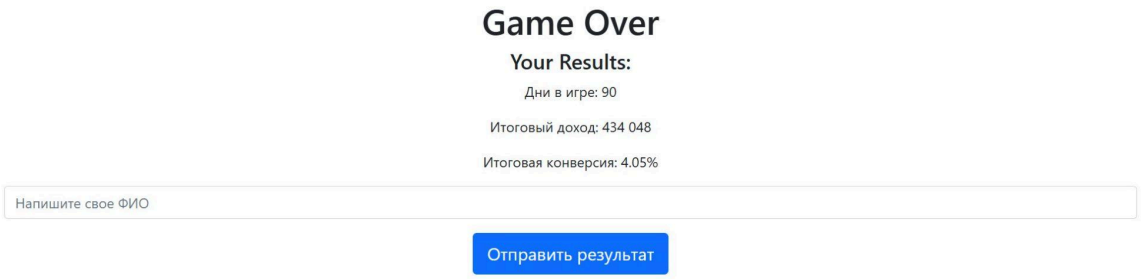
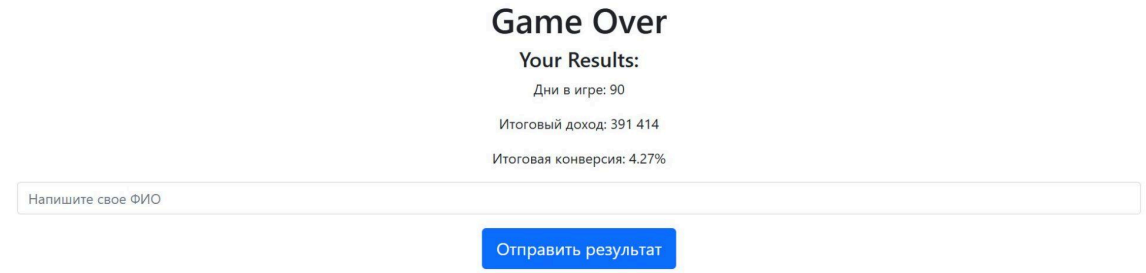
Домашнее задание по А/Б тестам.

Итоговые выводы внизу документа)

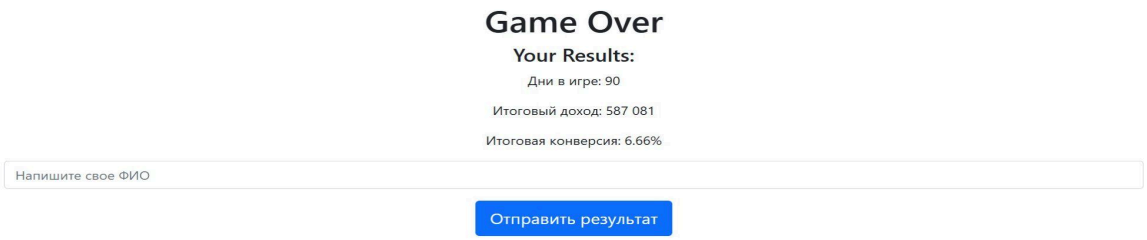
1. Стратегия “рандом” или как делать не надо.
   Рандомно берем задачи в разработку.
   Условие для внедрения эксперимента для всех пользователей: p-value < 5%. То есть, мы смотрим только на p-value, не обращаем внимание на то, сколько дней идет эксперимент и на конверсию. Метрика “прокрасилась” - выкатываем изменение в релиз. Рандомно отменяем эксперименты: можем отменить эксперимент через 1 день, а можем - через 20.

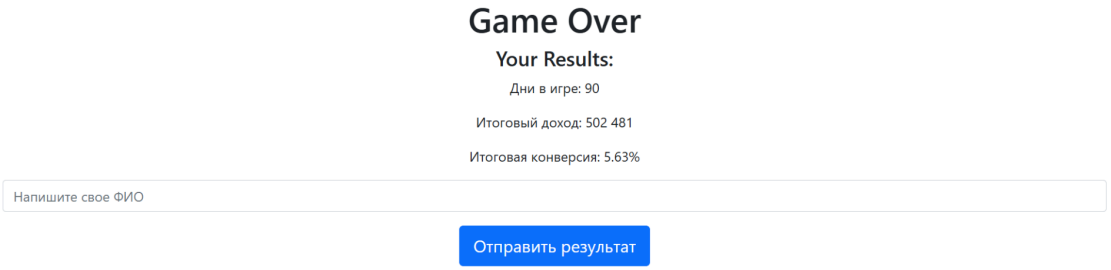


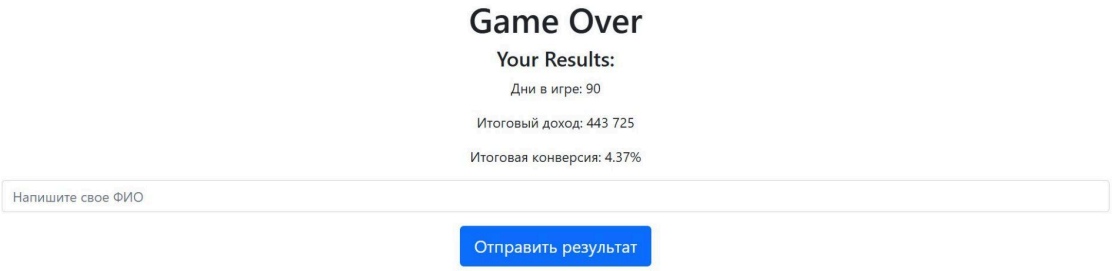
   Доход: 434048 < 450000. Результат плохой.

