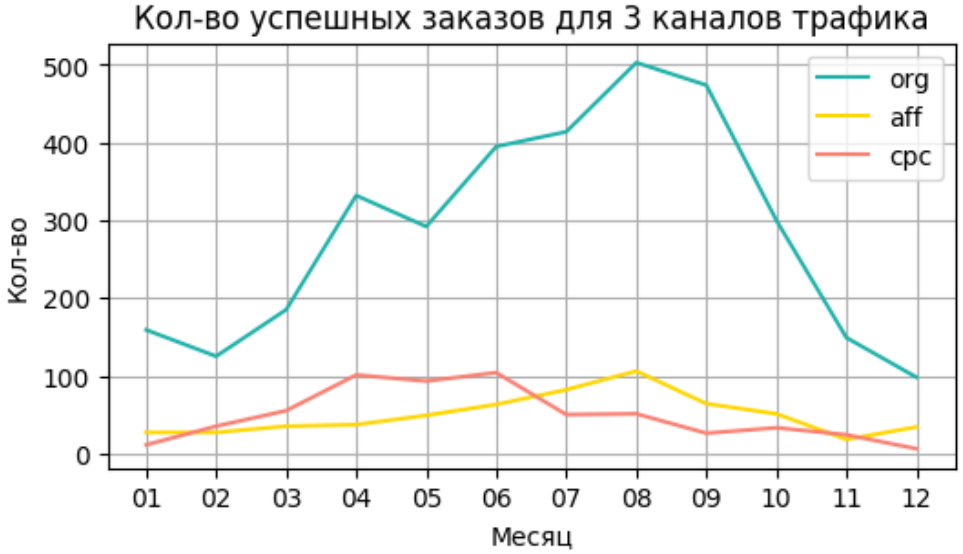
Далее проверим отношение этих двух метрик – появится ли явный лидер?



Да, по графику наиболее высокий показатель заинтересованности в экскурсиях у *партнерского канала*. Можно предположить, что пользователи, приходящие по партнерской ссылке, более вовлечены и намерены выбрать экскурсию, чем приходящие по остальным двум каналам. Самый низкий показатель у *органического канала* – стабильно более 40% пользователей не посетили страницу ни одной экскурсии.

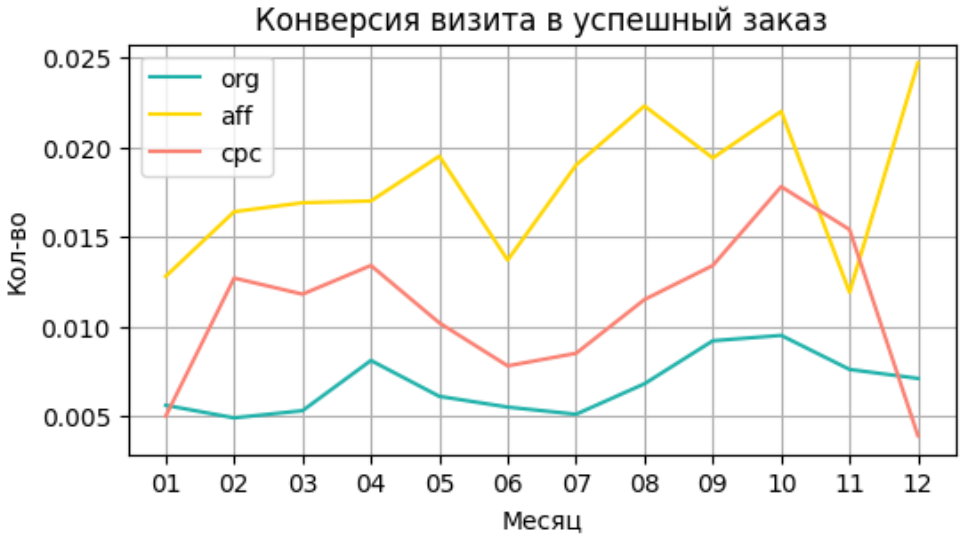
Далее проанализируем графики, связанные с заказами.





На приведенных выше графиках об общем количестве заказов и успешном наибольшее количество у *органического трафика*. *Cpc* и *партнерские каналы* примерно на одном уровне (cpc всё-же имеет больше заказов).

