|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме\_\_\_\_\_\_\_\_\_\_\_ «Adidas US Sales» \_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5-55Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Шакиров Т.М.

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  МаслениковК.Ю.

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Маслеников К.Ю.

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_Анализ данных по базе данных «Adidas US Sales» \_\_

Студент группы \_\_\_\_\_\_\_ИУ5-55Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Шакиров Тимур Маратович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследовательская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики(кафедра,предприятие,НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 3 нед., 50% к 9 нед., 75% к 12 нед., 100% к 15 нед.

***Техническое задание***  Спроектировать систему анализа и визуализации по теме «Adidas US Sales», провести визуализацию полученных данных посредством языка программирования Python

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 23 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «14» сентября 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Маслеников К.Ю.

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Шакиров Т.М.

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Аннотация**

По заданию требуется проанализировать базу данных, очистить её от

ненужных данных, также нужно выдвинуть гипотезы, проанализировать их и составить графики по данным гипотезам. Для данного задания была выбрана база данных «Adidas US Sales». В ходе работы будут использоваться такие библиотеки как: «Pandas» и «Seaborn». Анализ будет проводиться через программу «Jupyter Python».

Также будет производиться корреляционный анализ, агрегирование данных, оптимизация памяти, удаление дубликатов, очистка данных. Будет осуществлён анализ гипотез по известным данным и построены соответствующие графики и схемы.

В работе будут рассмотрены данные о продажах различных товаров бренда Adidas в США. В конце, будет сделан вывод о проделанной работе с данным датасетом.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_Toc184474190)

[Определение данных для анализа 6](#_Toc184474191)

[1. Описание данных 6](#_Toc184474192)

[2. Формулирование гипотез 6](#_Toc184474193)

[3. Подготовка данных для работы 7](#_Toc184474194)

[4. Изучение общей информации 8](#_Toc184474195)

[5. Исследовательский анализ данных 11](#_Toc184474196)

[Гипотеза 1 – Средняя цена за единицу продукции Adidas выше в регионах с высокой покупательской способностью, чем в центральных и южных регионах США 11](#_Toc184474197)

[Гипотеза 2 - Во время праздников количество проданных единиц и общая выручка значительно возрастают, в то время как средняя цена за единицу продукции снижается 12](#_Toc184474198)

[Гипотеза 3 –Уровень операционной прибыли от продаж продукции Adidas зависит от канала продаж – продажи онлайн имеют более высокую операционную прибыль по сравнению с оффлайн продажами 13](#_Toc184474199)

[Гипотеза 4 – Товары в ценовом диапазоне $50-$100 имеют больший средний объем продаж, но меньшую цену за единицу, чем премиальные товары, с вариациями по штатам 14](#_Toc184474200)

[Гипотеза 5 –Рентабельность продаж зависит от региона и сезона, и сезонное влияние на рентабельность варьируется в зависимости от региона 16](#_Toc184474201)

[Анализ корреляции 17](#_Toc184474202)

[ЗАКЛЮЧЕНИЕ 19](#_Toc184474203)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 20](#_Toc184474204)

# **ВВЕДЕНИЕ**

Необходимо исследовать данные и произвести анализ продаж компании Adidas.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

# **Определение данных для анализа**

В качестве данных для анализа был выбран Dataset «Adidas US Sales».

В датасете представлены все продажи компании Adidas, осуществляемые в США с информацией о каждой продаже и её характеристиках, а также о каждом товаре и его описание.

# **Описание данных**

Для анализа были собраны данные о всех продажах, включенных в датасет. В наборе данных содержатся:

* **sales\_id**: уникальный идентификатор продажи
* **retailer**: название розничного продавца
* **retailer\_id**: уникальный идентификатор розничного продавца
* **invoice\_date**: дата счета
* **region**: регион, в котором была совершена продажа
* **state**: штат, в котором была совершена продажа
* **city**: город, в котором была совершена продажа
* **product**: название или описание товара
* **price\_per\_unit**: цена за единицу товара
* **units\_sold**: количество проданных единиц
* **total\_sales**: общая сумма продаж
* **operating\_profit**: операционная прибыль от продажи
* **sales\_method**: метод продажи (например, онлайн, оффлайн)

# **Формулирование гипотез**

Гипотеза 1:Средняя цена за единицу продукции Adidas выше в регионах с высокой покупательской способностью, чем в центральных и южных регионах США.

Гипотеза 2: Во время праздников количество проданных единиц и общая выручка значительно возрастают, в то время как средняя цена за единицу продукции снижается.

Гипотеза 3: Уровень операционной прибыли от продаж продукции Adidas зависит от канала продаж – продажи онлайн имеют более высокую операционную прибыль по сравнению с оффлайн продажами.

