**Внимание!** Для выполнения тестовых заданий скачайте и откройте массив данных по ссылке:

<https://docs.google.com/spreadsheets/d/1EOEmGcBpokRfYbiNBDQs5XnWG9QGmOSwYKpKiOkhQR4/edit?usp=sharing>

1. Во вкладке "Данные об аудитории" информация о пользователях, посетивших наше приложение в ноябре. Чему равен MAU продукта?

\***MAU (Monthly Active Users)** — это метрика, используемая для измерения активности пользователей в течение одного месяца. Она показывает количество уникальных пользователей, которые взаимодействовали с продуктом, сервисом или приложением хотя бы один раз за последний месяц.

7639168141048216529

2. Используя вкладку "Данные об аудитории", посчитайте, чему будет равен DAU

\***DAU (Daily Active Users)** — это метрика, которая показывает количество уникальных пользователей, которые взаимодействовали с продуктом, приложением или сервисом хотя бы один раз в течение дня. DAU помогает понять, сколько пользователей активно пользуются продуктом каждый день.

255490560483

3. Используя вкладку "Данные об аудитории", посчитайте, чему будет равен retention первого дня у пользователей, пришедших в продукт 1 ноября

\*Retention (удержание пользователей) — это метрика, которая показывает, сколько пользователей продолжает пользоваться продуктом через определенный промежуток времени после первоначального взаимодействия. Retention можно рассчитать как процент пользователей, вернувшихся в продукт через определенное время (например, через 1 день, 1 неделю, 1 месяц) от количества всех новых пользователей.

28,3%26,6%38,5%32,7%

4. На графике изображены retention кривые 2 продуктов. Какие выводы можно сделать, глядя на них?



Ваш ответ:

Сравнение удержания:

Синий продукт имеет лучшее удержание пользователей на всех этапах (день 1 - день 7).

Красный продукт быстро теряет пользователей после дня 1.

Темп снижения:

У синего продукта темп снижения замедляется после дня 3, стабилизируясь около 30%.

У красного продукта снижение продолжается до почти 0% к дню 6.

Рекомендации:

Синему продукту следует изучить, что помогает удерживать пользователей после 3-го дня, и усилить эти механики.

Красному продукту стоит улучшить первые взаимодействия с пользователями (например, обучение или вовлечение).

Теперь напишу код для анализа и построения подобных кривых.

5. Во вкладке "Данные об аудитории" есть информация о том, сколько объявлений посмотрел каждый пользователь (view\_adverts). Посчитайте пользовательскую конверсию в просмотр объявления за ноябрь? (в пользователях)

\* Пользовательская конверсия — это метрика, которая показывает, какой процент пользователей выполнил целевое действие по отношению к общему количеству пользователей. В контексте веб-сайтов это может быть действие, такое как просмотр объявления или клик по рекламному баннеру.

41,8%54,7%46,3%39%

6. Используя информацию из вкладки "Данные об аудитории", посчитайте среднее количество просмотренных объявлений на пользователя в ноябре

4,96,25,32,9

7. Мы провели опрос среди 2000 пользователей. Из них 500 «критики», 1200 «сторонники» и 300 «нейтралы». Посчитайте, чему будет равен NPS

\*NPS (Net Promoter Score) — это метрика, которая измеряет лояльность пользователей к компании или продукту и делит их на три группы: Сторонники (Promoters) , Нейтралы (Passives), Критики (Detractors). NPS высчитывается как (% сторонников - % критиков).

30%43%40%35%

8. Во вкладке "Данные АБ-тестов" результаты трех несвязанных АБ тестов для ARPU (общая выручка/общее количество пользователей).  
Посмотрите на результаты тестов и интерпретируйте их. Напишите значения p-value, которые вы получили.  
Подготовьте выводы и рекомендации.   
  
experiment\_num - номер эксперимента  
experiment\_group - группа, в которую попал пользователь  
user\_id - id пользователя  
revenue - выручка, которую сгенерировал пользователь, купив платную услугу продвижения

Ваш ответ:

Эксперимент 1: p-value = 0.688966190624939

Эксперимент 2: p-value = 0.0011282266247294027

Эксперимент 3: p-value = 0.06031548002841902

9. По датасету с листерами посчитайте средний доход на пользователя

121.2156.470.930.7средняя здесь не применима

**158.46**

10. По датасету с листерами посчитайте медиану возраста пользователя

27,422827,9327медиана здесь не применима

11. Какой график лучше всего подходит для отображения разброса цен на товары в разных магазинах?  
\*возможно несколько вариантов ответа

Линейный графикКруговая диаграммаЯщик с усами (box plot)

Гистограмма  
Лучшая диаграмма для отображения разброса (дисперсии) цен в разных магазинах — это диаграмма ящиков (диаграмма ящиков и усов). Она наглядно визуализирует медиану, квартили и выбросы. Гистограмма также может быть полезна, показывая распределение цен.  
Необходимо заменить заполнители, такие как «income\_column», «age\_column», названия листов и имя файла Excel, вашими фактическими данными. Вам может потребоваться скорректировать статистические тесты в вопросе 8 в зависимости от структуры данных вашего AB-теста (например, используя ANOVA, если у вас больше двух групп). После запуска этого кода с вашими данными вы получите ответы и сможете сделать соответствующие выводы.