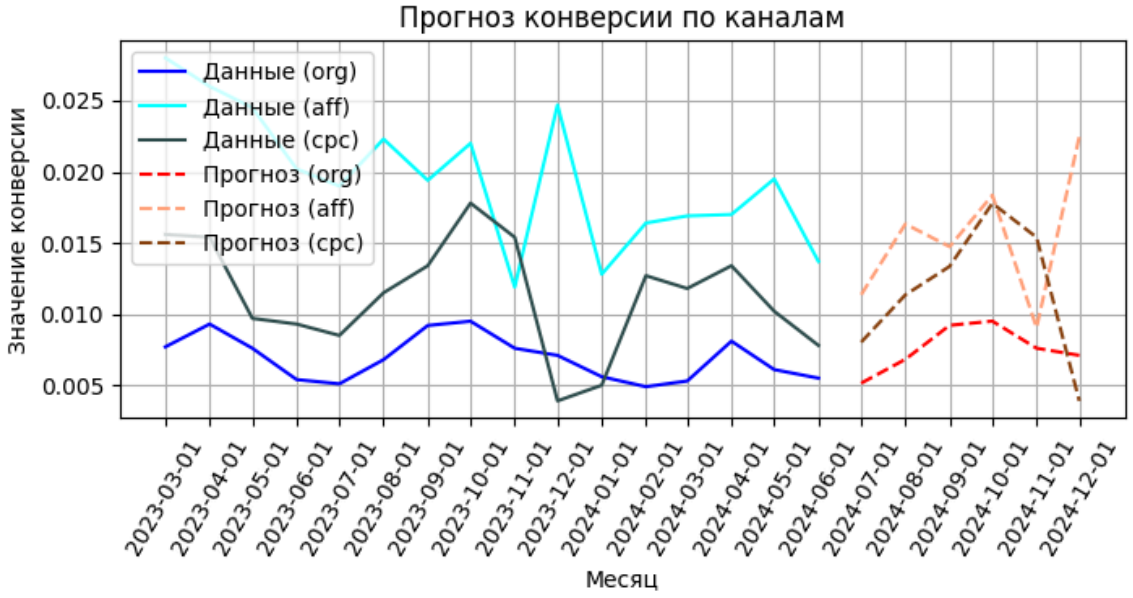
Теперь посмотрим на конверсию визитов в успешные заказы на графике ниже.



Нетрудно заметить, что наибольшей конверсией визитов в успешные заказы обладает партнерский канал с пиком ~0.025, на втором месте находится cpc канал, а наименьшей конверсией обладает самый популярный из каналов – органический.

Промежуточные выводы: На данном этапе видно, что *органический канал* – точно не претендент на канал с самым высоким потенциалом. *Канал cpc* имеет большее количество визитов, просмотров, и, соответственно, заказов, однако пользователи перешедшие по *партнерским ссылкам* просматривают больше всего экскурсий и наиболее вероятно оформят успешный заказ.

Рассматривать прогнозы основных метрик (визиты, просмотры, заказы) скорее всего не поможет выбрать канал с самым высоким потенциалом, так как они достаточно равномерны. Поэтому посмотрим на прогноз конверсии визита в успешный заказ на 6 месяцев вперед!
Получился следующий график:



Необходимо отметить, что я столкнулась с определенными трудностями в составлении этого прогноза. Скорее всего основными причинами являются: небольшое количество наблюдений и отсутствие видимого тренда (данные выглядят как будто не имеют никакого закона распределения). Подробные шаги можно посмотреть в Colab.

Но если допустить, что конверсия будет схожа прогнозу, то партнерский канал будет с самой высокой конверсией.

Выводы: Анализируя всё вышеописанное, **канал с самым высоким потенциалом – *партнерский (affiliate)***. Распишем характеристики партнерского канала:

- стабильное количество визитов и просмотров, нет явных спадов метрик;
- высокая заинтересованность посетителей пришедших по этому каналу – самое высокое кол-во просмотров экскурсий на визит;

и наконец,

- самая большая конверсия визитов в заказы с сильным отрывом от других каналов – канал с самым качественным трафиком.

Из характеристик упомянутых в начале 4. пункта не наблюдается только устойчивый рост успешных заказов, но судя по другим положительным трендам – при масштабировании партнерского канала и, возможно, оптимизирования пользовательского опыта, можно добиться значительного роста выручки!

