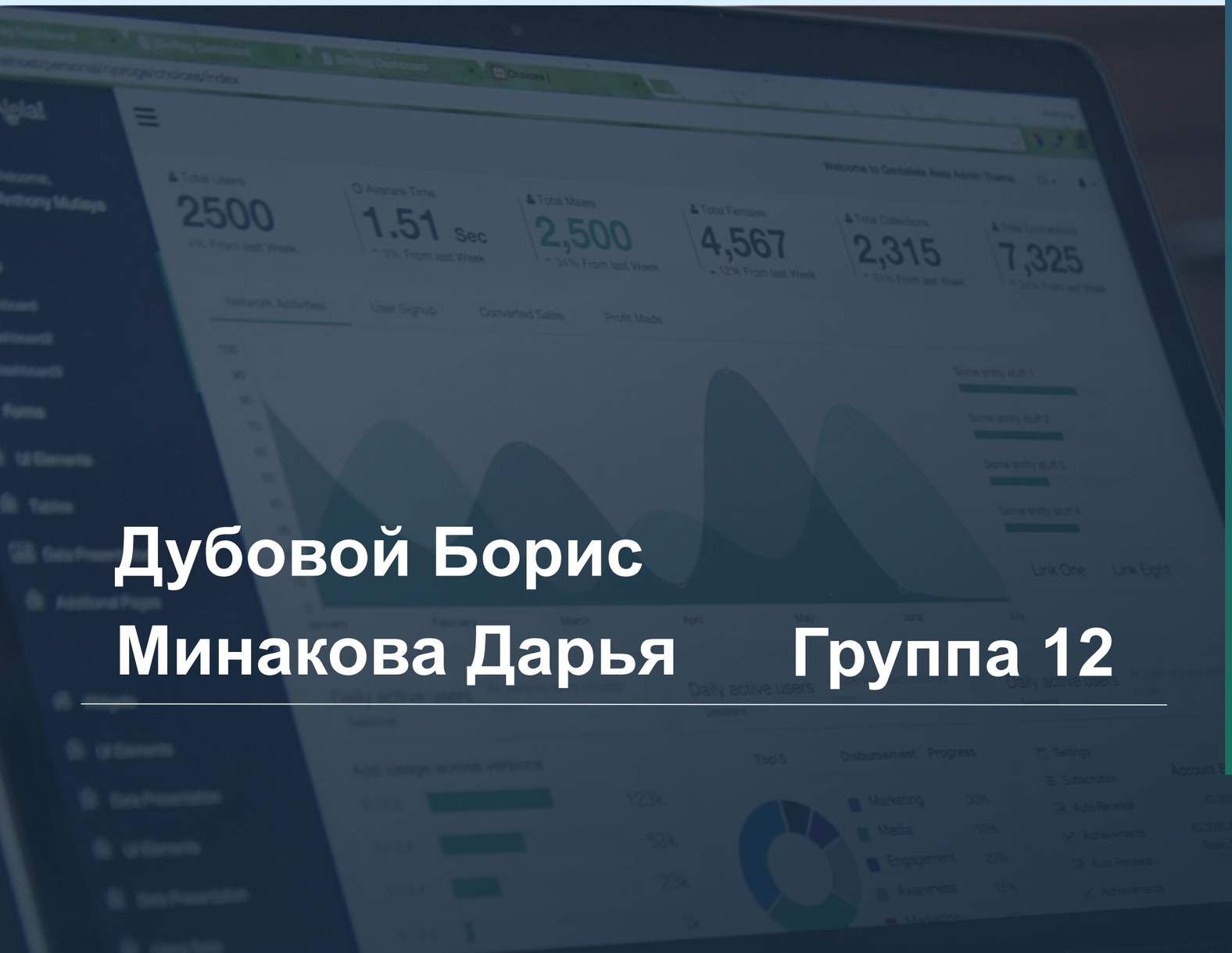
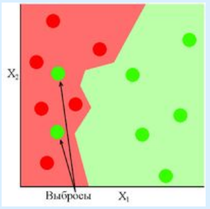


E-com анализ данных

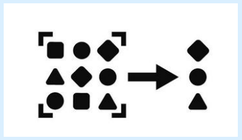
Дубовой Борис
Минакова Дарья Группа 12

Предобработка данных

Удалены пропуски у 13 пользователей. (отбор по кат. шкалам)



В колонках revenue, sessionduraion выбросы заменены на медиану по региону



Полные дубликаты удалены.