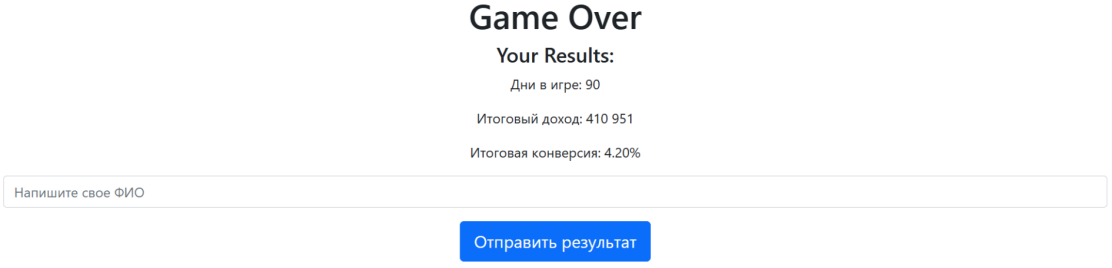


   Доход: 391414 < 450000. Результат еще хуже.
2. Немного совершенствуем стратегию.
   Стараемся брать по 3 задачи, выполнение которых занимает одинаковое время (так удобнее играть).
   Все a/b тесты запускаем на 7 дней.
   Если 7 дней прошло и p-value < 5 %, выкатываем эксперимент на всех.
   Если 7 дней прошло и p-value > 5% отменяем эксперимент.
   Все еще игнорируем конверсию.







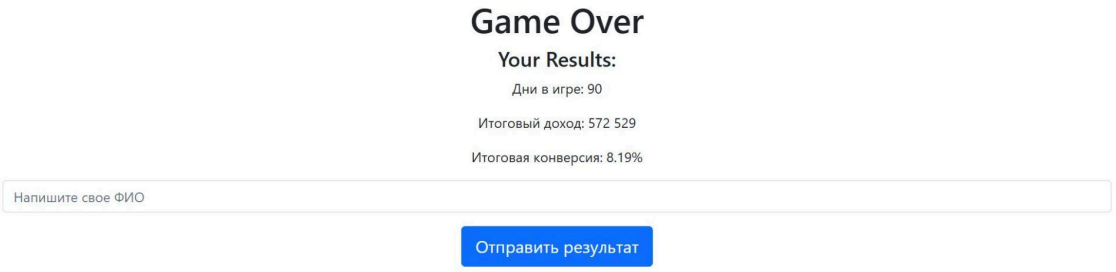


Получаем довольно разные результаты. Видим, что конверсия сонаправлена с доходом.

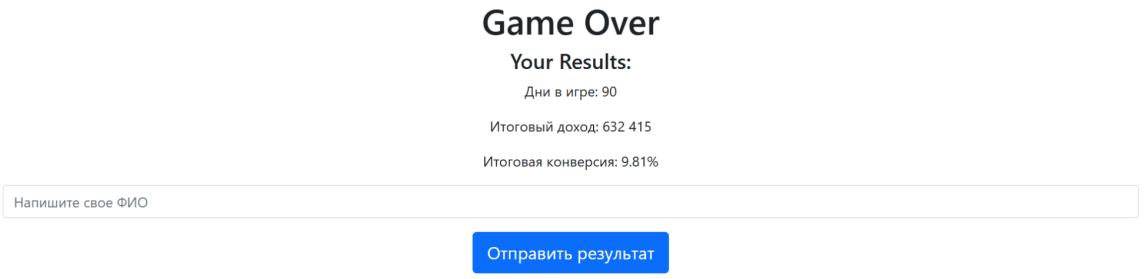
Очевидно, что если мы просто смотрим на p-value, когда релизим эксперимент на всю аудиторию, то мы можем релизнуть тот эксперимент, который влияет негативно. Поэтому получаются настолько разные результаты.

3. Смотрим не только на p-value, но и на конверсию.

