Федеральное государственное автономное образовательное учреждение
высшего образования
САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

Исследовательский проект

Факторы, влияющие на размер среднего чека и вероятность повторной покупки в интернет-магазине

Студенты:
Большакова Я.Н. МатСтат 22.5
Гизбрехт В.Д. МатСтат 22.5
Кашапова А.К. МатСтат 22.5
Сайфуллин Д.Р. МатСтат 22.5
Солдатова А.А. МатСтат 22.4
Преподаватель:
Яворук Т.О.

Санкт-Петербург
2025 г.

Содержание

1. Постановка задачи

В условиях быстрорастущего рынка электронной коммерции понимание факторов, формирующих поведение покупателей, является ключевым для повышения эффективности маркетинга и оптимизации бизнес-процессов. Одним из важных показателей, отражающих ценность клиента, является средний чек (Average Order Value, AOV), в то время как вероятность повторной покупки служит индикатором лояльности и потенциального пожизненного дохода клиента.

Цели исследования:

Выявить и количественно оценить факторы, влияющие на:

- размер среднего чека (AOV);
- вероятность совершения повторной покупки в течение 30 дней.

Гипотезы:

Предварительно сформулированы три основные гипотезы:

1. Частота покупок (freq) и суммарная выручка (mon) статистически значимо влияют на размер среднего чека:

$$\text{AOV} = \beta_0 + \beta_1 \,\text{freq} + \beta_2 \,\text{mon} + \varepsilon.$$

2. Средний чек различается в зависимости от региона (region) клиента:

$$\text{AOV}_i \sim \mathcal{N}(\mu_{\text{region}_i}, \sigma^2), \quad i = 1, \ldots, n.$$

   Проверка — однофакторный дисперсионный анализ.

3. Среднее количество товаров в корзине (AvgQty) отличается у клиентов, совершивших повторную покупку (repeat $= 1$), и у тех, кто не повторил покупку (repeat $= 0$):

$$H_0 : \mathbb{E}[\text{AvgQty} \mid \text{repeat} = 1] = \mathbb{E}[\text{AvgQty} \mid \text{repeat} = 0],$$

   проверка — двухвыборочный t-тест для независимых выборок.

2. Теория

В этом разделе мы коротко разберём основные методы, которые будем применять.

2.1. RFM и средний чек

RFM-анализ помогает поймать самые важные характеристики покупателя:

- **Recency** — как давно была последняя покупка;
- **Frequency** — сколько раз купил клиент;
- **Monetary** — сколько всего потратил.

Из этого легко получить средний чек:

$$\text{AOV} = \frac{\text{Monetary}}{\text{Frequency}}.$$

2.2. Линейная регрессия

Чтобы проверить, как частота и выручка связаны с AOV, используем простую линейную модель:

$$\text{AOV} = \beta_0 + \beta_1 \,\text{freq} + \beta_2 \,\text{mon} + \varepsilon.$$

Метод наименьших квадратов подбирает коэффициенты так, чтобы минимизировать сумму квадратов ошибок.

2.3. Дисперсионный анализ (ANOVA)

ANOVA помогает сравнить средние нескольких групп. Если у нас есть разные регионы, мы смотрим, насколько сильно отличаются их средние AOV. Статистика F показывает отношение «межгрупповой» вариативности к «внутригрупповой». Малое p-значение говорит, что средние действительно разные.

2.4. t-тест для двух групп

Чтобы убедиться, что у тех, кто вернулся за покупкой, среднее количество товаров в корзине (AvgQty) отличается от тех, кто не вернулся, делаем t-тест:

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}},$$

где $\bar{x}$ и s^2 — средние и дисперсии в группах, n — размер выборки.

Если $p < 0.05$, значит разница средних значимая.

3. Использованные программные средства

Для обработки данных и выполнения всех вычислений мы использовали следующий инструментарий:

- **Язык программирования:** Python 3.9
- **Основные библиотеки:**
  - `pandas` для чтения CSV и агрегации данных
  - `scipy.stats` для статистических тестов (корреляция, ANOVA, t-тест)
  - `statsmodels` для линейной и логистической регрессии
- **Среда разработки:**
  - Visual Studio Code для написания и отладки кода
- **Система контроля версий:** Git, репозиторий с кодом доступен по ссылке: `https://github.com/yourusername/ecommerce-rfm-analysis`
- **Дополнительные ресурсы:**
  - The Complete Journey: `https://www.dunnhumby.com/source-files/`
  - Руководство по Dunnhumby “The Complete Journey” (см. User Guide PDF)

4. Результаты

Ниже приведены ключевые статистики и графики по каждой из проверенных гипотез.