Ошибки в разных шкалах

Регион

Ошибка ввода: ‘Unjted States’ и много вариаций ‘France’

Канал

Ошибка ввода: 'контексная реклама'

Девайс

Регистр буквы: android

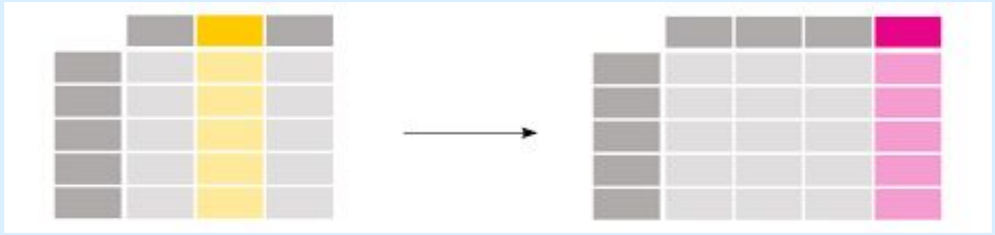
Промо-код

[0.0, 1.0, 0.8627150632632511]

Бинарная шкала

Новые колонки:

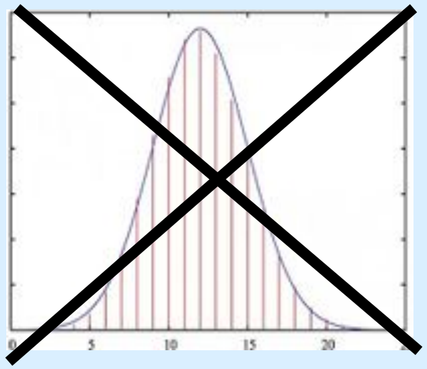
1) Сумма покупки + промо-код
2) Время суток
3) Платил ли клиент (Yes / No)



Форматирование данных

Дата и время в формате datetime

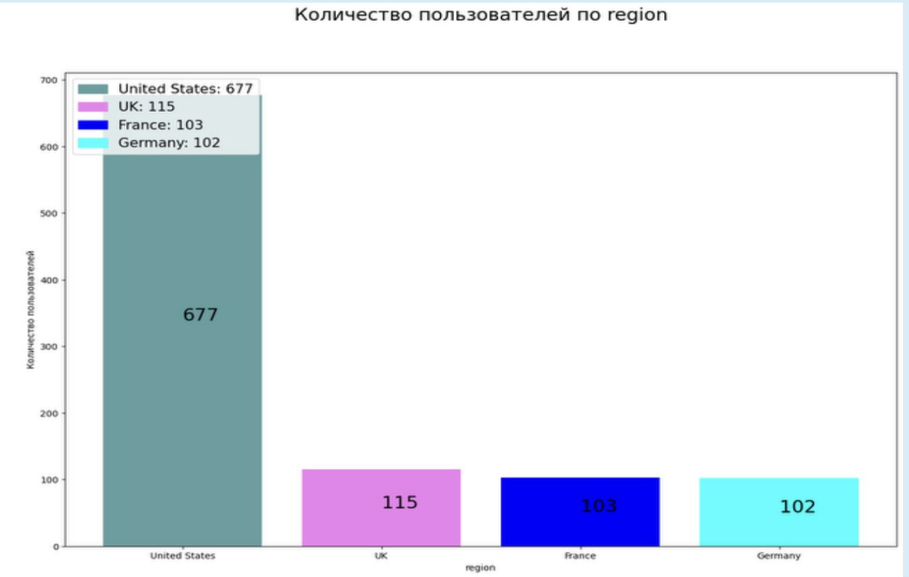
Название колонок приведен к формату PEP8

