**«Аналитик данных»**

**Пояснительная записка**

Выполнил

Вислов Иван

**Содержание**

[1. Исходные данные 3](#_Toc134959696)

[2. Задача 4](#_Toc134959697)

[3. Описание АБ-Теста 6](#_Toc134959698)

[3.1 Цель эксперимента 6](#_Toc134959699)

[3.2 Дизайн эксперимента 6](#_Toc134959700)

[3.3 Сплит-система 6](#_Toc134959701)

[3.4 Таргет-метрики 6](#_Toc134959702)

[3.5 Описание данных 6](#_Toc134959703)

[4. Инструменты 8](#_Toc134959704)

[4.1 Excel 8](#_Toc134959705)

[4.2 Jupyter 8](#_Toc134959706)

[4.3 Microsoft Word (MS Word) 9](#_Toc134959707)

[5. Теоретическая часть 10](#_Toc134959708)

[6. Решение 12](#_Toc134959709)

[6.1 Импорт и анализ таблиц 12](#_Toc134959710)

[6.2 Объединение таблиц 14](#_Toc134959711)

[6.3 Автоматизация статистических вычислений 15](#_Toc134959712)

[6.4 Чистка неверно заполненных точек 15](#_Toc134959713)

[6.5 Расчет общих результатов АБ Теста 16](#_Toc134959714)

[6.6 Сегментация результатов АБ Теста 17](#_Toc134959715)

[6.7 Отчет по АБ Тесту 17](#_Toc134959716)

[6.8 Калькулятор 18](#_Toc134959717)

[7. Выводы 19](#_Toc134959718)

[Приложение А. Датасет](#_Toc134959719)

[Приложение Б. Python](#_Toc134959720)

[Приложение В. Калькулятор SkyLenta](#_Toc134959721)

# Исходные данные

Вы работаете аналитиком данных в крупном ритейлере SkyLenta, который присутствует во многих российских городах.

Вы работаете в отделе продуктовой аналитики, и ваша текущая задача - проанализировать АБ-Тест, проведенный во всех городах.

# Задача

Вам необходимо проанализировать и визуализировать результаты, провести сегментацию, а также сделать выводы и сформулировать рекомендации для дальнейших запусков АБ Теста.

Также надо построить таблицу, которая будет в удобной форме хранить результаты АБ Теста.

Таблица, которая должна быть сформирована в результате исследования должна показывать для каждой торговой точки результат исследования. Перед составлением таблицы должны быть очищены данные, то есть удалены торговые точки, имеющие нулевые платежи или нулевое количество клиентов в контрольной или тестовой группе.

Поля таблицы должны быть следующими:

* city - город
* id\_trading\_point - уникальный идентификатор торговой точки, к которой прикреплен данный пользователь
* count\_test - кол-во наблюдений в тестовой группе
* count\_control - кол-во наблюдений в контрольной группе
* count\_all - суммарное кол-во наблюдений
* percent\_count - процент кол-ва наблюдений данной ТТ от всех наблюдений
* avg\_payment\_test - средний платеж в тестовой группе
* avg\_payment\_control - средний платеж в контрольной группе
* diff - разница между средними платежами
* sigma\_test - стандартное отклонение платежей в тестовой группе
* sigma\_control - стандартное отклонение платежей в контрольной группе
* ttest - значение статистического критерия (критерий Стьюдента для сравнения средних платежей)
* pvalue\_ttest - pvalue статистического критерия (критерий Стьюдента для сравнения средних платежей)

В таблице также должен присутствовать флаг (лейбл), по которому можно будет разделить торговые точки по результатам эксперимента

* если результат (поле diff) положительный, и разница на самом деле есть (на основании p\_value), то это положительный исход
* если результат (поле diff) отрицательный, и разница на самом деле есть (на основании p\_value), то это отрицательный исход
* если разницы на самом деле нет (на основании p\_value), то это нейтральный исход

Полученную таблицу необходимо выгрузить в Excel.

Три различных исхода должны быть выгружены на отдельные листы.

На основании этого Excel-файла будет построен калькулятор (на каждом листе свой калькулятор).