5. пункт

(Не обязательно, но будет плюсом) *Любопытны любые твои комментарии:*

a. *Посмотри на сайт и на полученные данные, и предложи, как можно улучшить метрики эффективности разных каналов.*
b. *Как ты думаешь, почему каждый канал отличается по эффективности?*
c. *Достаточно ли данных для выводов, почему? Какие дополнительные метрики бизнеса или улучшения сайта ты можешь предложить и почему?*

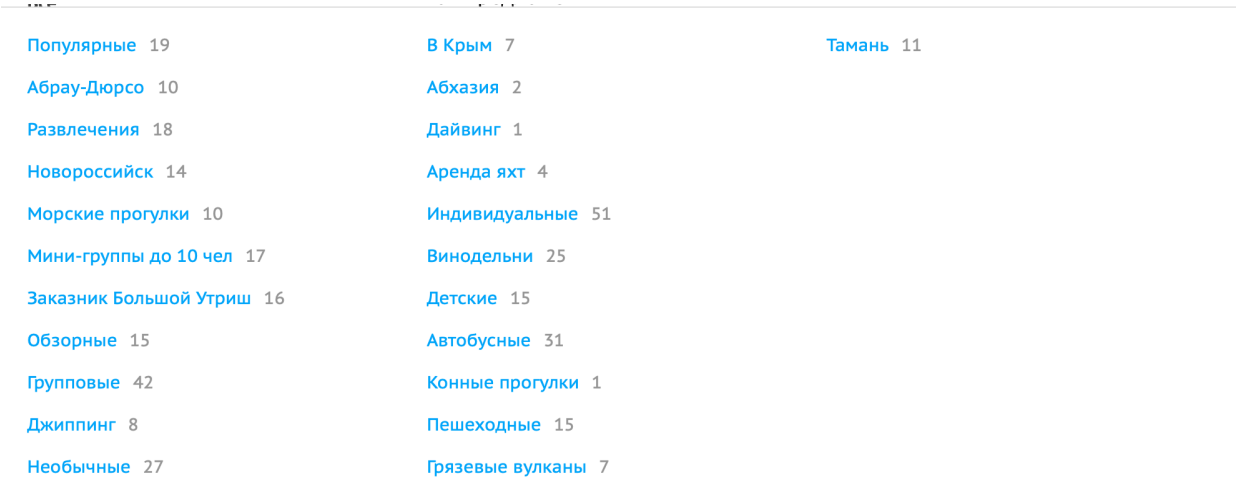
5.a. Попунктно рассмотрим каждый из каналов:

- **Органический канал** – основной поток трафика идёт именно с него, при этом конверсия визита в успешный заказ и количество просмотренных экскурсий у него наименьшая. В основном “целевая аудитория” этого канала – люди, которые едут в путешествие и хотят забронировать экскурсию.
    a. Потенциальный пользователь пользуется поисковой системой,
    b. заходит на первые попавшиеся сайты (в том числе на sputnik8.com),
    c. листает экскурсии,
    d. кликает на заинтересовавшие его.

По результатам исследования видим большое проседание в пункте d. Какие могут быть причины?

1. Объявление не цепляет (кажется неинтересным, неподходящим).
2. Посетитель теряется в большом количестве экскурсий (сложно сделать выбор, всё смывается в одно).
3. Потенциальный покупатель экскурсии не нашел нужную экскурсию (не воспользовался фильтрами, нет нужной ему фильтрации).
4. Человек искал что-то абсолютно другое, чего Sputnik8 не предоставляет.

На мой взгляд в отношении 1-го и 2-го пунктов “кликабельность” может улучшить введение ярких ТОП баннеров, например “ТОП-3 экскурсии в Анапе”, который красочно выделен на первой строке из всех экскурсий. Скорее всего, большинство нажмет на хотя бы одну из этих трех экскурсий. Также можно добавить количество купленных экскурсий на карточку экскурсии (в добавление к отзывам). Также я бы предложила более продуманную категоризацию, например, в этой раздвигающейся вкладке сделать столбец “Место экскурсии”: [Абрау-Дюрсо, Новороссийск, Заказник Большой Утриш,...], “Тип экскурсии”: [Морские прогулки, Обзорные, Джиппинг, Дайвинг, Аренда яхт, …], “Атмосфера”: [Развлечения, Необычные, (можно добавить Романтические,...)] , вкладку “Популярные” можно сделать отдельно, так как, я думаю, это самая популярная категория. Вкладки “Мини-группы” и “Индивидуальные” доступны по фильтрации по группам, но можно и повторно добавить столбец “Тип группы”.