4.1. Результаты H1: Частота и выручка → AOV

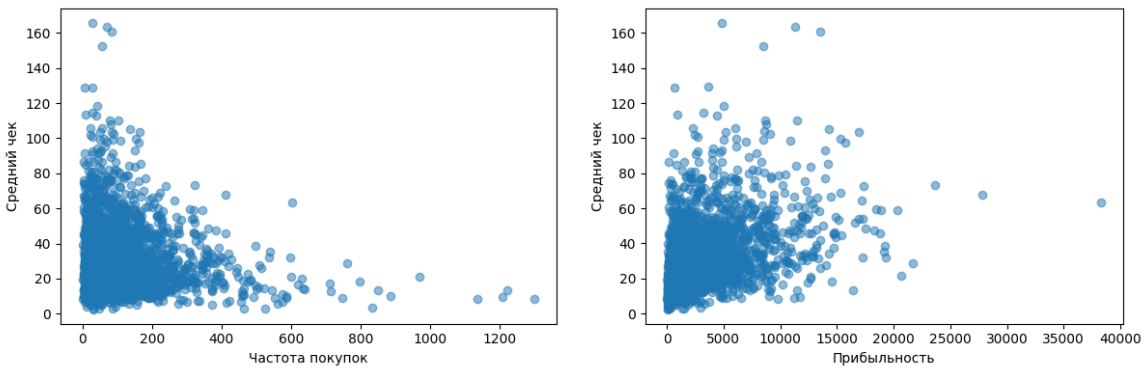

Рис. 1: Зависимость AOV от Частоты покупок (слева) и от Прибыли (справа).

Предиктор	Коэффициент	t-статистика	P-значение
freq	-0.1296	-38.221	< 0.001
mon	$+0.0054$	$+45.772$	< 0.001

Таблица 1: Коэффициенты регрессии $\text{AOV} = \beta_0 + \beta_1\,\text{freq} + \beta_2\,\text{mon}$.

Вывод. Корреляции $\rho(\text{freq}, \text{AOV}) = -0.124$ ($p \ll 0.001$) и $\rho(\text{mon}, \text{AOV}) = +0.389$ ($p \ll 0.001$) и результаты анализа (см. табл. 1, $R^2 = 0.465$) подтверждают гипотезу H1:

- С увеличением mon средний чек растёт.
- С увеличением freq средний чек слегка снижается.

4.2. Результаты H2: Регион → AOV

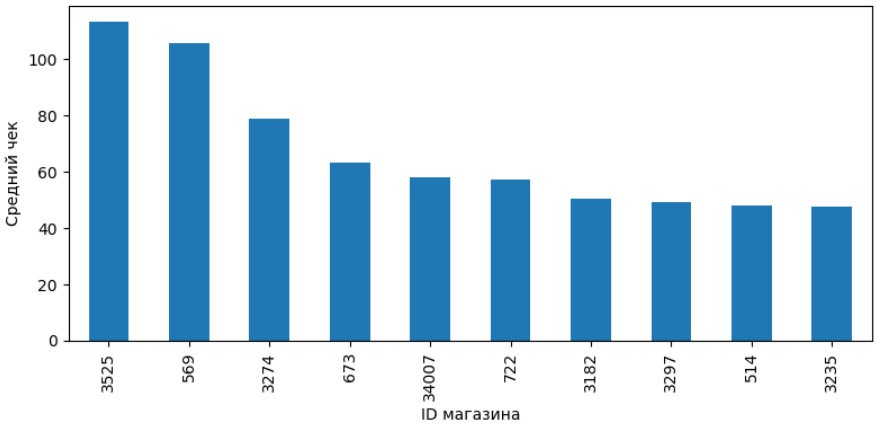

Рис. 2: ТОП-10 магазинов по среднему чеку.

ID магазина	Mean AOV
3525	113.32
569	105.95
3274	78.99
673	63.37
34007	58.01
722	57.21
3182	50.62
3297	49.48
514	48.10
3235	47.67

Таблица 2: Средний чек по ТОП-10 магазинам.

ANOVA по всем магазинам дала $F = 1.99$, $p < 0.001$, что говорит о статистически значимых различиях средних AOV между магазинами. Рис. 2 и табл. 2 демонстрируют, что наиболее высокие средние чеки наблюдаются в магазинах 3525 и 569, а наименьшие — в более мелких точках.