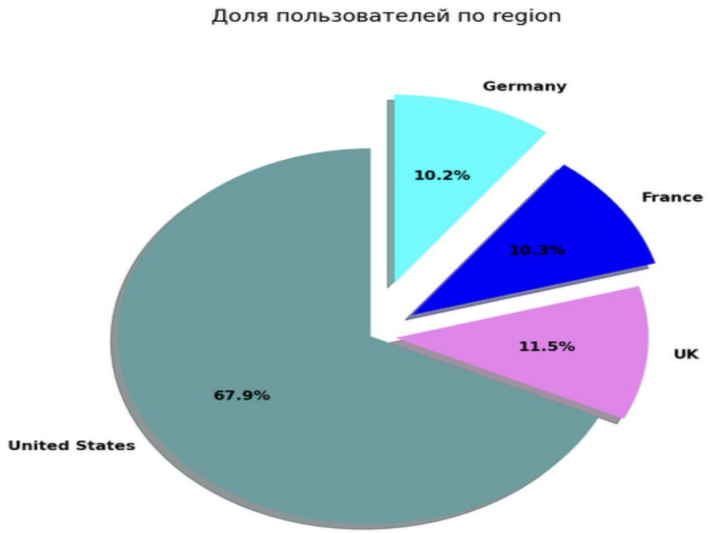


Все количественные данные распределены ненормально

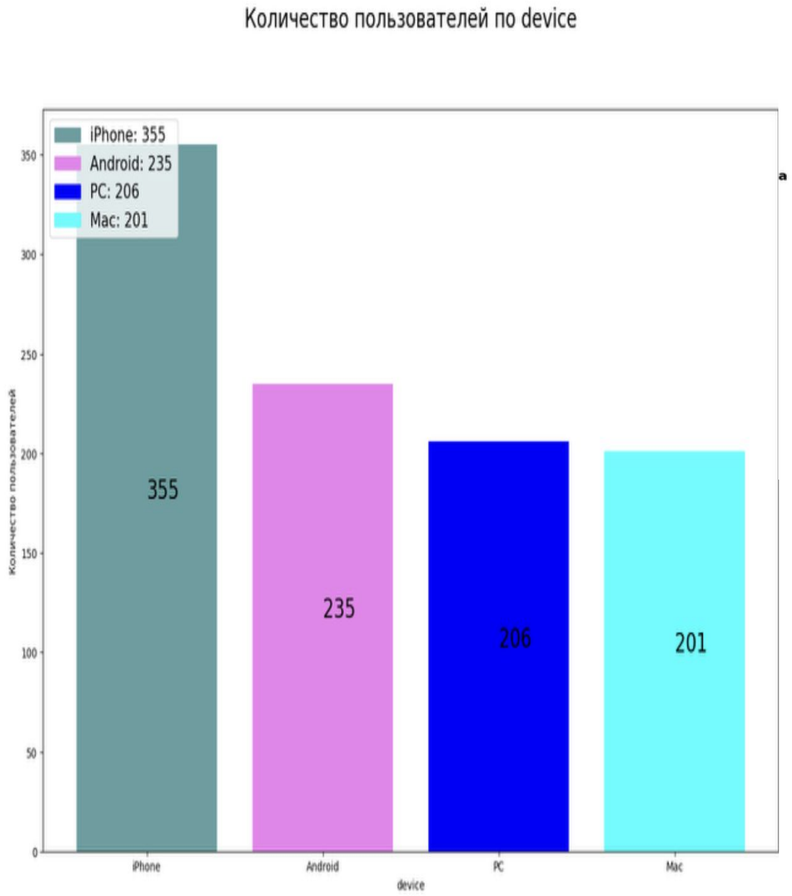
Графический анализ категориальных шкал

Регион

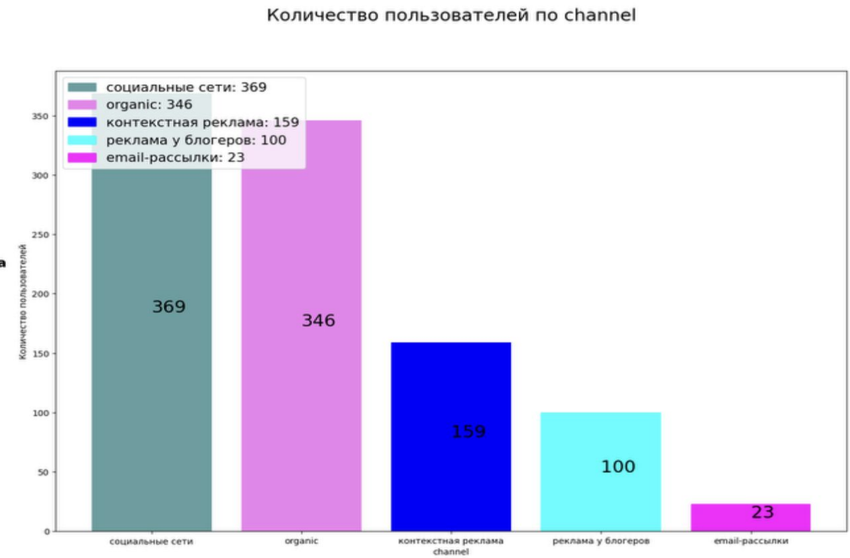