**ТЗ для калькулятора SkyLenta**

1. Положительные исходы

Внешним параметром должно быть количество потенциальных клиентов-покупателей за период времени *N*. Опираясь на поля *diff*, а также на долю покупателей в этой торговой точке среди всех, определите, какая выгода может быть получена от замены механики *A* на *B* при условии, что ей воспользуются *N* клиентов.

1. Отрицательные исходы

Конфигурация калькулятора такая же, как на положительных исходах. Необходимо рассчитать, какая сумма может быть потеряна, если мы заменим механику *А* на механику *В*.

c) Нейтральные исходы

Мы предполагаем, что результаты на эти торговых точках не видны, так как мы собрали слишком мало наблюдений. Мы хотим понять, какое количество наблюдений нам потребуется в каждой из этих торговых точек, чтобы разглядеть определенную разницу в средних платежах. Внешним параметром должно быть MDE - сумма в рублях (разница между средними платежами в группах), которая считается минимально значимой с точки зрения бизнеса. Для каждой торговой точки должно быть рассчитано кол-во наблюдений, необходимое для обнаружения разницы масштаба MDE (рассчитывать на основании стандартного отклонения платежей).

# Описание АБ-Теста

**3.1 Цель эксперимента**

Исследование альтернативного метода воздействия на клиентские покупки с помощью пуш-уведомлений.

Воздействие “контроль” - уведомление о новых товарах и скидках с помощью баннера в приложении

Воздействие “тест” - уведомление с помощью пуша (сообщение о товарах и скидках появится в уведомлениях приложения).

**3.2 Дизайн эксперимента**

Длительность эксперимента - 3 месяца.

География: в эксперименте задействованы все города присутствия в России.

**3.3 Сплит-система**

Клиенты разбиты на две группы одинакового размера случайным образом.

**3.4 Таргет-метрики**

* Конверсия из рекламы в покупку
* Средний чек

**3.5 Описание данных**

Исходный датасет представлен в приложении А.

Вкладка “Данные”:

* id\_order - уникальный идентификатор покупки.
* id\_client - уникальный идентификатор клиента.
* amt\_payment - размер платежа.
* dtime\_pay - дата и время оплаты.

Вкладка “Clients”:

* id\_client - уникальный идентификатор клиента.
* dtime\_ad - дата и время показа рекламного объявления (или баннером в приложении, или пуш-уведомлением, в зависимости от группы).
* nflag\_test - группа эксперимента (0 - контроль, 1 -тест).
* id\_trading\_point - идентификатор торговой точки, к которой прикреплен данный пользователь.

Вкладка “Region\_dict”:

* id\_trading\_point - уникальный идентификатор торговой точки, к которой прикреплен данный пользователь.
* city - название города.

# Инструменты

В данном разделе приведен перечень пакетов и программных компонентов, используемых для выполнения работы.

**4.1 Excel**

[Программа](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0) для работы с [электронными таблицами](https://ru.wikipedia.org/wiki/%D0%AD%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D1%82%D0%B0%D0%B1%D0%BB%D0%B8%D1%86%D0%B0) (входит в пакет [Microsoft Office](https://ru.wikipedia.org/wiki/Microsoft_Office))

**4.2 Jupyter**

[Интерактивный блокнот](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%82%D0%B5%D1%80%D0%B0%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B9_%D0%B1%D0%BB%D0%BE%D0%BA%D0%BD%D0%BE%D1%82), веб-реализация [Python](https://ru.wikipedia.org/wiki/IPython)

Для реализации работы использованы следующие библиотеки и модули [Python](https://ru.wikipedia.org/wiki/IPython):

Pandas - [программная библиотека](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%B0%D1%8F_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B0) на языке [Python](https://ru.wikipedia.org/wiki/Python) для обработки и [анализа данных](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85). Работа pandas с данными строится поверх библиотеки [NumPy](https://ru.wikipedia.org/wiki/NumPy), являющейся инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для манипулирования числовыми [таблицами](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)) и [временными рядами](https://ru.wikipedia.org/wiki/%D0%92%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D0%BE%D0%B9_%D1%80%D1%8F%D0%B4). Название библиотеки происходит от [эконометрического](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0) термина [«панельные данные»](https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BD%D0%B5%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5), используемого для описания многомерных структурированных наборов информации.