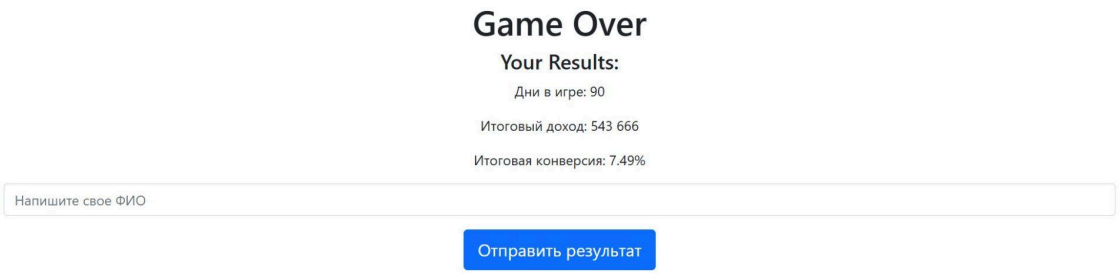
3.1. Также стараемся брать по 3 задачи, выполнение которых занимает одинаковое время.
A/B тест делаем на 7 дней. Выводим эксперимент в релиз на всю аудиторию, если p-value меньше 5% и конверсия в тестовой группе больше, чем в контрольной. Если же хотя бы одно из этих условий не выполняется, то останавливаем a/b тест.



Наш доход 572529. Это на 122529 больше, чем base line. Также у нас довольно высокая конверсия - 8.19%.
3.2. Закрепим результат. Сыграем по предыдущей стратегии еще раз.



Наш доход 632415. Это на 182415 больше, чем base line. Также наша конверсия равна 9.81%!
3.3. Теперь попробуем делать a/b тест не на 7 дней, а на 10.



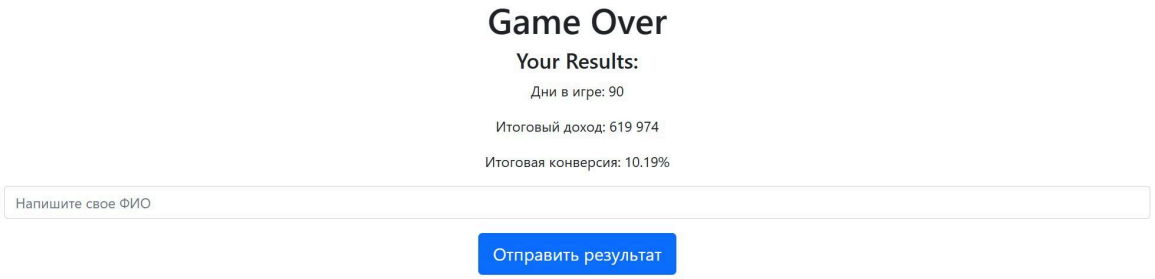
Наш доход 543666 и конверсия 7.49%. По деньгам мы заработали на 93666 больше, чем base line - неплохой результат. Но он ниже того варианта, где a/b тест был на 7 дней.

3.5. Теперь попробуем брать 1% уровень значимости вместо 5%. Также запускаем a/b тест на 7 дней.

Если мы снижаем уровень значимости с 5% до 1%, то у нас:

Меньше ошибок 1 рода (Меньше вероятность ложноположительного результата).

Выше порог значимости, следовательно, сложнее найти эффект.



Наш доход 619974 и конверсия 10.19%! Хороший результат.По деньгам мы заработали на 169974 больше, чем base line. Можно предположить, что конверсия вышла больше из-за того, что мы “отфильтровали” слабые эффекты.

Оптимальная стратегия:

Стараемся брать по 3 задачи, выполнение которых занимает одинаковое время (так удобнее играть, но можно и рандомно брать задачи). Единственное, лучше к концу нашего периода (90 дней) не давать разработчикам большие задачи, потому что они долго делаются, и мы просто впоследствии не успеем завершить выполнение a/b теста до окончания периода.

A/B тест делаем на 7 дней.

Уровень значимости можно взять 5% или 1%.
Выводим эксперимент в релиз на всю аудиторию, если p-value меньше уровня значимости и конверсия в тестовой группе больше, чем в контрольной. Если же хотя бы одно из этих условий не выполняется, то останавливаем a/b тест.

Выводы:

- Для успешной игры нужно смотреть на p-value и конверсию.
- 7 дней в нашей игре кажется оптимальным для выполнения a/b теста.
- Если через 7 дней эксперимента p-value меньше уровня значимости (отвергается нулевая гипотеза о том, что изменений нет) и конверсия в тестовой группе больше, чем в контрольной, мы выкатываем фичу в релиз на всех пользователей. Так как получается, что разница между тестовой и контрольной группами значима и влияния эксперимента положительно.
- Если через 7 дней p-value меньше уровня значимости, но конверсия в тестовой группе меньше, чем в контрольной, то получается, что у нашего эксперимента отрицательный эффект, мы не выкатываем фичу в релиз и прекращаем a/b тест.
- Если через 7 дней p-value чуть больше уровня значимости и конверсия в тестовой группе больше, чем в контрольной, то у нас есть большой соблазн подождать 1-2 для и, возможно, что p-value станет меньше уровня значимости. И вроде как мы можем выкатить эксперимент в релиз. Но так делать не надо, потому что если мы будем дальше просто ждать и продолжать смотреть на p-value, то рано или поздно случайные колебания дадут нам "значимый" результат, даже если на самом деле эффекта нет. Что приведет к ложным выводам (я прочитала, что у такого процесса даже есть название - "p-hacking" или "ловля значимости").

P.S. Спасибо за такое задание, было правда очень интересно его выполнять и в более “прикладном” смысле познакомиться с a/b тестами!