Девайс

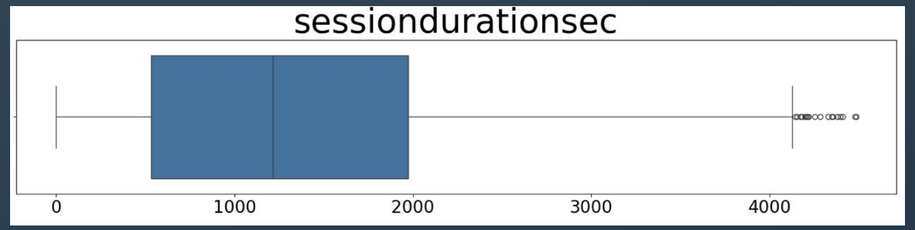


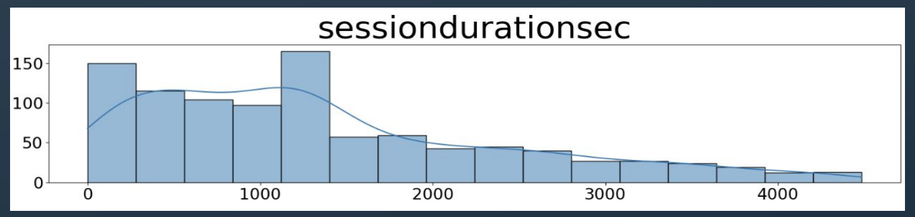
Тип канала



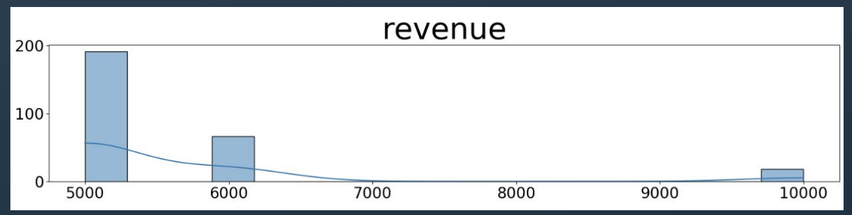
Графический анализ количественных шкал.