NumPy - библиотека с [открытым](https://ru.wikipedia.org/wiki/%D0%9E%D1%82%D0%BA%D1%80%D1%8B%D1%82%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D0%B5%D1%81%D0%BF%D0%B5%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) [исходным кодом](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D1%85%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BE%D0%B4) для [языка программирования](https://ru.wikipedia.org/wiki/%D0%AF%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) [Python](https://ru.wikipedia.org/wiki/Python). Возможности: поддержка многомерных [массивов](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%B4%D0%B5%D0%BA%D1%81%D0%BD%D1%8B%D0%B9_%D0%BC%D0%B0%D1%81%D1%81%D0%B8%D0%B2) (включая [матрицы](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0))); поддержка [высокоуровневых](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D1%81%D0%BE%D0%BA%D0%BE%D1%83%D1%80%D0%BE%D0%B2%D0%BD%D0%B5%D0%B2%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) математических функций, предназначенных для работы с многомерными массивами.

Matplotlib - [библиотека](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B0_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)) на языке программирования [Python](https://ru.wikipedia.org/wiki/Python) для визуализации данных [двумерной](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0#%D0%94%D0%B2%D1%83%D0%BC%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0_(2D)) и [трёхмерной](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0#%D0%A2%D1%80%D1%91%D1%85%D0%BC%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0) графикой.

Seaborn - библиотека для создания статистических графиков на Python. Она построена на основе matplotlib и тесно интегрируется со структурами данных pandas. Seaborn помогает изучить и понять данные. Его функции построения графиков работают с датасетами и выполняют все необходимы преобразования для создания информативных графиков.

Scipy - библиотека для [языка программирования](https://ru.wikipedia.org/wiki/%D0%AF%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) [Python](https://ru.wikipedia.org/wiki/Python) с [открытым](https://ru.wikipedia.org/wiki/%D0%9E%D1%82%D0%BA%D1%80%D1%8B%D1%82%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D0%B5%D1%81%D0%BF%D0%B5%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) [исходным кодом](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D1%85%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BE%D0%B4), предназначенная для выполнения научных и инженерных расчётов

Xlsxwriter — это API Python с открытым исходным кодом для записи файлов в формате Excel 2007+ XLSX. Используя API, можно записывать текст, формулы, числа и гиперссылки на несколько рабочих листов. Кроме того, API позволяет вставлять диаграммы, объединять ячейки, форматировать ячейки, применять фильтры, данные проверки, вставлять изображения PNG/JPEG/BMP/WMF/EMF, использовать многоформатные строки и многое другое.

**4.3 Microsoft Word (MS Word)**

[Текстовый процессор](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2%D1%8B%D0%B9_%D0%BF%D1%80%D0%BE%D1%86%D0%B5%D1%81%D1%81%D0%BE%D1%80), предназначенный для создания, просмотра, редактирования и форматирования текстов статей, деловых бумаг, а также иных [документов](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2%D1%8B%D0%B9_%D1%84%D0%B0%D0%B9%D0%BB), с локальным применением простейших форм [таблично](https://ru.wikipedia.org/wiki/%D0%A2%D0%B0%D0%B1%D0%BB%D0%B8%D1%86%D0%B0)-[матричных](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)) алгоритмов (входит в пакет [Microsoft Office](https://ru.wikipedia.org/wiki/Microsoft_Office)).

# Теоретическая часть

В основе выполнения настоящей работы лежит анализ АБ-теста через анализ статистических критериев.

А/B-тест (эксперимент) — это аналитическая техника, цель которой — сравнение нескольких конфигураций/механик в процессах с помощью контрольной и тестовой группы.

Нормальное распределение — «идеальное распределение», свойства которого имеют корни в физике и теории вероятности. Также его называют “Распределение Гаусса” или “Распределение Муавра”

Кривая нормального распределения по форме похожа на купол, симметричный относительно своей вершины.

