|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **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\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_Система оперативного анализа данных \_\_\_\_\_***

***\_\_\_\_\_\_\_«Пицца в американских заведениях»\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_ИУ5-53Б\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Д.К. Пермяков\_\_\_\_\_**

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

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

(Подпись, дата) (И.О.Фамилия)

*2023 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 \_\_\_ИУ5\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_В.И. Терехов\_\_

(И.О.Фамилия)

«\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_\_\_\_Анализ данных по базе данных «Pizza restaurants and the pizza they sell». \_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_ИУ5-53Б\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Пермяков Дмитрий Кириллович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ИССЛЕДОВАТЕЛЬСКАЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

График выполнения НИР: 25% к \_\_\_\_ нед., 50% к \_\_\_\_ нед., 75% к \_\_\_ нед., 100% к \_\_\_\_ нед.

***Техническое задание*** \_\_\_\_*Спроектировать систему анализа и визуализации по теме*\_\_\_\_\_\_\_\_\_\_\_\_ *“Пицца в американских заведениях”, провести визуализацию полученных данных посредством языка программирования Python.*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

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

Расчетно-пояснительная записка на \_\_16\_ листах формата А4.

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

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

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

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

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

Дата выдачи задания «\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

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

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_**Д.К. Пермяков**\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

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



**Содержание**

[Введение 4](#_bookmark0)

[Цель 4](#_bookmark1)

[Задачи 4](#_bookmark2)

[Определение данных для анализа 5](#_bookmark3)

[Формулирование гипотез 5](#_bookmark4)

[Анализ данных 7](#_bookmark5)

[Загрузка данных в Python 7](#_bookmark6)

[Очистка данных 9](#_bookmark7)

[Агрегирование данных 9](#_bookmark8)

[Анализ поставленных гипотез 11](#_bookmark9)

[Корреляционный анализ 16](#_bookmark10)

[Заключение 17](#_bookmark11)

[Список используемых источников 17](#_bookmark12)

# Введение

## Цель:

Исследовать базу данных и изучить данные о продажах пиццы в американских заведениях, выявить зависимость продажи различных видов пиццы от ресторанов и местоположения.

## Задачи:

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. Корреляционный анализ
11. Формулирование ограничений и выводов.

# Определение данных для анализа

Для анализа был выбран DataSet «Pizza restaurants and the pizza they sell». Набор данных содержит информацию об американских заведениях, в которых продаются разные виды пиццы, об их местоположении, о цене пиццы и её описании.

Таблица имеет 3510 записей. Каждая запись содержит информацию об адресе ресторана, его категории, городе, штате/провинции, широте и долготе, валюте, цене, названии ресторана, названии пиццы, описании пиццы, минимальной и максимальной цене пиццы.

# Формулирование гипотез

1. Самой популярной пиццей в Америке является “Маргарита”.
2. Наибольшее количество пиццы продаётся на востоке США.
3. Самая дорогая пицца продаётся в Аляске.
4. Больше всего видов пиццы продаётся в Калифорнии.

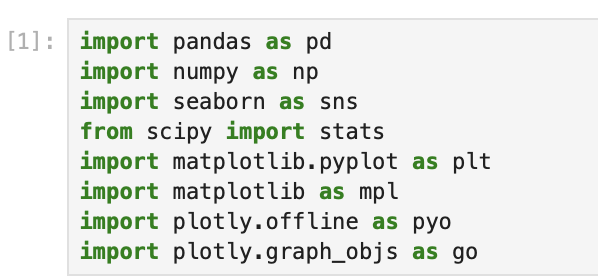
**Обоснование сформулированных гипотез:**

* 1. Самой популярной пиццей в Америке является “Маргарита”. Эта пицца является одним из классических и простых блюд с минимумом ингредиентов и универсальным вкусом, поэтому вполне логично предположить, что она будет самой популярной.
  2. Наибольшее количество пиццы продаётся на востоке США. В начале 20 века восточное побережье США привлекало значительное количество итальянских иммигрантов. Итальянская кухня, включая пиццу, стала частью местной кулинарной культуры под влиянием этой иммиграции.
  3. Самая дорогая пицца продаётся в Аляске. Это связано с удалённым местоположением этого штата и суровыми климатическими условиями, что влияет на доступность ингредиентов стоимость их транспортировки из других регионов.
  4. Больше всего видов пиццы продаётся в Калифорнии. Кроме того, что это самый населённый штат в США, Калифорния известна своим культурным разнообразием, что отражается и на продаваемой еде.