Длительность сессии

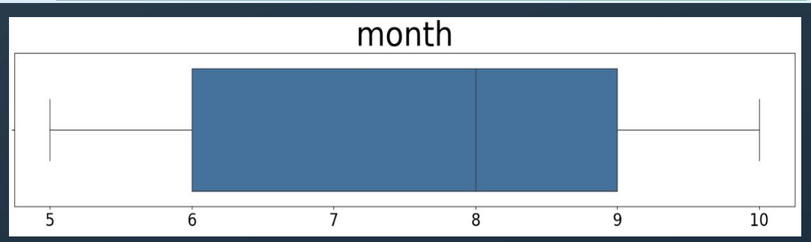


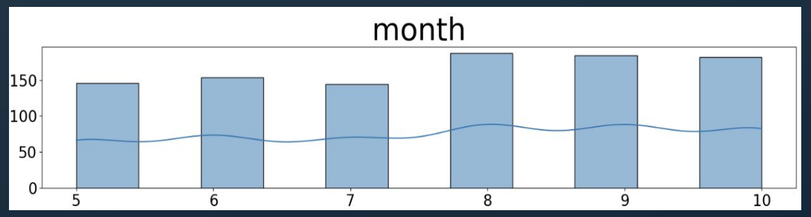


Доход

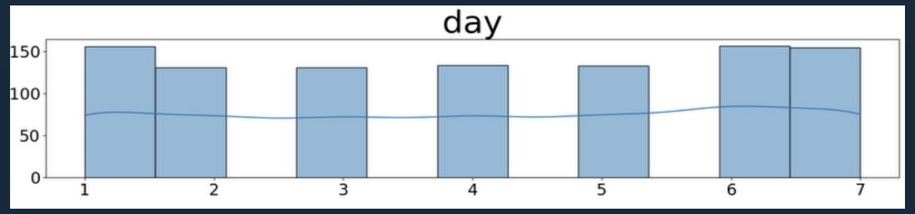


Месяц сессии

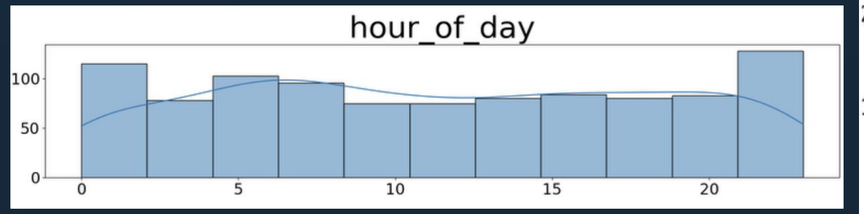




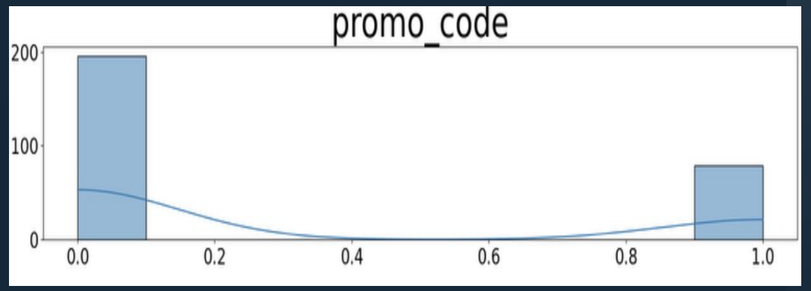
День сессии



Время дня сессии



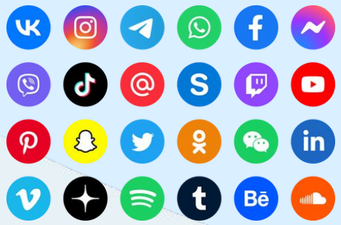
Наличие промо-кода



Основные расчеты по каналам

Соц. сети

- **Конверсия: 40%**
- **Средний чек: 5674 $**
- **Кол-во пользователей: 369**
- Доходность: 629 889 $
- Ср. время сессии: 22.5 мин



Реклама у блогеров

- **Конверсия: 11%**
- **Средний чек: 5447 $**
- **Кол-во пользователей: 100**
- Доходность: 157 971 $
- Ср. время сессии: 23.7 мин

Контекстная реклама

- **Конверсия: 15%**
- **Средний чек: 5332 $**
- **Кол-во пользователей: 159**
- Доходность: 223 958$
- Ср. время сессии: 25.4 мин

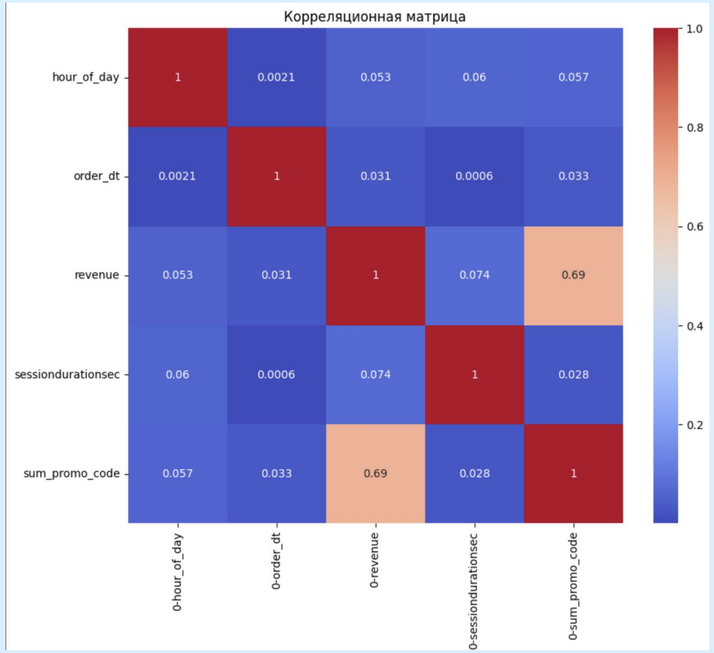
Органический

- **Конверсия: 32%**
- **Средний чек: 5585 $**
- **Кол-во пользователей: 346**
- Доходность: 485 913$
- Ср. время сессии: 21.8 мин

Количество пользователей в базе данных: 997

Всего покупок в базе данных: 275

Анализ взаимосвязанности количественных факторов

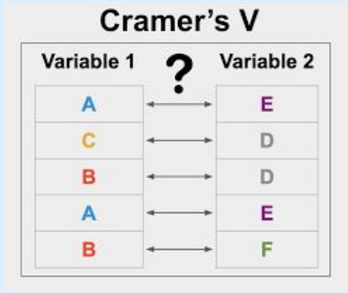


Значимых численных корреляций выявлено не было.

