|  |  |
| --- | --- |
|  | **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\_\_\_\_\_\_\_\_\_\_\_ «Coffee Shop Sales» \_

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

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

Студент ИУ5-52Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Насруллаев А.К.

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

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

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(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 «Coffee Shop Dataset» \_\_

Студент группы \_\_\_\_\_\_\_ИУ5-52Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Насруллаев Арсен Камильевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(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 «Coffee Shop Dataset», провести визуализацию полученных данных посредством языка программирования Python*

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

Расчетно-пояснительная записка на 20 листах формата А4.

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

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

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

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

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

Дата выдачи задания «10» сентября 2024 г.

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

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Насруллаев А.К.

(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 “Coffee Shop Dataset”. 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 транзакциях в кофейнях. В конце будет сделан вывод о проделанной работе с данным датасетом.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_Toc183866530)

[1. Определение данных для анализа 6](#_Toc183866531)

[2. Описание данных 7](#_Toc183866532)

[3. Формулирование гипотез 8](#_Toc183866533)

[4. Подготовка данных для работы 9](#_Toc183866534)

[5. Изучение общей информации 10](#_Toc183866535)

[6. Исследовательский анализ данных 11](#_Toc183866536)

[. 11](#_Toc183866537)

[Гипотеза 1 - Продажа кофе составляет основную часть прибыли заведения 11](#_Toc183866538)

[Гипотеза 2 - В заведении, которое находится в сердце Нью-Йорка, средний чек будет выше, чем в остальных заведениях 13](#_Toc183866539)

[Гипотеза 3 - Большая часть покупок происходит утром 15](#_Toc183866540)

[Гипотеза 4 - Горячие напитки чаще покупают в холодное время года, нежели в теплое 16](#_Toc183866541)

[Гипотеза 5 - Люди изредка покупают выпечку вечером 17](#_Toc183866542)

[Анализ корреляции 18](#_Toc183866543)

[ЗАКЛЮЧЕНИЕ 19](#_Toc183866544)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 20](#_Toc183866545)

# **ВВЕДЕНИЕ**

Необходимо исследовать базу данных и выявить закономерности.

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.

# **1. Определение данных для анализа**

В качестве данных для анализа был выбран Dataset «Coffee Shop».

В датасете представлены транзакции и их подробности.

# **Описание данных**

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

* **Transaction ID: номер транзакции**
* **Transaction Date: дата транзакции**
* **Transaction Time: время транзакции**
* **Store Number: номер заведения**
* **Store Location: расположение заведения**
* **Unit price: цена за единицу товара**
* **Product Category: категория продукта**
* **Product Type: тип продукта**
* **Product Name: название продукта**
* **Revenue: сумма транзакции**
* **Month: месяц (числовое значение от 1 до 12))**
* **Weekday: день недели ( числовое значение от 1 до 7)**
* **Hour: час (числовое значение от 1 до 24)**

# **Формулирование гипотез**

Гипотеза 1:Продажа кофе составляет основную часть прибыли заведения

Гипотеза 2: В заведении, которое находится в сердце Нью-Йорка, средний чек будет выше, чем в остальных заведениях

Гипотеза 3: Большая часть покупок происходит утром

Гипотеза 4: Горячие напитки чаще покупают в холодное время года, нежели в теплое

Гипотеза 5: Люди изредка покупают выпечку вечером

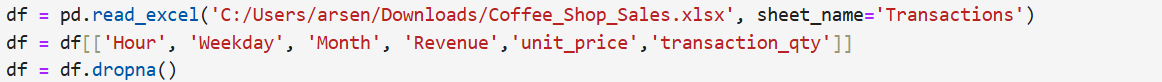
# **Подготовка данных для работы**

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

Проводим очистку данных

Изображение выглядит как текст, снимок экрана, программное обеспечение

Автоматически созданное описание



# **Изучение общей информации**

Рассмотрим доход кофеен с течением времени



Вывод: видим, что с течением времени доход только увеличивался. Возможно, это связано с тем, что со временем тренд на времяпрепровождение в кофейнях возрастает и способствует развитию данного сектора.