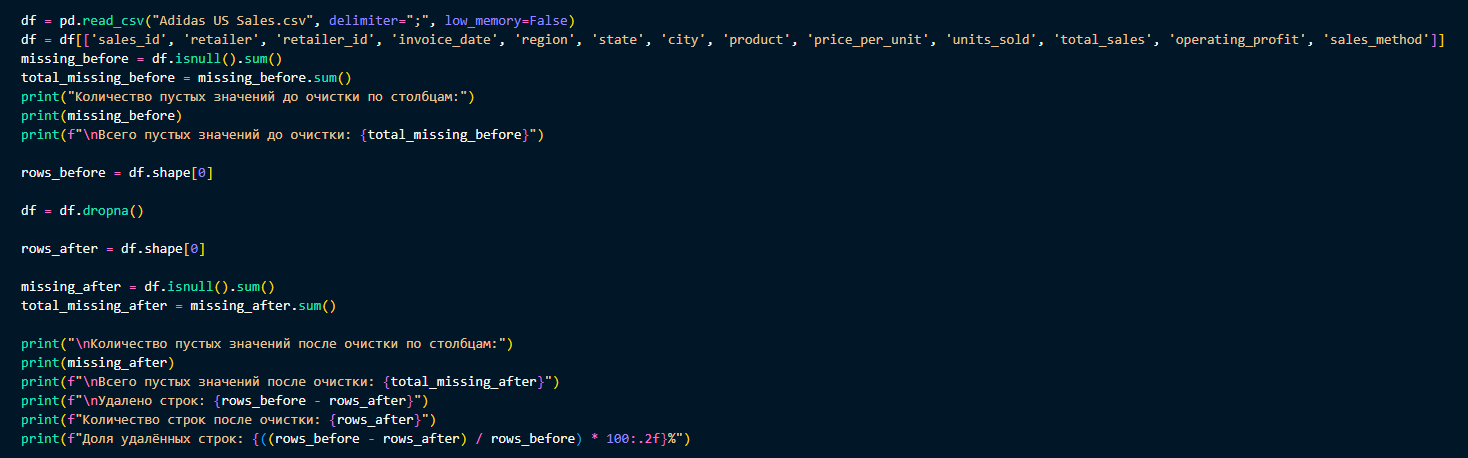
Гипотеза 4: Товары в ценовом диапазоне $50-$100 имеют больший средний объем продаж, но меньшую цену за единицу, чем премиальные товары, с вариациями по штатам.

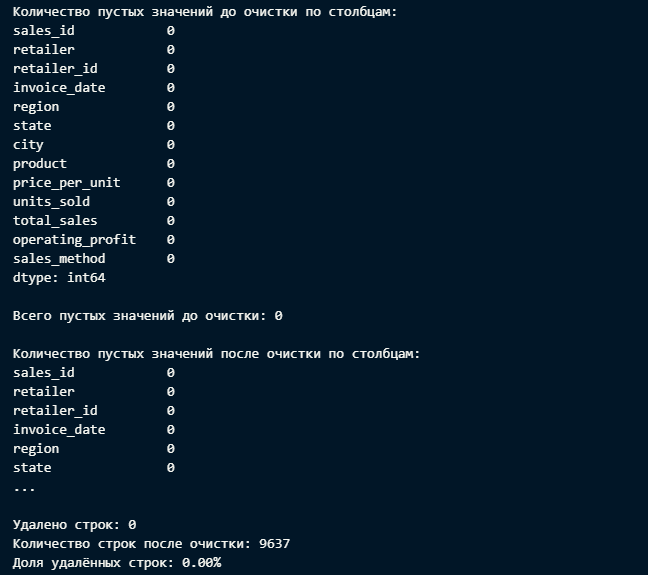
Гипотеза 5: Рентабельность продаж (операционная прибыль на единицу) зависит от региона и сезона, и сезонное влияние на рентабельность варьируется в зависимости от региона.

# **Подготовка данных для работы**

Загружаем датасет и подключаем необходимые библиотеки  
Выбираем из датасета поля, необходимые для анализа