Общая формулировка гипотез (в одинаковых шкалах):

- **<u>Нулевая гипотеза:</u>** **Категориальные переменные независимы.**
- **<u>Альтернативная гипотеза:</u>** **Категориальные переменные связаны.**

Анализ взаимосвязанности категориальных факторов



Все категориальные факторы в датасете: ['region', 'device', 'channel', 'month', 'day', 'payment_type', 'promo_code', 'time_of_day']

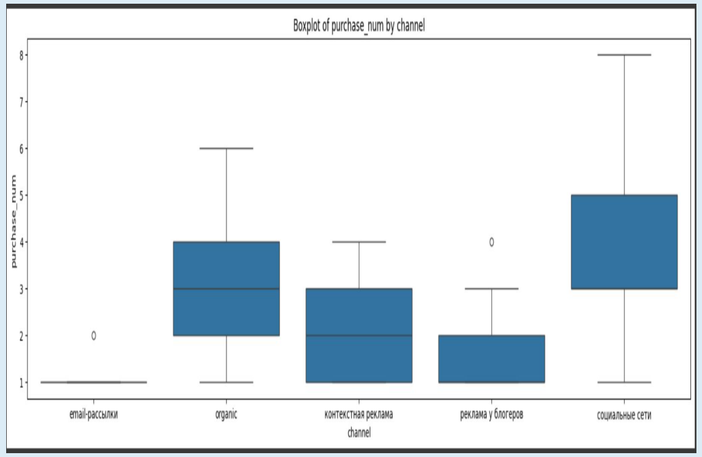
Замеченная средняя корреляция:

'region' воздействует на 'channel' с коэффицентом 0.4042222857755066

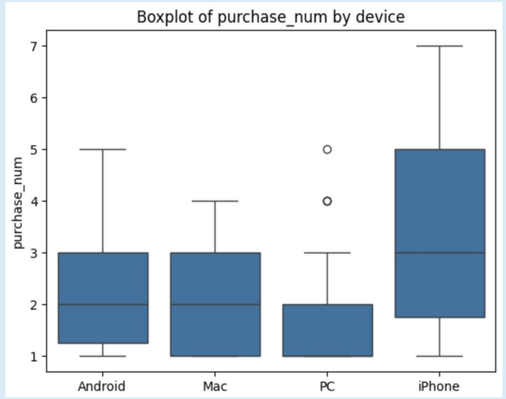
Общая формулировка гипотез (в разных шкалах):

*** <u>Нулевая гипотеза:</u> Средние значения {количественной переменной} в группах {категориальной переменной} одинаковы, нет статистически значимых различий между группами.**

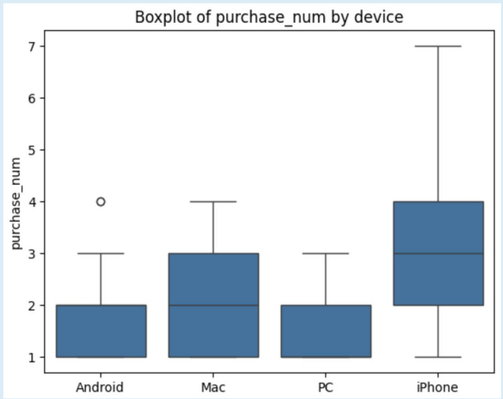
*** <u>Альтернативная гипотеза:</u> Средние значения {количественной переменной} в группах {категориальной переменной} различны, существуют статистически значимые отличия между группами.**



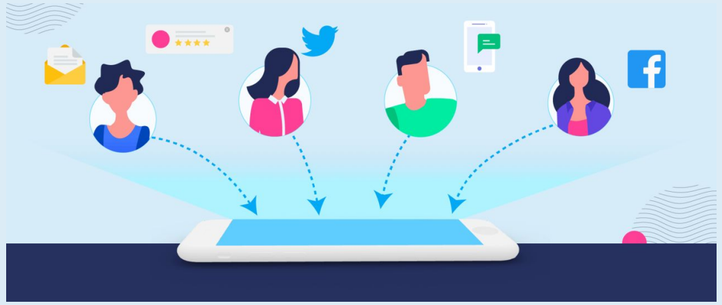
Фактор channel влияет на количество покупок в день (Отвергаем нулевую гипотезу)