12. На каком графике бимодальное распределение?

№1

№2

№3

* **График №1:** Это нормальное (гауссово) распределение, имеющее один пик.
* **График №2:** Это unimodal (одномодальное) распределение со скошенностью вправо (положительная асимметрия).
* **График №3:** Это приблизительно unimodal (одномодальное) распределение с умеренной асимметрией вправо (положительная асимметрия). Может показаться, что есть два пика, но это скорее следствие малого объема выборки и группировки данных в гистограмме. На самом деле это, скорее всего, одномодальное распределение.

**Бимодальное распределение** характеризуется наличием *двух* отчетливых пиков (мод). Ни один из представленных графиков не соответствует этому описанию.

№4

На данном графике бимодального распределения нет. График показывает тенденцию к положительной корреляции, то есть, по мере увеличения значений на оси X, значения на оси Y (продажи) также растут. Однако, это не бимодальное распределение, так как отсутствуют два отдельных, четко выраженных пика.

That scatter plot shows a **positive correlation** between the independent variable (which is not labeled on the x-axis) and sales.

Here's why:

* **Upward Trend:** As the values on the x-axis increase, the sales values (on the y-axis) also tend to increase. This general upward trend indicates a positive relationship.
* **Not Perfect:** The correlation isn't perfect; the points aren't exactly on a straight line. There's some scatter or variability, which is typical in real-world data. A perfect positive correlation would have all points falling precisely on a straight, upward-sloping line.

**In summary:** While not perfectly linear, the overall pattern demonstrates that higher values on the x-axis are associated with higher sales figures. To quantify the strength of this positive correlation, you would need to calculate a correlation coefficient (like Pearson's r).

13. Какая случайная величина имеет наибольшую дисперсию данных по следующим графикам плотности распределения?

Наибольшую дисперсию имеет случайная величина, представленная на **графике №3**.

Дисперсия отражает степень разброса данных вокруг среднего значения. На графике №3 данные распределены по значительно более широкому диапазону значений по сравнению с остальными гистограммами. Это указывает на то, что дисперсия данных на этом графике больше, чем на остальных. Графики 1, 2 и 4 демонстрируют более узкое, компактное распределение данных, что говорит о меньшей дисперсии.

№1

№2

№3

№4

14. На каком графике можно посчитать коррелцияю?  
\*возможно несколько вариантов ответа

Корреляцию можно посчитать по двум графикам:

1. **Scatter Plot (Диаграмма рассеяния):** Этот тип графика напрямую отображает взаимосвязь между двумя переменными (X и Y). Наблюдаемая кластеризация точек (как в данном случае, где точки образуют примерно восходящую линию) указывает на корреляцию. Для количественного определения силы и направления корреляции используют коэффициент корреляции (например, Пирсона).
2. **Correlation Heatmap (Тепловая карта корреляции):** Этот график непосредственно показывает коэффициенты корреляции между переменными. В данном примере тепловая карта показывает корреляцию между переменными X, Y, и Z. Чем темнее красный цвет, тем сильнее положительная корреляция; чем темнее синий — тем сильнее отрицательная.

Линейный график (Line Plot) не подходит для оценки корреляции между двумя переменными, так как он отображает изменение одной переменной во времени.

thumb\_upthumb\_down

Корреляцию можно посчитать, используя **тепловую карту корреляции (Correlation Heatmap)**. Эта карта напрямую показывает коэффициенты корреляции между переменными X, Y и Z. Значения в ячейках — это коэффициенты корреляции, которые варьируются от -1 (совершенная отрицательная корреляция) до +1 (совершенная положительная корреляция). Значение 0 указывает на отсутствие линейной корреляции.









15. Что значит, если при проверке гипотез мы получили p-value = 0.05?

Это означает, что нет никакой статистически значимой разницы между группами

Есть 5% вероятность случайно получить такой или еще более экстремальный результат, если нулевая гипотеза верна

Это означает, что результаты эксперимента на 95% точны

Это говорит о том, что альтернативная гипотеза верна с вероятностью 95%

**Есть 5% вероятность случайно получить такой или еще более экстремальный результат, если нулевая гипотеза верна.**

**Объяснение**:

* p-value = 0.05 означает вероятность 5% того, что наблюдаемые результаты (или более экстремальные) могут произойти случайно при условии, что **нулевая гипотеза (H₀)** верна.
* Это не означает, что результаты эксперимента точны на 95%, и это не подтверждает альтернативную гипотезу с вероятностью 95%. Это просто уровень статистической значимости.