Проводим очистку данных

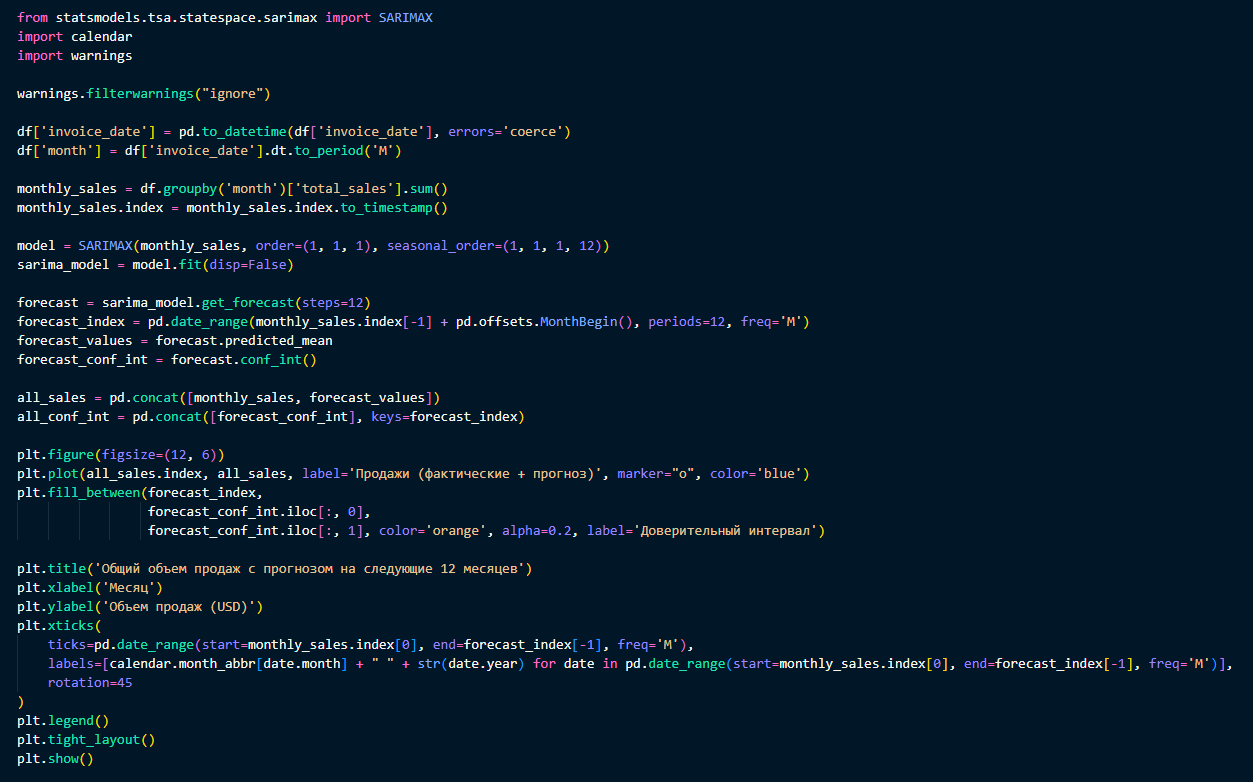


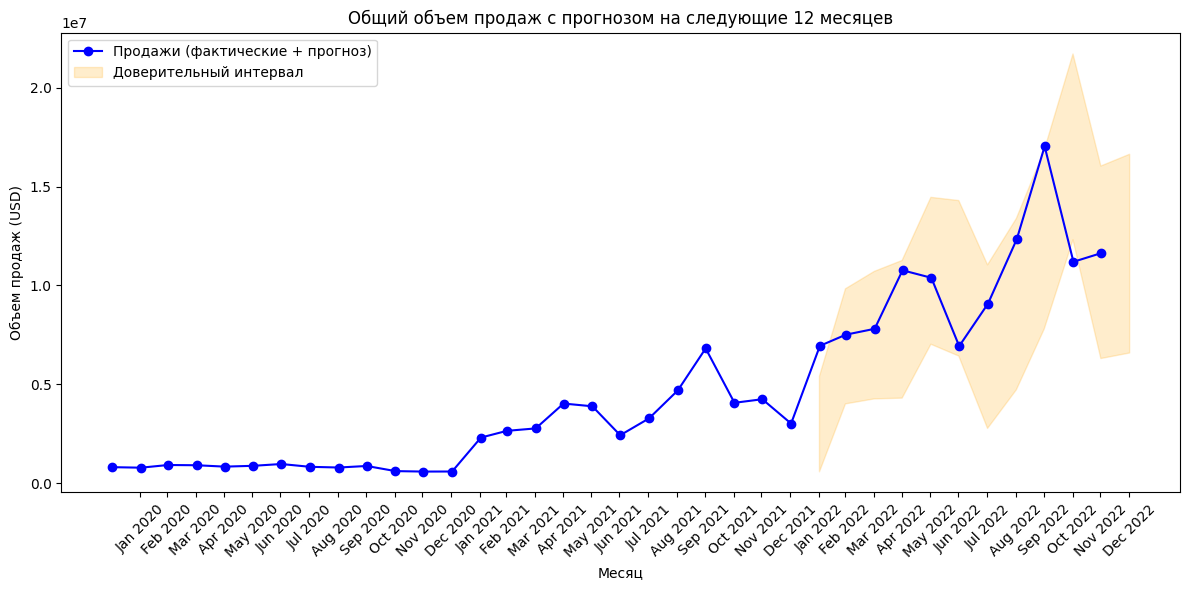


# **Изучение общей информации**

Рассмотрим общий объём продаж и сделаем прогноз на ближайшие 12 месяцев.

Прогноз продаж будем делать с использованием SARIMA. Это расширение модели ARIMA, которое учитывает сезонные колебания временного ряда. Она добавляет сезонные параметры к классической модели ARIMA, позволяя прогнозировать временные ряды, которые имеют регулярные сезонные циклы.