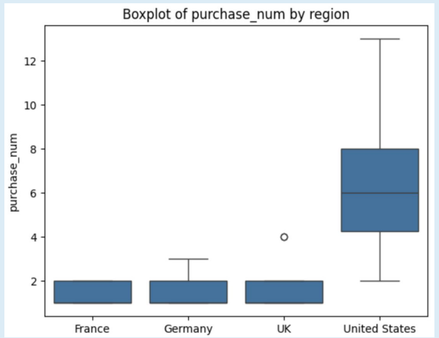
Фактор device влияет на количество покупок в день (Отвергаем нулевую гипотезу)



United States:
Фактор device влияет на количество покупок в день в отобранном сегменте



France: Количественный фактор channel имеет влияние на количество покупок в день недели в отобранном сегменте



Фактор region влияет на количество покупок в день (Отвергаем нулевую гипотезу)

Проверка гипотез

Подтвержденные альтернативные гипотезы (факторы взаимосвязаны)

Регрессионная модель

Важные колонки:

region, device, sessiondurationsec

Наилучшие модели, по результатам rmse и r^2, состоят из колонок:

['region', 'device', 'sessiondurationsec', 'day', 'sum_promo_code'] -> rmse = 1017.88, r^2 = 0.137

['region', 'sessiondurationsec', 'day', 'sum_promo_code'] -> rmse = 1019.36, r^2 = 0.134