𝑋 ~ 𝑁 (𝜇, 𝜎)

У нормального распределения два параметра:

𝜇 — точка центра распределения,

𝜎 — разброс распределения, который определяет, насколько приплюснутым или, наоборот, высоким будет форма кривой распределения.

Нормальное стандартное распределение — это нормальное распределение с фиксированными параметрами Nu (𝜇 = 0, 𝜎 = 1).

Для анализа АБ-теста в работе использовано Распределение Стьюдента и t-критерий. t-распределение Стьюдента — это нормальное стандартное распределение, деленное на нормированный на свои степени свободы хи-квадрат.

Сравним нормальное стандартное распределение и распределение Стьюдента:

- они оба симметричны

- они оба центрированы в нуле

- разница заключается в “толщине хвостов”

Для принятия или отклонения теста в пользу контроля нами сравниваются статистические критерии.

Статистический критерий — это некое значение, с помощью которого мы можем принять или отклонить нулевую гипотезу H₀ (по сравнению с H₁).

Если значение статистического критерия попадает в область принятия, то наша гипотеза H₀ подтверждается по сравнению с H₁.

Если значение статистического критерия попадает в критическую область, то наша гипотеза H₁ подтверждается по сравнению с H₀ (или можно сказать, что H₀ отвергается по сравнению с H₁).

Критическая зона, которая в нашем случае составляет 5% (α) от всей площади графика под кривой плотности стандартного нормального распределения, разбита на две симметричные части — по 2.5% с каждой стороны (рисунок 5.1).

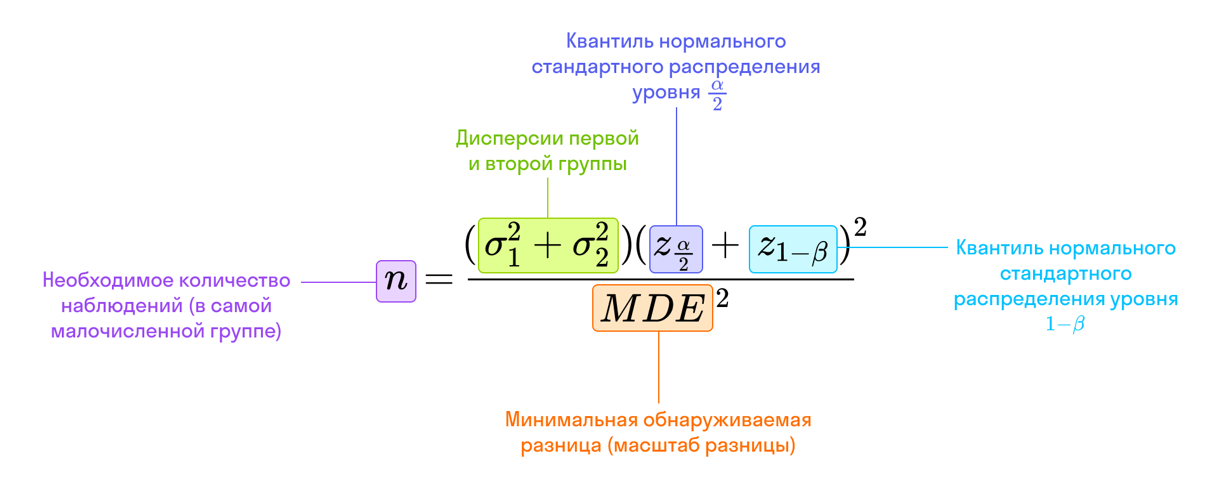


Рисунок 5.1

P\_value — это накопленная доля площади под кривой плотности, которая соответствует нашему рассчитанному статистическому критерию. Его можно воспринимать как «относительную силу нулевой гипотезы». Если p\_value, соответствующее нашему статистическому критерию, больше, чем уровень значимости α, то «нулевая гипотеза сильная» и ее нужно принять по сравнению с альтернативной гипотезой.

Для построения нормального распределения требуется статистически значимое количество данных (измерений). В некоторых случаях для оценки результатов АБ-теста не хватает данных, в этом случае устанавливается MDE для корректности эксперимента.