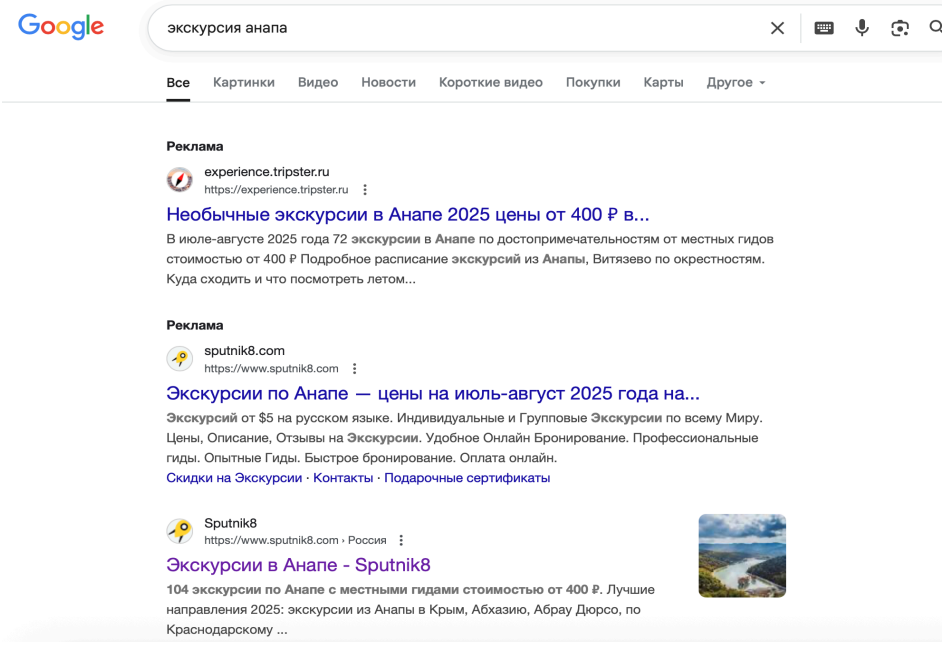
Мне кажется такие обновления помогут лучше ориентироваться пользователю на сайте и повысят метрики эффективности!

- **Партнерский канал** – по результатам исследования оказался каналом с самым высоким потенциалом. Обычно люди, переходящие по партнерским ссылкам уже “прогретая” аудитория, то есть они более заинтересованы в покупке экскурсии именно на сайте Sputnik8, чем на каком-то другом. Если грамотно подобран партнер, то его аудитория будет имеет отношение (интерес) к

путешествиям и является целевой аудиторией для покупки экскурсий, поэтому мы и получаем такие хорошие показатели конверсии.

Чтобы улучшить метрики этого канала необходимо его масштабировать: чаще запускать рекламу у блогеров/travel-инфленсеров или приглашать для сотрудничества более популярных инфлюенсеров, больше размещаться у партнеров-компаний на сайтах/в подборках (например, на сайте билетов самолетов, автобусов, на сайтах типа “10 мест Анапы, где должен побывать каждый!” что-то в таком духе :)) Фокус на расширение этого канала должен дать рост в прибыли и повысить его эффективность.

- **Cpc канал** – трафик из платной поисковой рекламы (как я понимаю, в основном, это клики на рекламные баннеры в поисковых системах - прилагаю пример на скрине).



Успех такого канала напрямую зависит от качества его настройки (правильный анализ ключевых слов, сразу ли выпадает нужная экскурсия, если поиск конкретизирован). Попробовав разные запросы, я вижу, что такая реклама появляется только при прямом запросе – “экскурсия в анапе”, при конкретизации выходит уже просто сайт без рекламы. В целом можно попробовать донастроить ключевые слова и расширить поле рекламы. Правда я не вижу в этом сильного профита, так как поисковая система хорошо справляется сама – в большинстве случаев выдает первым сайт Sputnik8, а такие клики просто забирают деньги.

5.b. В пункте 5.a достаточно подробно рассмотрены цели человека переходящего по каждому из каналов. Это как раз и есть основная причина, почему каждый канал отличается по эффективности – разная аудитория!

5.c. Безусловно существует множество метрик, данных о которых хотелось бы иметь для более широкого и точного анализа. Перечислю основные, о которых я задумывалась в процессе выполнения задания и почему они полезны:

- более длинный временной отрезок для построения более точных прогнозов,
- время посетителя на сайте, время активного просмотра экскурсии (насколько сайт удерживает внимание/интерес пользователя – где наблюдаются проседания, чтобы оценить, что можно улучшить);
- данные о зарегистрированных пользователях – кол-во новых пользователей, ”retention” – как часто пользователи возвращаются за новыми заказами, персональная информация о пользователе для оценки целевой аудитории (более детальный анализ пользователей);
- доход в цифрах от успешных заказов, допустим, средняя сумма заказа (могу предположить, что это процент от стоимости экскурсий, тогда необходимы данные об экскурсии для каждого заказа – чтобы корректно анализировать выручку компании);
- причина отмены заказа (лично пользователем или ошибка обработки оплаты).

Для начала хотелось бы рассмотреть именно эти метрики, но думаю есть еще множество, которые я не упомянула! Насчет улучшениях сайта – основные предложения описаны в пункте 5.a:)