['region', 'day', 'sum_promo_code'] -> rmse = 1020.21, r^2 = 0.133

['day', 'sum_promo_code'] -> rmse = 1023.70, r^2 = 0.126

['sessiondurationsec', 'day', 'sum_promo_code'] -> rmse = 1024.69, r^2 = 0.125

Количество пользователей
997

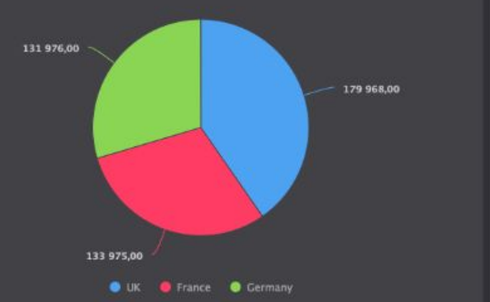
Количество платящих пользователей
275

Средняя продолжительность сессии
22 минуты

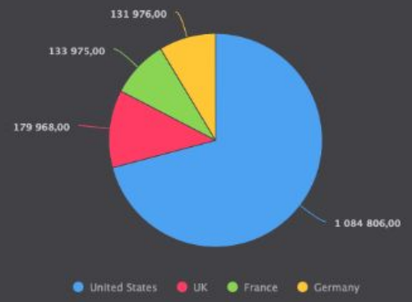
Средний чек
5 566,3

Сумма продаж
1 530 725

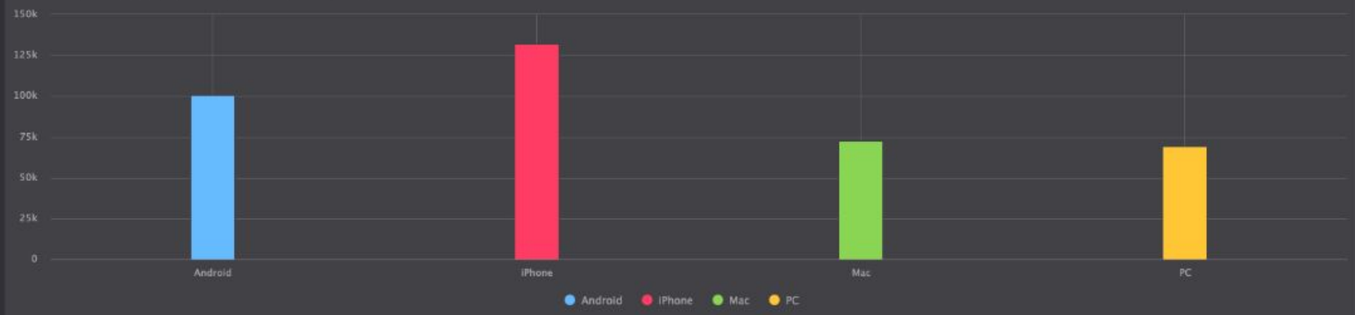
Распределение доходов в Европе



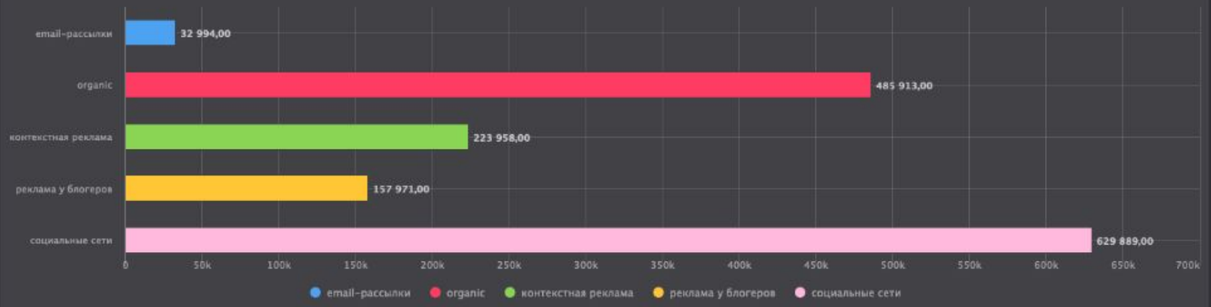
Распределение доходов в зависимости от региона



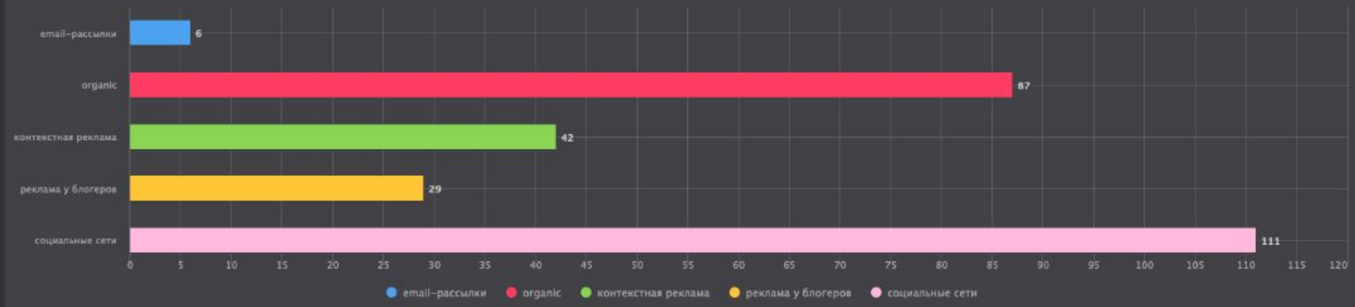
Зависимость деятельности сессии от устройства



Сумма покупок в зависимости от вида рекламы



Количество покупок в зависимости от вида рекламы



Дашборд

Дашборд


Сумма продаж по месяцам
450k
400k
350k
300k
250k
200k
150k
100k
50k
0
5
6
7
8
9
10
France
Germany
UK
United States

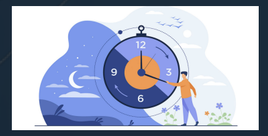
Общий вывод после анализа.

Фокус на ключевые регионы:

США привлекает самое большое количество пользователей, как в общем числе, так и среди платящих клиентов. Развитие бизнеса в этом регионе, возможно, будет наиболее результативным.

Сезонные акции и промокоды:

Лето - самый активный период для покупок. Акции и промокоды в этот период могут стимулировать продажи.



Оптимизация времени суток:

Ночью совершается наибольшее количество покупок. Учитывая это, добавление ночной темы на сайт может привлечь больше пользователей.

Оптимизация под устройства:

iPhone является наиболее популярным устройством среди пользователей, стоит рассмотреть оптимизацию сайта под продукцию Apple для улучшения пользовательского опыта.

Эффективные рекламные каналы:

Реклама в социальных сетях и органическая канал показали наилучшие результаты. Наилучшая конверсия (40%) достигается в соц.сетях , но при этом затраты тоже довольно большие, в среднем 5000 рублей в одном канале (согласно интернету), в то время конверсия органического сектора - 32%, при этом компания ничего не тратит. Возможно аналитикам стоит пересмотреть вложения.