MDE (Minimum Detectable Effect) — это минимальная разница между двумя группами, которая считается значимой (обычно с точки зрения бизнеса).



# Решение

Решения по обработке и работе с данными представлены в приложении Б и В.

**6.1 Импорт и анализ таблиц**

6.1.1 Импортируем данные в окружение Jupyter Notebook.

Импорт данных осуществляется с помощью функции “.read\_exel”

Выгрузим информацию с листа “Данные” в датафрейм “df\_date”, “Clients” в “df\_clients”, “Region\_dict” в датафрейм “df\_reg\_dict”

6.1.2 Изучаем все три таблицы и проверяем их на наличие пустых значений.

Для изучения таблиц используем функции “.describe()” и “.info()”

Результаты исследования приведены на рисунках 6.1.2.1…6.1.2.3

|  |  |
| --- | --- |
|  |  |

Рисунок 6.1.2.1

|  |  |
| --- | --- |
|  |  |

Рисунок 6.1.2.2

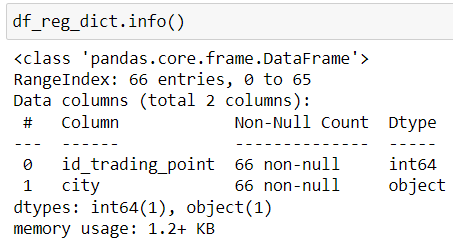


Рисунок 6.1.2.3

6.1.3 Исключаем из таблиц все строки, в которых есть “нулловые” значения.

По итогам очистки количество строк в таблице “df\_date” сократилось с 38 691 до 36 519.

Для таблицы “df\_clients” аналогичные значения составили 55 605 и 55 356 соответственно.

Таблица “df\_reg\_dict” без изменений

6.1.4 Исследуем количество торговых точек в каждом городе - строим группировку по количеству в каждом городе и визуализируем с помощью гистограммы.

Проводим группировку с помощью функции “.groupby”.

Гистограмма с группировкой приведена на рисунке 6.1.4.

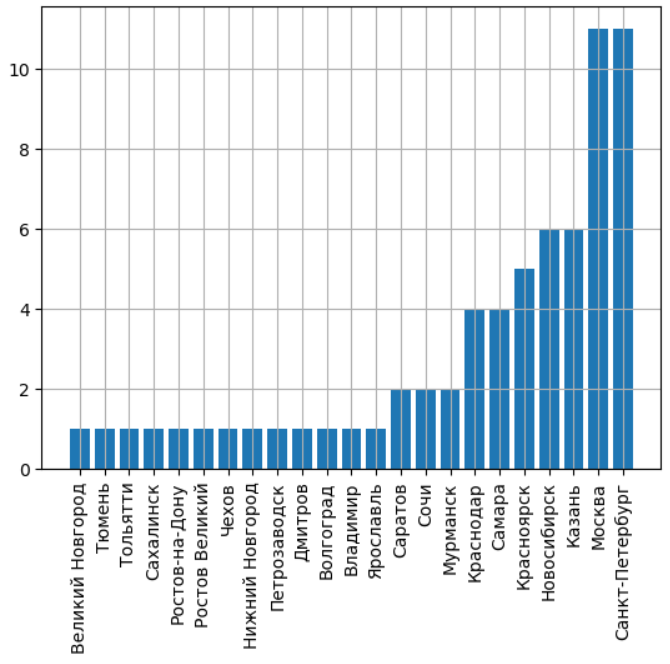


Рисунок 6.1.4

**6.2 Объединение таблиц**

6.2.1 У одного клиента может быть несколько платежей. Строим агрегацию таблицы с платежами, где вычисяем сумму платежей на каждого клиента.

Проводим группировку с помощью функции “.groupby”.

6.2.2 Соединяем (по клиенту) сгруппированную таблицу с платежами с клиентской таблицей. Убеждаемся, что в таблице остались все клиенты из клиентской таблицы.