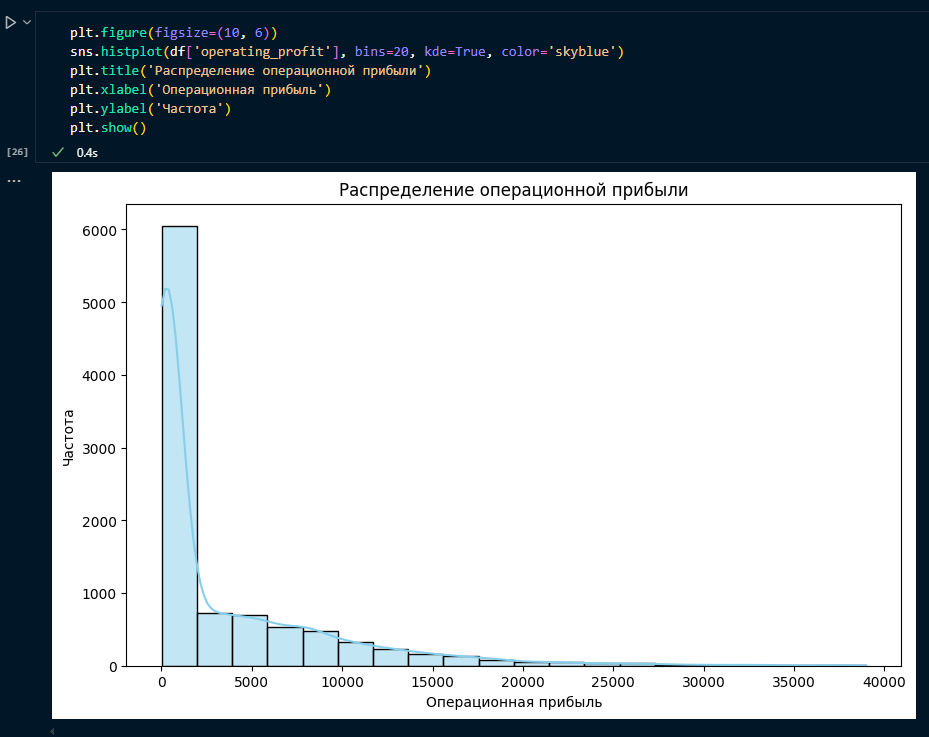




Вывод: Продажи Adidas подвержены сезонным колебаниям, с наибольшей активностью весной и осенью. Для максимизации выручки можно ориентироваться на эти месяцы для проведения маркетинговых акций и повышения запасов.

Продажи демонстрируют выраженную сезонность с пиком осенью, за которым следует снижение к концу года. Эти колебания могут быть связаны с сезонными факторами, такими как повышенный спрос в преддверии осенних и зимних праздников. Общий тренд в продажах остается относительно стабильным, что говорит о постоянном интересе к продукту в течение года, однако с заметными колебаниями.

Рассмотрим распределение операционной прибыли

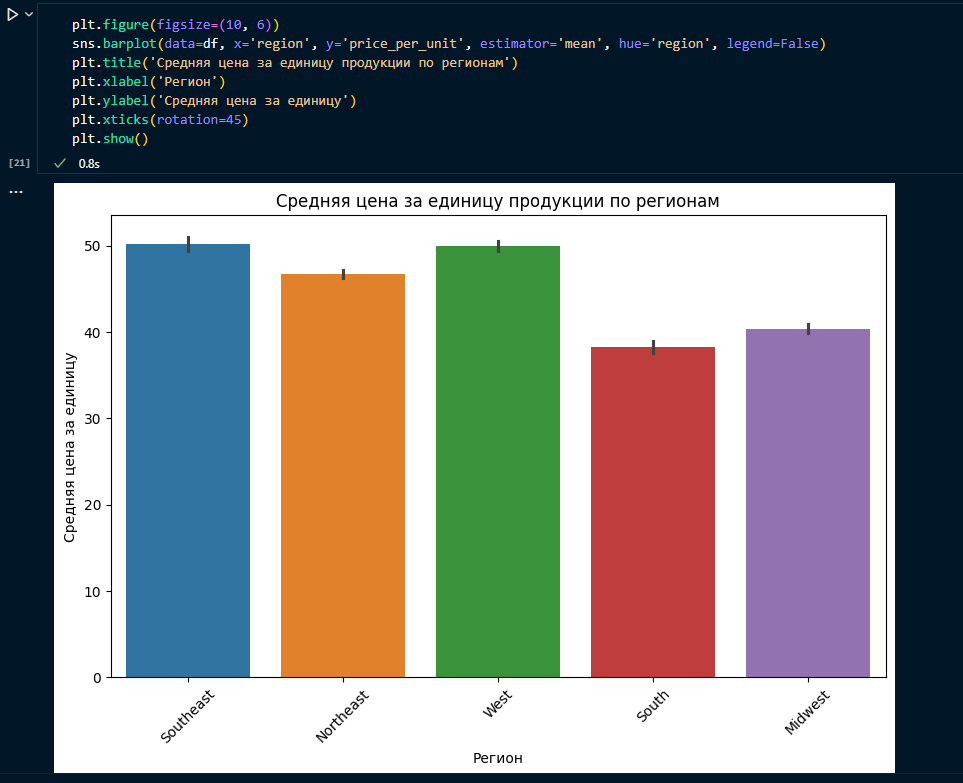


Вывод: График демонстрирует, что большая часть продаж приносит низкую операционную прибыль, в то время как небольшое количество сделок отличается высокой прибыльностью. Такой характер распределения часто встречается в продажах и указывает на необходимость анализа факторов, влияющих на высокоприбыльные сделки, чтобы лучше понять, как увеличить долю таких операций.

# **Исследовательский анализ данных**

## **Гипотеза 1 – Средняя цена за единицу продукции Adidas выше в регионах с высокой покупательской способностью, чем в центральных и южных регионах США**

Для этого составим график, по которому сможем увидеть среднюю цену за единицу продукции по регионам.