# Анализ данных

Для анализа данных было использовано программное обеспечение

«Pandas», «numpy», «matplotlib», «seaborn», «plotly» и «scipy».



**Загрузка данных в Python**

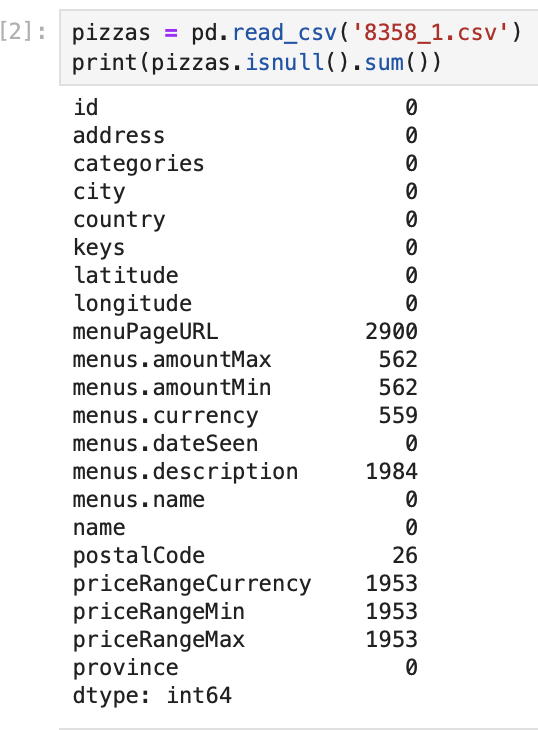
Для загрузки .csv файла с данными в pandas была использована функция read\_csv().

pizzas = pd.read\_csv('8358\_1.csv')

**Проверка данных**

Набор данных был проверен на наличие отсутствия данных по столбцам.

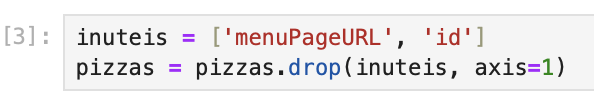
Для этого был использован метод isnull() и sum() для объектов pizzas.



**Очистка данных**

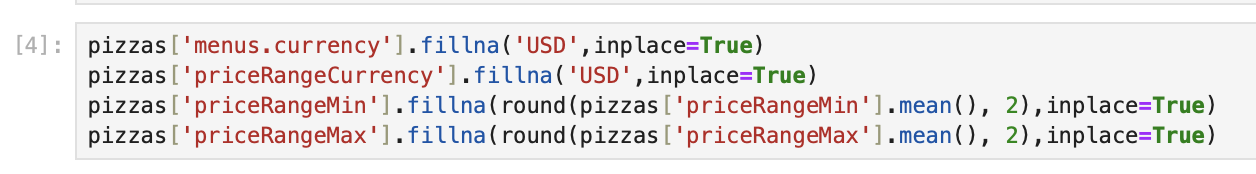
**Удаление ненужных столбцов.**

При помощи метода drop удалим столбцы “menuPageURL” и “id”, так как в анализе они не используются.