Соединяем таблицы с помощью функции “ .merge”

6.2.3 Заполняем нулями суммы покупок тех клиентов, которые их не совершали

6.2.4 С помощью словаря регионов подтягиваем к каждой торговой точке город, в котором она находится.

Соединяем таблицы с помощью функции “ .merge”

6.2.5 Создаем поле флаг платежа, который принимает значения 0 или 1 в зависимости от того, заплатил клиент или нет.

Для создания условия используем пакет “numpy”

**6.3 Автоматизация статистических вычислений**

6.3.1 Создаем функцию test\_calc, которая будет вычислять значение t-критерия (критерия Стьюдента) и p\_value для сравнения средних и с помощью функции print выводить сообщение о том, существует ли разница между средними (на основании p\_value).

Аргументы функции: r1 (первая выборка), r2 (вторая выборка), alpha (уровень значимости - имеет значение по умолчанию 5%).

Теоретическая часть по распределениям и статистическим гипотезам приведена в разделе 5.

6.3.2 Создаем функцию mann\_whitney\_func, которая будет рассчитывать значение критерия Манна Уитни и p\_value для сравнения распределений и с помощью функции print выводить сообщение о том, существует ли разница между средними (на основании p\_value).

Аргументы функции: r1 (первая выборка), r2 (вторая выборка), alpha (уровень значимости - имеет значение по умолчанию 5%).

Теоретическая часть по распределениям и статистическим гипотезам приведена в разделе 5.

**6.4 Чистка неверно заполненных точек**

6.4.1 Создаем пустой список. Запускаем цикл по всем торговым точкам и добавляем к этому списку все торговые точки, в которых не было заплачено ни одного рубля ни одним клиентом.

Перечень торговых точек: [739.0, 866.0, 46.0, 1099.0, 1015.0, 603.0, 810.0, 800.0, 228.0, 26.0, 7.0, 23.0, 4.0, 1.0, 13.0]

6.4.2 Создаем еще один пустой список. Запускаем цикл по всем торговым точкам и добавляем к этому списку все торговые точки, в которых пустует или контрольная, или тестовая группа.

Перечень торговых точек: [739.0, 866.0, 1099.0, 1015.0, 603.0, 810.0, 800.0, 228.0]

**6.5 Расчет общих результатов АБ Теста**

6.5.1 Отбрасываем все торговые точки, которые были обнаружены в пункте 6.4.

По итогам операции количество записей в таблице сократилось до 45 318.

6.5.2 Изображаем гистограмму платежей, на которой различными цветами изображены группы “тест” и “контроль”.

Гистограмма распределения платежей по группам приведена на рисунке 6.5.2. Данные очищены от нулевых значений, являющимися статистическими выбросами.

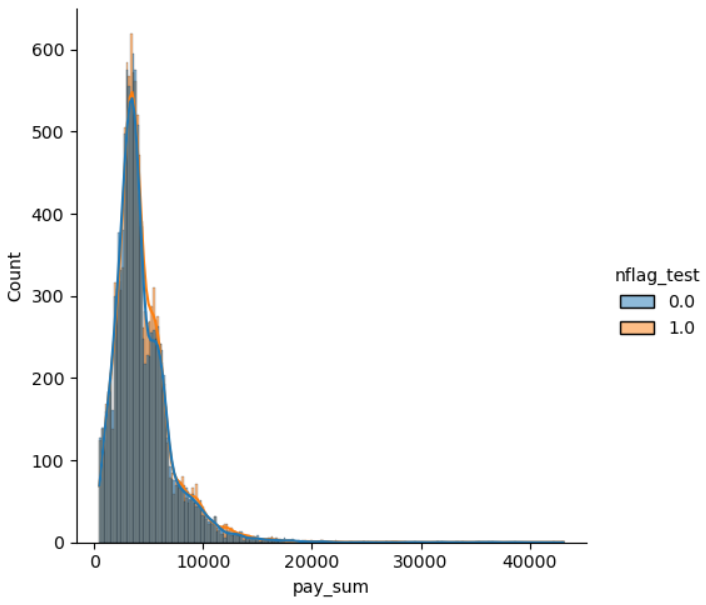


Рисунок 6.5.2

6.5.3 Применяем функцию test\_calc (сравниваем средние платежи)

Результат применения функции test\_calc приведен на рисунке 6.5.3

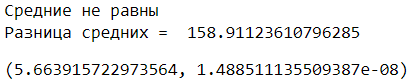


Рисунок 6.5.3

6.5.4 Применяем функцию test\_calc (сравниваем конверсию в платеж, то есть r1 и r2 — это ряды нулей и единиц, в зависимости от того, оплатил ли что-то клиент или нет).