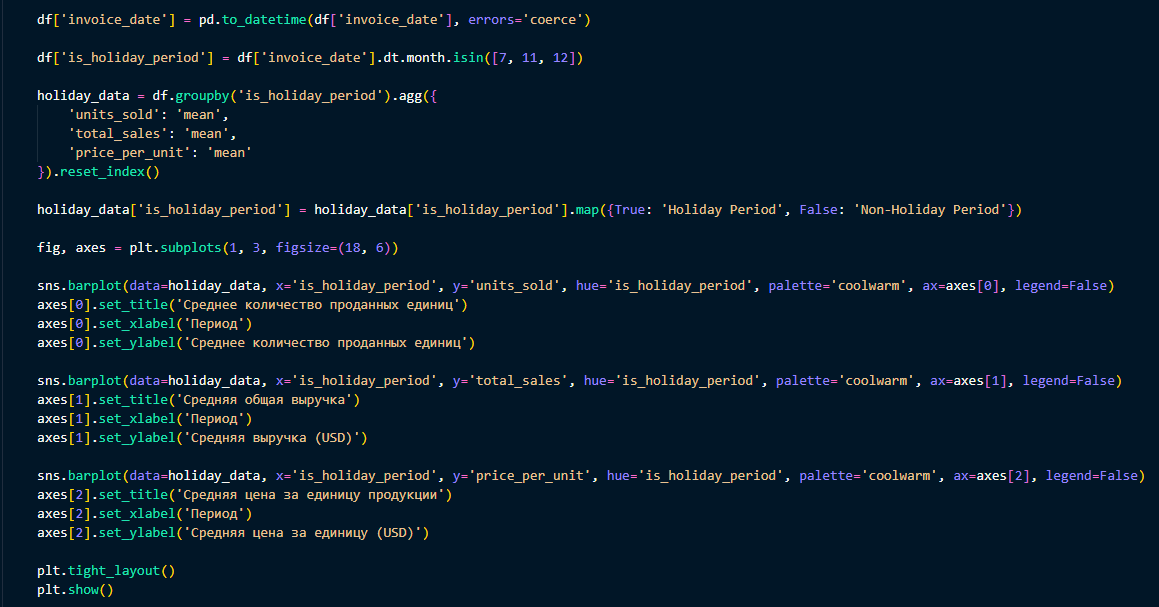
**Вывод:**

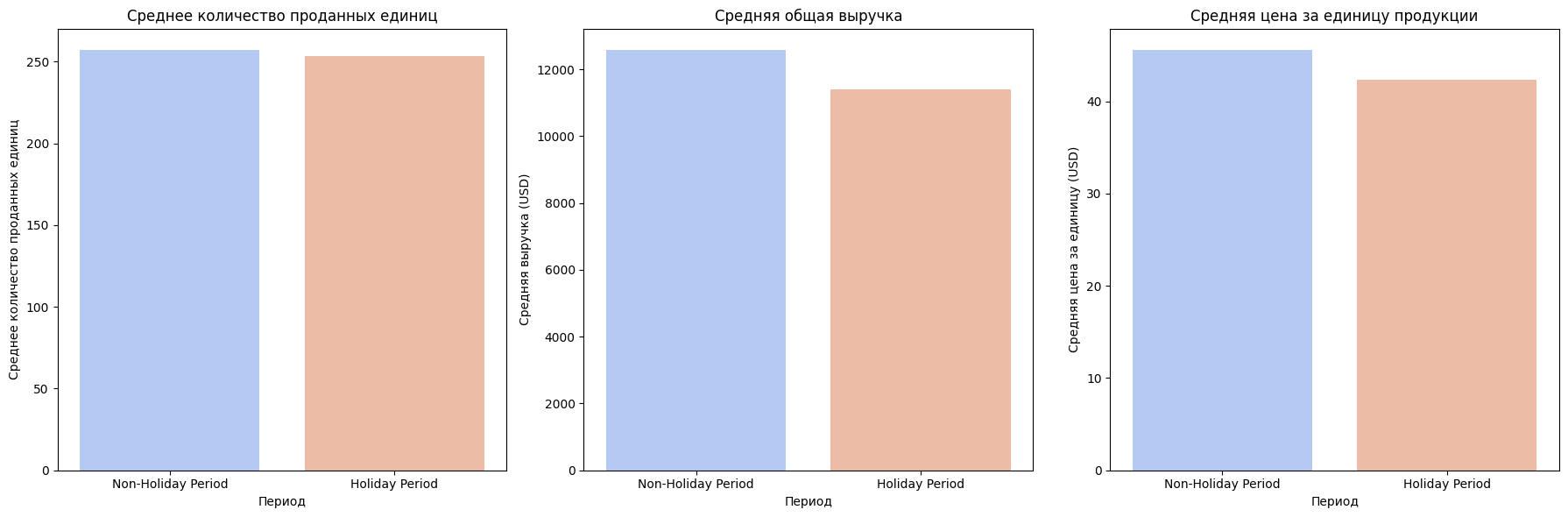
Таким образом, видим, что средняя цена за единицу продукции компании Adidas в регионах с высокой покупательской способностью действительно немного выше, чем в центральных и южных регионах США. Значит гипотеза подтверждается.

## **Гипотеза 2 - Во время праздников количество проданных единиц и общая выручка значительно возрастают, в то время как средняя цена за единицу продукции снижается**

Для проверки этой гипотезы нам необходимо:

1. Определить праздничные периоды: можно выделить такие периоды, как ноябрь-декабрь (черная пятница, рождественские распродажи) и июль (летние распродажи);
2. Сравнить метрики в праздничные и непраздничные периоды:
   1. Вычислить среднее количество проданных единиц и общую выручку;
   2. Вычислить среднюю цену за единицу продукции в каждом периоде.
3. Отобразить все данные на графике, чтобы наглядно сравнить необходимые показатели.