**Заполнение пропущенных значений.**

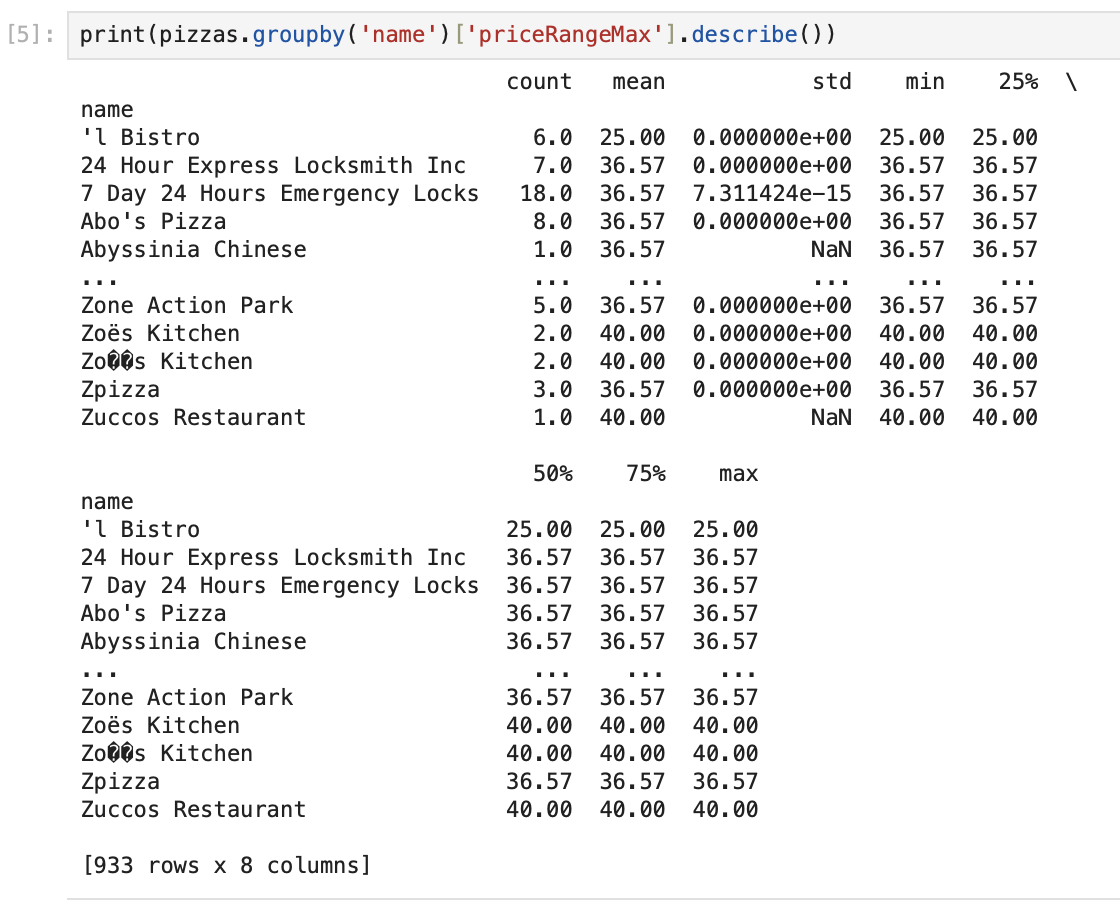
Заменим пустые ячейки в столбцах “menus.currency” и “priceRangeCurrency” на “USD”, а также пустые значения в столбцах “priceRangeMin” и “priceRangeMax” заменим на средние значения этих столбцов с помощью метода fillna():