16. Какой метод наиболее подходит для проверки гипотезы о равенстве средних двух выборок из нормального распределения?

t-тестХи-квадрат тестАнализ дисперсии (ANOVA)Корреляция Пирсона

**Объяснение**:

* **t-тест** используется для проверки равенства средних двух выборок из нормального распределения.
* **Хи-квадрат тест** применяется для категориальных данных.
* **ANOVA** используется для сравнения средних более чем двух групп.
* **Корреляция Пирсона** измеряет линейную зависимость между двумя переменными.

17. Как интерпретировать квартили в распределении доходов пользователей?

Показывают максимальный и минимальный доход

Делят данные на четыре равные части

Указывают на наиболее часто встречающийся доход

График плотности распределения вещества во вселенной

**Объяснение**:

* Квартили — это точки, которые делят данные на 4 равные части, каждая из которых содержит 25% наблюдений.
* Например, 25-й процентиль (1-й квартиль) показывает, где находятся 25% самых низких значений.

18. Были получены следующие результаты. Коллеги просят вас подтвердить их и сделать окончательный вывод по эксперименту.

* + - * Вариант A (контрольная группа) — 100 047 501 посетитель, 1003 платежа.
      * Вариант B (тестовая группа) — 100 001 055 посетителей, 1099 платежей.

Какие рекомендации вы бы дали, основываясь на этих данных?

Ваш ответ:

**Рекомендация**: Z-статистика: 2.10

p-значение: 0.035

Различие статистически значимо. Вариант B лучше.

* Если p-value < 0.05: разница статистически значима, можно применять изменения.
* Если p-value ≥ 0.05: разница незначима, требуется больше данных для подтверждения.

Для того, чтобы дать рекомендации по результатам А/В теста, необходимо провести статистический анализ, чтобы определить, является ли наблюдаемое различие в конверсии статистически значимым или просто случайным. Простой сравнения количества платежей недостаточно.

**Необходимые шаги для анализа:**

1. **Рассчитать конверсию:** Вычислите конверсию (CR - Conversion Rate) для каждой группы:
   * **Вариант A:** CR\_A = (1003 платежа / 100047501 посетитель) \* 100% ≈ 0.001%
   * **Вариант B:** CR\_B = (1099 платежа / 100001055 посетитель) \* 100% ≈ 0.0011%
2. **Провести статистический тест:** Используйте статистический тест, подходящий для сравнения двух пропорций (конверсий). Наиболее подходящим будет **Z-тест** или **хи-квадрат тест**. Эти тесты определят, является ли наблюдаемое различие в конверсии статистически значимым, учитывая размеры выборок. Нужно учесть, что эти выборки очень большие.
3. **Установить уровень значимости (альфа):** Обычно используется уровень значимости α = 0.05. Это означает, что существует 5% вероятность отвергнуть нулевую гипотезу (нет различия между вариантами), когда на самом деле она верна.
4. **Интерпретация результатов:** Если p-значение (полученное в результате Z-теста или хи-квадрат теста) меньше уровня значимости (α = 0.05), то различие в конверсии статистически значимо. В этом случае можно сделать вывод, что вариант B (тестовая группа) показывает лучшие результаты.

**Пример использования Z-теста (приблизительный, требует библиотеки stats):**

Этот расчет требует использования статистического пакета, например, scipy.stats в Python. Поскольку числа очень большие, приближение с использованием Z-теста допустимо.

**Рекомендации:**

До тех пор, пока не проведен статистический тест, окончательные рекомендации дать невозможно. Если тест покажет статистически значимое улучшение конверсии в варианте B, то рекомендуется:

* **Перейти на вариант B:** Запустить вариант B в качестве основного варианта.
* **Дальнейший анализ:** Провести более детальный анализ, чтобы понять, почему вариант B показал лучшие результаты. Какие элементы дизайна или функциональности повлияли на увеличение конверсии?

Если тест покажет, что различие не значимо, то:

* **Дальнейшее тестирование:** Продолжить A/B тестирование с большим объемом трафика или с другими вариантами.
* **Пересмотреть гипотезу:** Пересмотреть предположения и гипотезы, лежащие в основе эксперимента. Возможно, нужный эффект слишком мал, чтобы его достоверно обнаружить при текущем объеме трафика.

**Важно:** Необходимо использовать статистический пакет для более точного анализа, чем представленный здесь приближенный расчет. Этот пример показывает только общий подход.