****

Полученные графики выглядят так 

**Вывод:**

На графике видно, что среднее количество проданных единиц в праздничные периоды примерно такое же, как и в непраздничные. Это может указывать на то, что количество проданных товаров в праздничные дни не сильно отличается от остальных периодов.

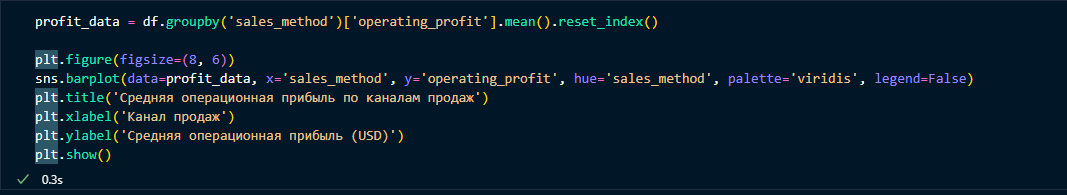
График также показывает, что средняя выручка немного ниже в праздничные периоды по сравнению с непраздничными периодами, хотя разница незначительна. Это может говорить о том, что повышение продаж в праздничные периоды не приводит к значительному увеличению выручки.

Средняя цена за единицу продукции немного ниже в праздничные периоды, что соответствует предположению о снижении цен в праздничные дни. Однако разница опять же незначительна.

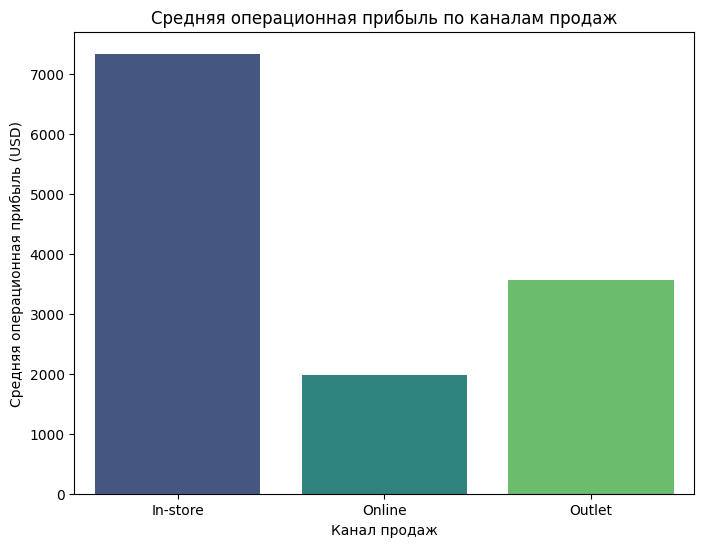
На основе графиков можно сделать вывод, что гипотеза о значительном росте количества проданных единиц и выручки в праздничные периоды не подтверждается. Мы видим лишь небольшое снижение средней цены за единицу продукции, но этот эффект не является существенным.

## **Гипотеза 3 –Уровень операционной прибыли от продаж продукции Adidas зависит от канала продаж – продажи онлайн имеют более высокую операционную прибыль по сравнению с оффлайн продажами**

Для проверки гипотезы необходимо сравнить средние значения операционный прибыли для онлайн и оффлайн заказов.



Полученный график выглядит так



**Вывод:**

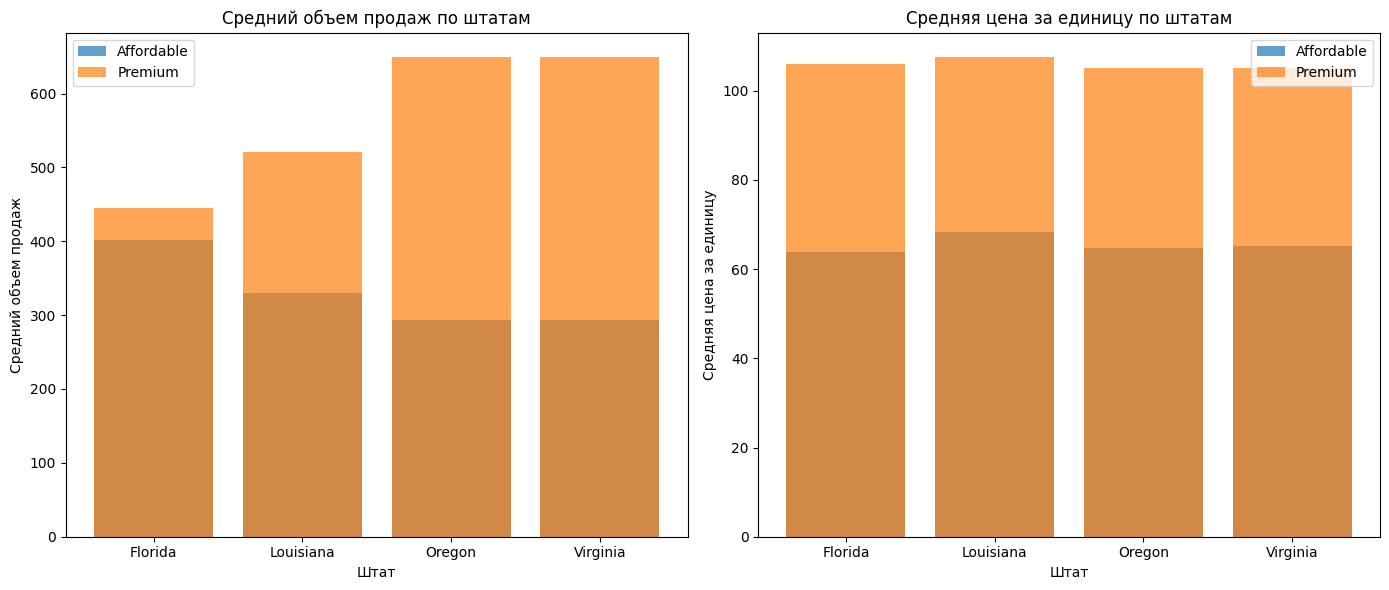
Гипотезане подтверждается. Наоборот, оффлайн-продажи демонстрируют наибольшую прибыль, а онлайн-продажи — наименьшую. Это может быть связано с различиями в ценовой политике, операционных расходах или другими факторами, которые снижают прибыльность онлайн-продаж по сравнению с оффлайн-магазинами.