**Агрегирование данных**

**Нахождение аномалий при помощи описательной статистики**

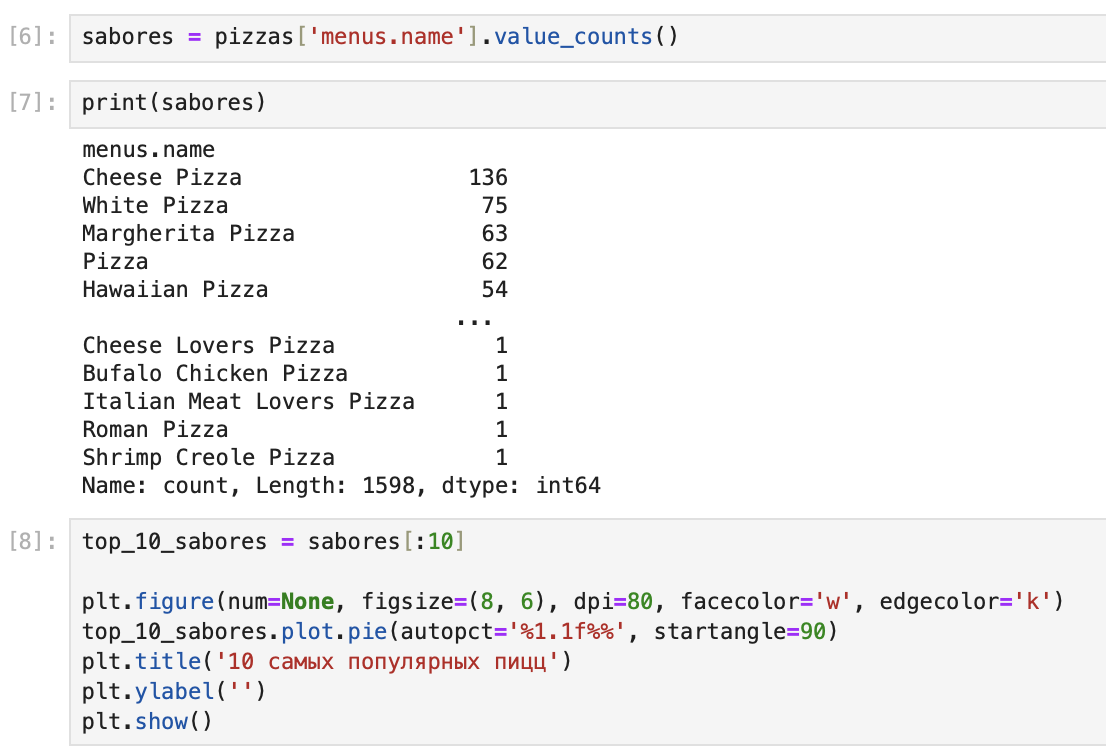
При помощи метода describe() было обнаружено, что в датасете представлено 933 ресторана. Нахождение аномалий для такого большого объёма данных не будет являться объективным, так как python не может вывести все 933 строки. Но из имеющихся данных видим, что самая дорогая пицца стоит 40 долларов, больше всего видов пиццы продаётся в ресторане “7 Day 24 Hours Emergency Locks”.



**Анализ поставленных гипотез**

1. **Самой популярной пиццей в Америке является “Маргарита”.**

Отобразим 10 самых продаваемых вкусов пиццы на круговой диаграмме:

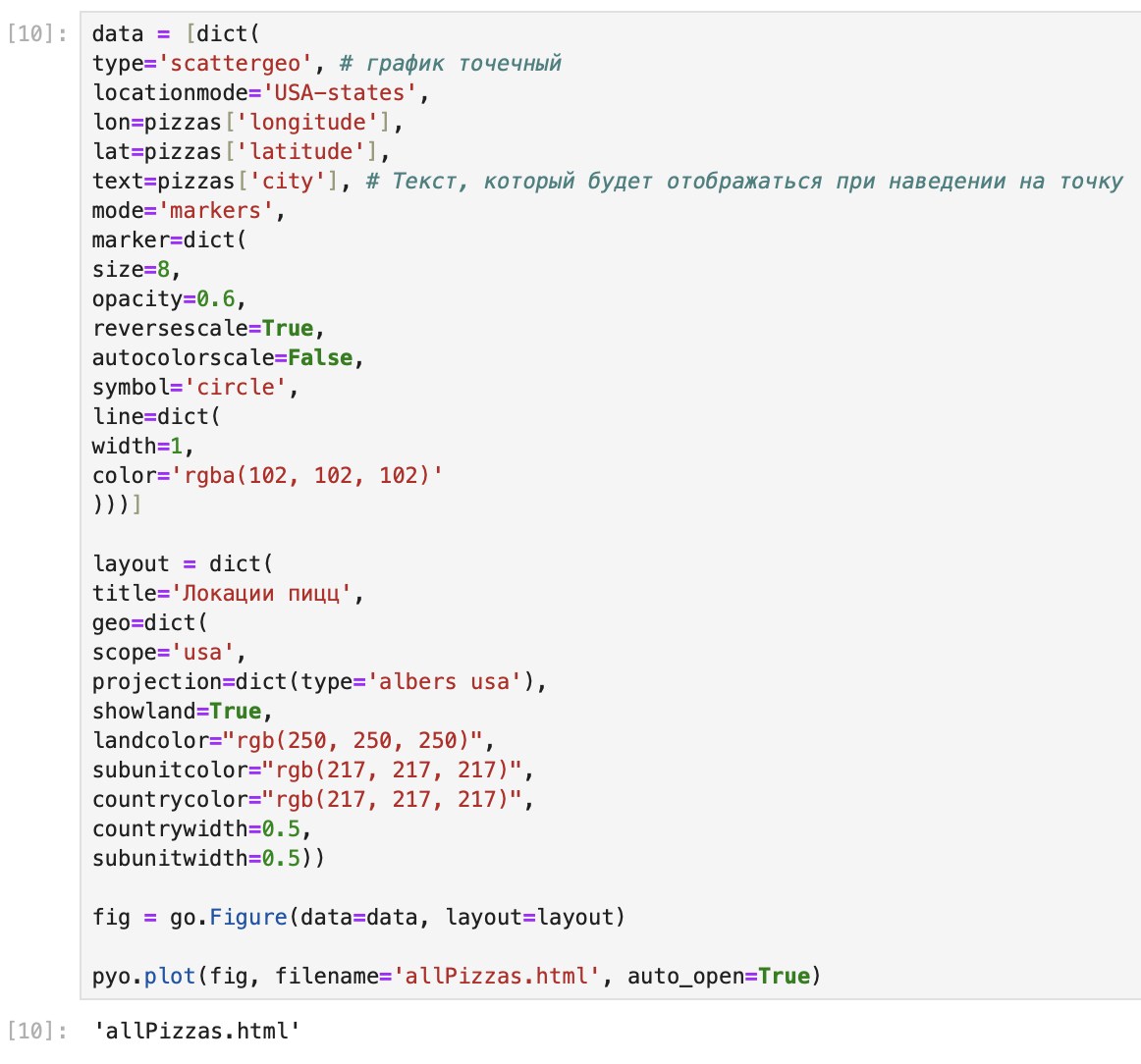


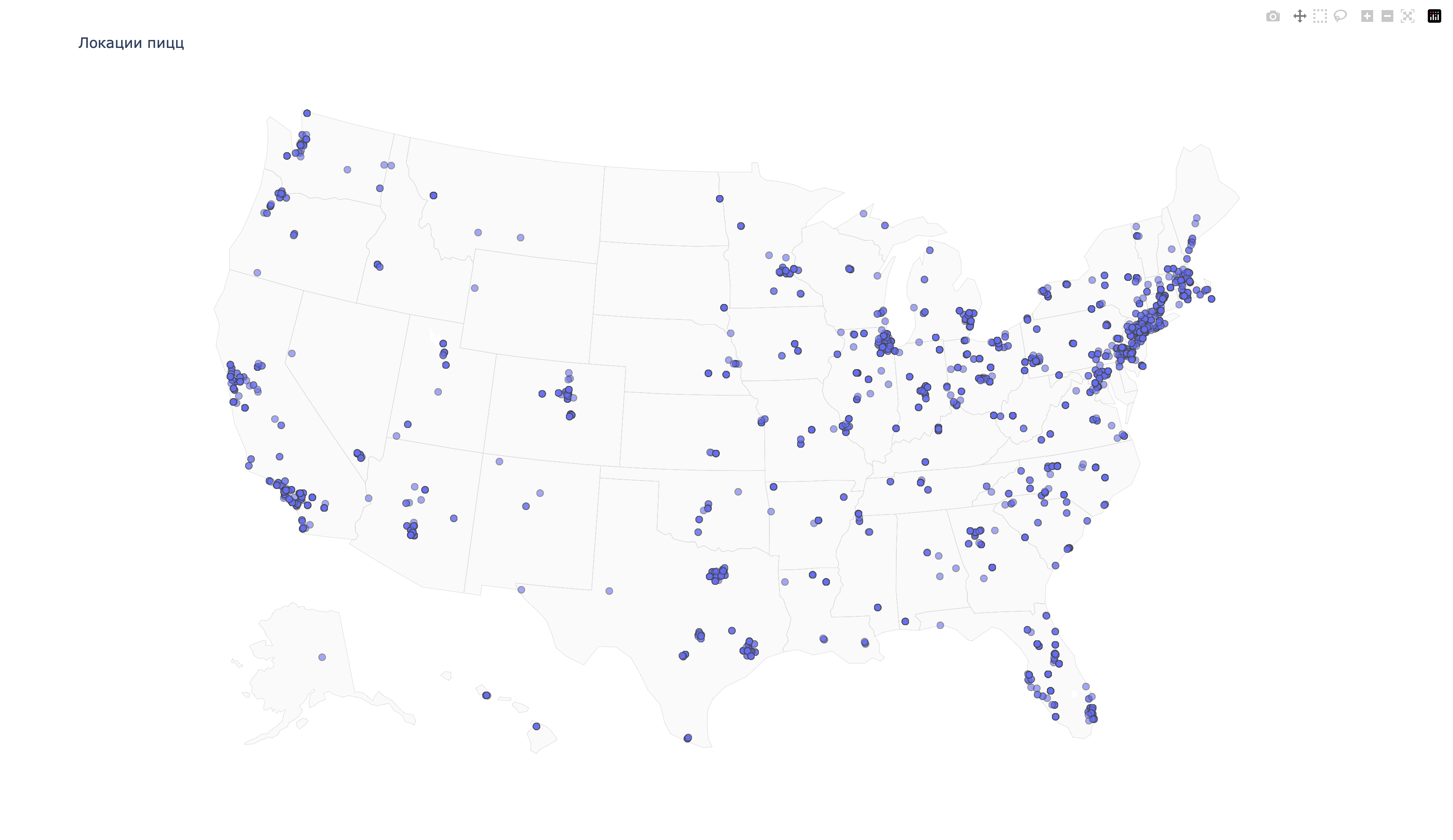


Посмотрим на диаграмму. Видно, что самым популярным видом пиццы является сырная пицца. А “Маргарита” только на третьем месте. Таким образом, гипотеза о том, что самой популярной пиццей в Америке является “Маргарита”, оказалась неверна.

1. **Наибольшее количество пиццы продаётся на востоке США.**

Для проверки данной гипотезы мы можем воспользоваться библиотекой plotly, а именно вывести интерактивную карту с нанесёнными на неё точками ресторанов.