Рассмотрим ежемесячный доход

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, линия, диаграмма, снимок экрана

Автоматически созданное описание

Вывод: видим, что ежемесячный доход также растет со временем.

Рассмотрим доход кофеен

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Автоматически созданное описание

Вывод: Доход практически одинаковый, с небольшим отрывом лидирует Hell’s Kitchen.

# **Исследовательский анализ данных**

## **.**

## **Гипотеза 1 - Продажа кофе составляет основную часть прибыли заведения**

Для этого составим круговую диаграмму, в которой будет видно распределение доходов по категориям продуктов

Полученный график выглядит так  
**Изображение выглядит как текст, снимок экрана, диаграмма, Шрифт

Автоматически созданное описание**

**Вывод:**

Таким образом, видим, что основной доход приносит кофе.

Гипотеза 1 – подтверждена

## **Гипотеза 2 - В заведении, которое находится в сердце Нью-Йорка, средний чек будет выше, чем в остальных заведениях**

Для этого составим график, на котором рассмотрим зависимость среднего чека от расположения заведения

**Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание**

Полученный график выглядит так Изображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание

**Вывод:**

На графике видим, что средние чеки примерно равны. Но все же в Манхэттене, ядре Нью-Йорка, средний чек несколько выше, чем в заведениях, расположенных на других локациях

Гипотеза 2 – подтверждена

## **Гипотеза 3 - Большая часть покупок происходит утром**

Для этого составим график корреляции между временем суток и суммой выручки. Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Полученный график выглядит так

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

**Вывод:**

Видим, что корреляция отрицательная: чем позже время суток, тем меньше доход. Отсюда делаем вывод, что утром происходит большая часть транзакций.

Гипотеза 3 – подтверждена.

## **Гипотеза 4 - Горячие напитки чаще покупают в холодное время года, нежели в теплое**

Для этого составим график зависимости количества покупок горячих напитков от месяца

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Полученный график выглядит так

Изображение выглядит как текст, График, диаграмма, линия

Автоматически созданное описание

**Вывод:**

Неожиданно, но в весенние месяцы количество покупок горячих напитков сильно больше, чем в зимнее время. Возможно, причиной является то, что люди чаще гуляют, попутно посещая кофейни.

Гипотеза 4 – опровергнута.

## **Гипотеза 5 - Люди изредка покупают выпечку вечером**

Для этого составим график, показывающий зависимость выручки с выпечки от времени дня

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Полученный график выглядит так