Результат применения функции test\_calc приведен на рисунке 6.5.4

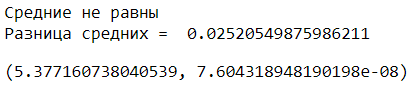


Рисунок 6.5.4

6.5.5 Применяем функцию mann\_whitney\_func.

Результат применения функции mann\_whitney\_func приведен на рисунке 6.5.5



Рисунок 6.5.5

**6.6 Сегментация результатов АБ Теста**

Повторяем аналитику из пункта 6.5 в отношении каждого города в отдельности, запуская цикл.

Результаты анализа приведены в приложении Б – раздел 6. Данные очищены от нулевых значений, являющимися статистическими выбросами.

**6.7 Отчет по АБ Тесту**

Создаем пустой датафрейм. Запускаем цикл по всем торговым точкам.

Результаты анализа приведены в приложении Б – раздел 7.

Заполняем для каждой торговой точки все поля, указанные в разделе 2.

Задаем один из трех лейблов, в зависимости от результата теста в данной торговой точке.

Для создания условия используем пакет “numpy”

Выгружаем полученные результаты в Excel. Разносим результаты по трем разным листам в зависимости от лейбла.

Полученные результаты представлены в приложении Б – раздел 7 и приложении В.

**6.8 Калькулятор**

ТЗ на построение калькулятора описано в разделе 2.

Калькулятор представлен в приложении В.

В качестве параметров для расчета прибыли и потерь используем количество клиентов.

Положительные исходы: умножение diff на кол-во клиентов (параметр) и умножение на долю данной торговой точки среди всех торговых точек.

Потенциальная прибыль, зависящая от количества клиентов, указана на листе «Положительные исходы» - приложение В.

Отрицательные исходы: такая же механика, только с отрицательными и статистически подтвержденными diff.

Потенциальные потери, зависящие от количества клиентов, указаны на листе «Отрицательные исходы» - приложение В.

Нейтральные исходы: воспользуемся формулой минимального кол-ва наблюдений для АБ Теста. В качестве сигмы берем сигму контроля или минимум из сигм контроля и теста.

В связи с тем, что для некоторых торговых точек затруднительно посчитать прибыль или потери ввиду недостаточности количества измерений, то на листе «Нейтральные исходы» - приложение В, приведено рекомендуемое количество наблюдений в зависимости от статистически значимой средней разности чеков (параметр – MDE).

# Выводы

По результатам первичного анализа обнаружены пустые значения на листах “Данные” и “Clients” в исходном датасете, которые исключены для снижения вероятности получения выбросов и построения распределения, по форме близкого к нормальному.

На основании сравнения средних платежей в контрольной и тестовой группе, а также конверсии в платежи внутри групп, обнаружена статистически значимая разница.

Визуализация результатов по каждой торговой точке приведена в приложении Б.

На основании данных теста, построен калькулятор – приложение В, в котором указаны потенциальная прибыль и убытки по каждой торговой точке.

Потенциальная прибыль превышает потенциальные убытки, но в связи с тем, что для некоторых торговых точек затруднительно посчитать прибыль или потери ввиду недостаточности количества измерений, то на листе «Нейтральные исходы» - приложение В, приведено рекомендуемое количество наблюдений в зависимости от статистически значимой средней разности чеков (параметр – MDE).

Тест имеет скорее позитивный, чем негативный эффект для компании, но для комплексной оценки рекомендуется его продление.

# Приложение А. Датасет

# Приложение Б. Python

# Приложение В. Калькулятор SkyLenta