## **Гипотеза 4 – Товары в ценовом диапазоне $50-$100 имеют больший средний объем продаж, но меньшую цену за единицу, чем премиальные товары, с вариациями по штатам**

Для этого разделим товары на ценовые категории, рассчитаем средние цены за единицу и среднего объёма продаж для каждой категории по штатам, а далее сравним результаты



Полученный график выглядит так



**Вывод:**

Для «доступных» товаров (в диапазоне $50-$100) средний объем продаж выше, чем для «премиальных» товаров (цена выше $100) во всех штатах. Особенно заметен разрыв в штате Орегон, где объем продаж «доступных» товаров значительно превышает объем продаж «премиальных» товаров.

Средняя цена за единицу ожидаемо выше для «премиальных» товаров, чем для «доступных» во всех штатах. Разница в цене между категориями стабильна по всем штатам.

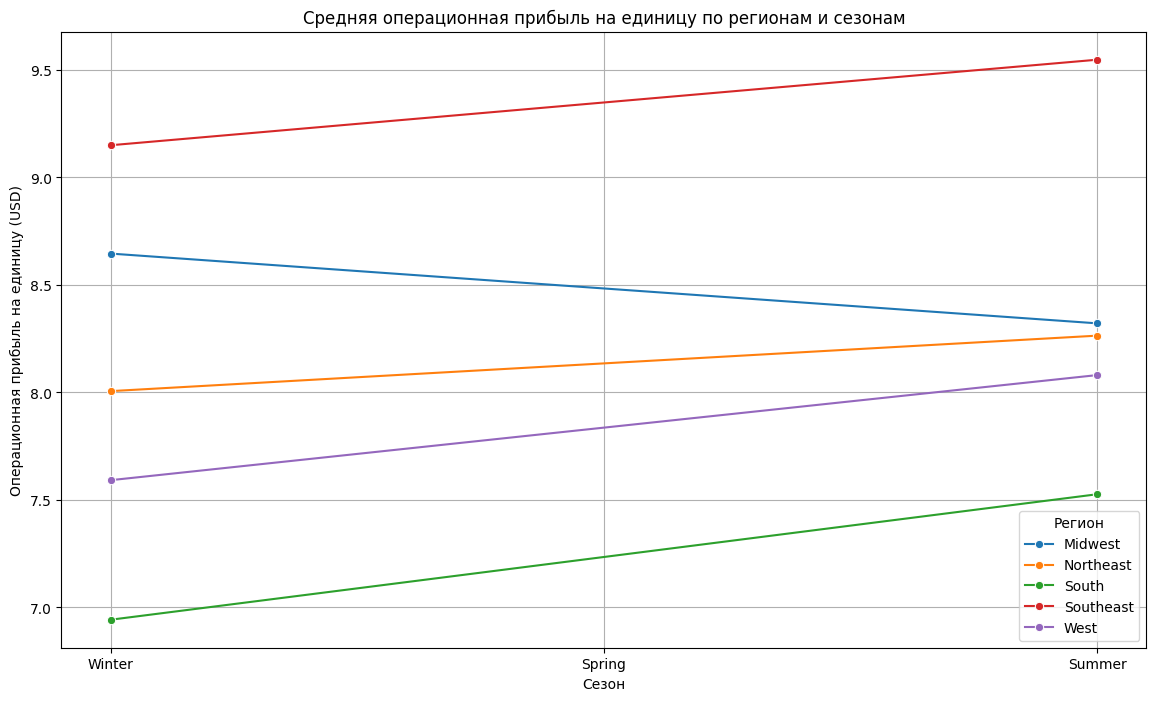
Гипотеза подтверждается: товары в диапазоне $50-$100 действительно имеют больший средний объем продаж, но меньшую цену за единицу, чем "премиальные" товары.