Изображение выглядит как диаграмма, текст, линия, График

Автоматически созданное описание

**Вывод:**

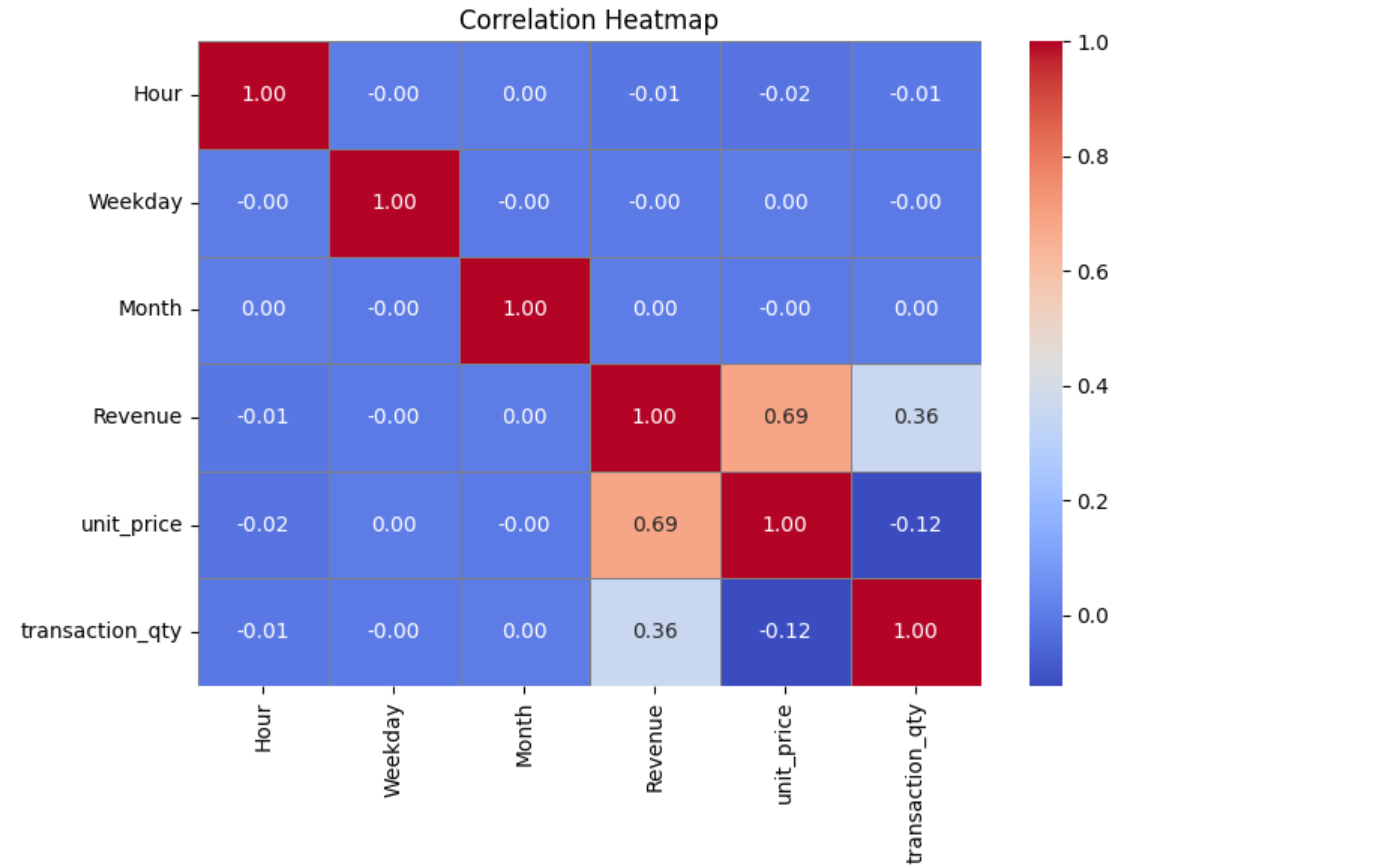
Видим, что люди покупают выпечку в первой половине дня, даже преимущественно утром.

Гипотеза 5 – подтверждена.

## **Анализ корреляции**

**Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание**

****

**Выводы корреляционного анализа:**

* 1. Доход сильно коррелирует с ценой товара
  2. Доход коррелирует с количеством позиций в чеке

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «Coffee Shop Dataset», выявлены важные статистические данные.

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. БУРНАШЕВ Р.А. Методические указания по программному обеспечению «Pandas». - 1 изд. - Казань: КФУ, 2022. - 25 с.
2. Методические указания по программному обеспечению «Seaborn» // [Электронный ресурс]: seaborn.pydata URL: https://seaborn.pydata.org/tutorial.html (дата обращения: 14.09.2024).
3. Методические указания по программному обеспечению «Plotly» // [Электронный ресурс]: habr.com URL: https://habr.com/ru/companies/skillfactory/articles/506974/ (дата обращения: 15.09.2024).
4. «Pandas. Работа с данными» // [Электронный ресурс]: vk.com URL: https://vk.com/wall-159224823\_92912 (дата обращения: 15.09.2024).
5. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi» // [Электронный ресурс]: vk.com URL: https://vk.com/wall-192648009\_320 (дата обращения: 15.09.2024).