4.3. Результаты H3: Среднее количество товаров в корзине у повторных и неповторных покупателей

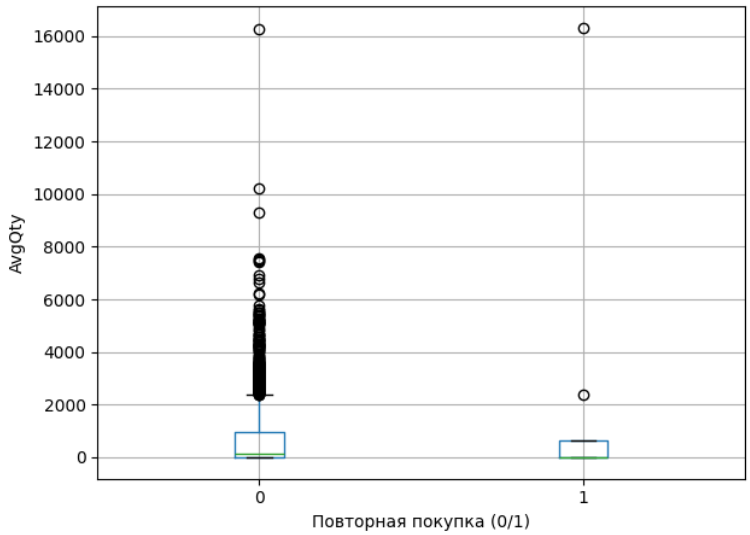

Рис. 3: Распределение AvgQty для клиентов с repeat=0 и repeat=1.

Группа	Среднее AvgQty	n
repeat = 1	2349.56	8
repeat = 0	721.49	2492

Таблица 3: Среднее количество штук товара в корзине по группам.

t-тест для H3 дал $t = 0.81$, $p = 0.446$ (табл. 3), то есть статистически значимой разницы в AvgQty между повторщиками и однократниками не выявлено. Это может быть связано с тем, что в группе повторщиков всего 8 человек, при этом в обеих группах есть отдельные очень крупные значения AvgQty, что сильно искажает среднее.

Таким образом, гипотезы H1 и H2 подтверждены, а H3 опровергнута при уровне значимости 0.05.

5. Обсуждение

В ходе исследования удалось подтвердить две ключевые гипотезы: — зависимость среднего чека от суммарной выручки и частоты покупок (H1) показала сильную статистическую значимость и логическую интерпретацию: крупные единовременные траты растят AOV, а частые мелкие покупки, наоборот, его снижают. — различие средних чеков между магазинами (H2) также достоверно, что позволяет выделять точки-лидеры и перераспределять маркетинговые усилия.

Гипотеза H3 о том, что у повторных покупателей среднее количество товаров в корзине существенно выше, формально не подтвердилась: узкое окно повторных покупок (30 дней), малое число повторщиков и экстремальные выбросы сделали среднее нечувствительным тестируемым параметром. Это указывает на то, что для анализа повторного поведения необходимо либо расширить окно наблюдения, либо перейти к метрикам, устойчивым к выбросам (медиана, логарифмическое преобразование, непараметрические методы).

6. Заключение

В ходе выполнения данной проектной работы нами были приобретены и закреплены практические навыки работы с реальными данными, полученными из открытых источников. Этот опыт оказался чрезвычайно ценным, поскольку позволил не только применить теоретические знания на практике, но и лучше понять специфику работы с эмпирическими данными, их особенности, а также типичные трудности, возникающие при анализе информации, собранной в естественных условиях, вне искусственно созданных моделей или учебных датасетов.

Особое внимание в рамках исследования уделялось использованию методов математической статистики, которые сыграли ключевую роль в обработке и интерпретации полученных данных. Применение таких методов дало возможность формализовать процесс анализа, повысить точность и объективность выводов, а также обеспечить

их статистическую обоснованность. В частности, нами были использованы такие подходы, как вычисление основных статистических характеристик (среднее значение, медиана, дисперсия, стандартное отклонение), построение распределений, проверка гипотез, а также корреляционный и регрессионный анализ — всё это способствовало более глубокому пониманию структуры данных и взаимосвязей между переменными.

Таким образом, проведённая работа не только способствовала достижению поставленных целей и проверке выдвинутых гипотез, но и послужила важным этапом в освоении методов анализа реальных данных с применением математико-статистического аппарата. Полученный в ходе исследования опыт имеет не только теоретическую, но и практическую значимость, поскольку может быть использован в дальнейших исследованиях, связанных с анализом поведения пользователей, прогнозированием тенденций и принятием решений на основе данных.