## **Гипотеза 5 –Рентабельность продаж зависит от региона и сезона, и сезонное влияние на рентабельность варьируется в зависимости от региона**

Для этого вычислим среднюю операционную прибыль на единицу продукции для каждого региона и сезона. Затем построим график, показывающий рентабельность по сезонам для каждого региона, чтобы увидеть различия в сезонных изменениях по регионам.



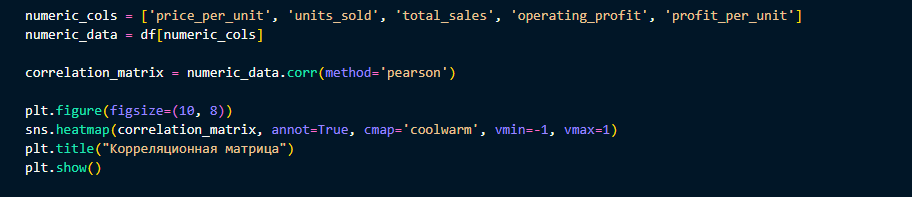
Полученный график выглядит так

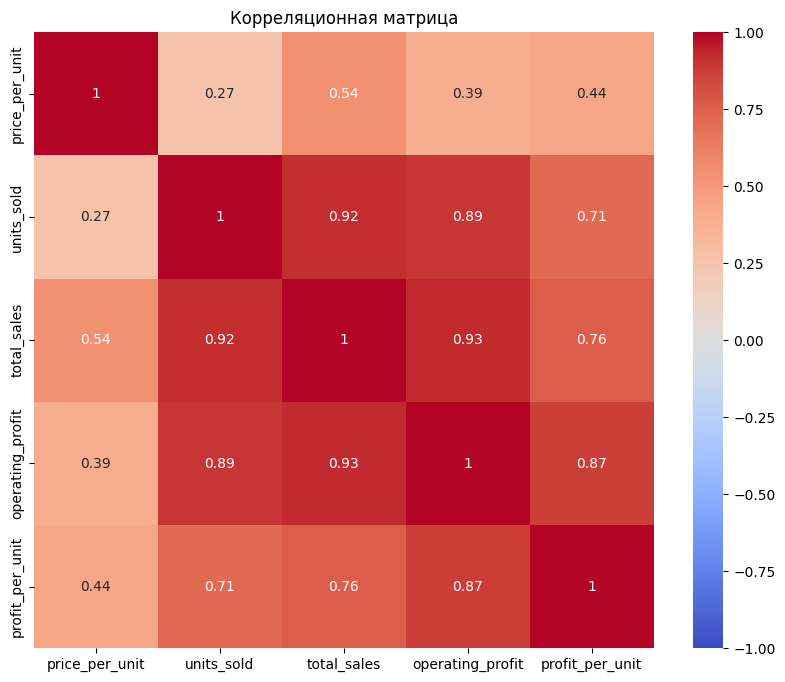


**Вывод:**

Гипотеза о том, что влияние сезона на рентабельность варьируется в зависимости от региона, подтверждается. Региональные отличия в сезонной рентабельности могут быть связаны с климатическими условиями, культурными предпочтениями и спросом на различные категории товаров в разные времена года.

## **Анализ корреляции**

****

****

**Выводы корреляционного анализа:**

* 1. Основным драйвером выручки и прибыли в данном датасете является количество проданных единиц (units\_sold), а не цена за единицу (price\_per\_unit);
  2. Чтобы увеличить общую выручку и операционную прибыль, компании, возможно, стоит фокусироваться на увеличении объемов продаж;
  3. Цена на единицу товара имеет умеренное влияние на прибыль на единицу, но не определяет ключевые метрики, такие как общая выручка и операционная прибыль.

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «Adidas US Sales», выявлены важные статистические данные, которые могут помочь при выборе стратегии будущих продаж.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas», «Seaborn» и «Plotly» на базе языка программирования Python. Также мы пользовались «Jupyter Python»: в нём мы писали код и анализировали диаграммы. В ходе работы я получил новые навыки в области анализа данных и познакомился с вышеперечисленными программами.

Была осуществлена оптимизация использования памяти, агрегирование данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Методические указания по программному обеспечению Pandas / Сост. И. В. Кузнецов. — СПб.: Университет ИТМО, 2021. — 84 с.;
2. Методические указания по программному обеспечению Seaborn / Под ред. А. С. Павлова. — М.: Изд-во НИУ ВШЭ, 2021. — 92 с.;
3. Методические указания по программному обеспечению Plotly / Сост. Н. Ю. Смирнова. — Новосибирск: НГУ, 2020. — 76 с.;
4. Абдрахманов М. И. Pandas. Работа с данными: пособие. — 2-е изд. — М.: Изд-во «Альфа-пресс», 2020. — 171 с.;
5. Крылов П. Н. Python. Визуализация данных: Matplotlib, Seaborn, Mayavi: справочное пособие. — Екатеринбург: УРФУ, 2020. — 254 с.;
6. DevPractice. [Электронный ресурс] — URL: <https://devpractice.ru> (дата обращения: 30.10.2024);
7. CodeRoad Wiki. [Электронный ресурс] — URL: <https://coderoad.wiki> (дата обращения: 30.10.2024);
8. DelftStack. [Электронный ресурс] — URL: <https://www.delftstack.com> (дата обращения: 30.10.2024).