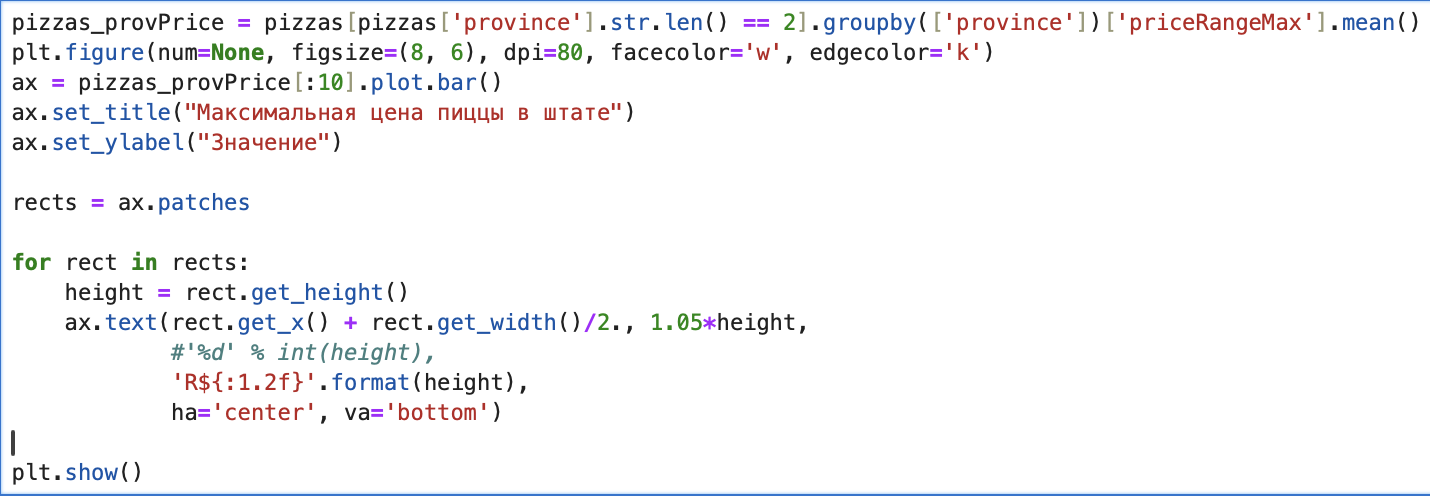




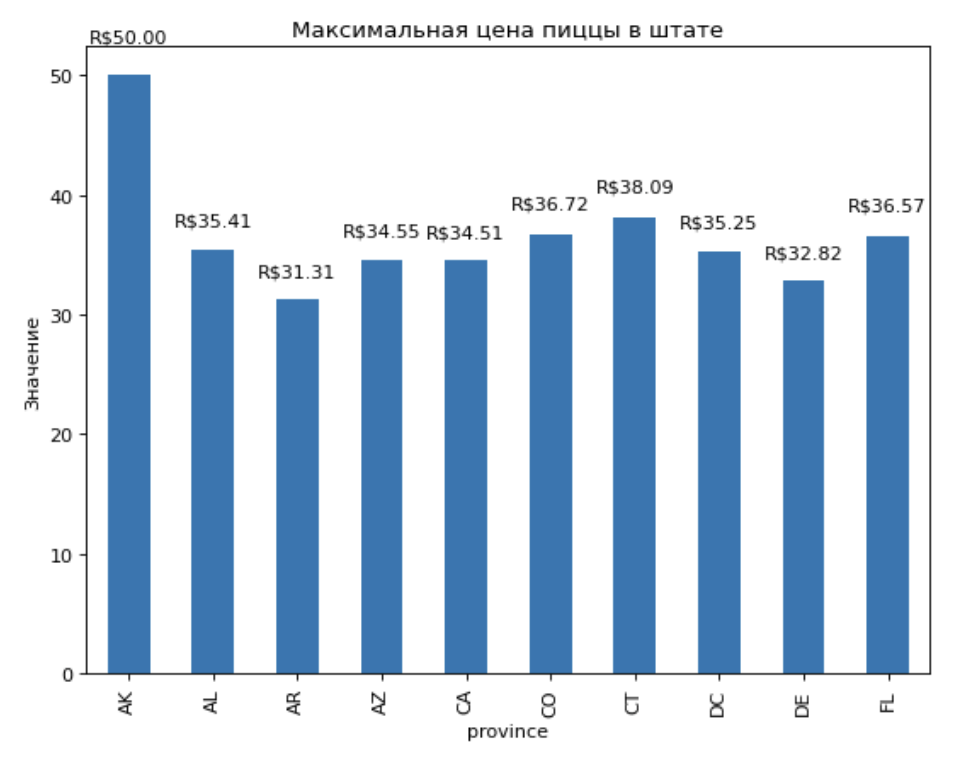
Судя по полученной карте, больше всего пиццы действительно продаётся на востоке США. Следовательно, гипотеза верна.

1. **Самая дорогая пицца продаётся в Аляске.**

Чтобы проверить эту гипотезу, можно построить столбчатую диаграмму.



На диаграмме хорошо видно, что самая дорогая пицца стоит 50 долларов и она продаётся в Аляске, причём она там стоит намного дороже, чем в остальных штатах.

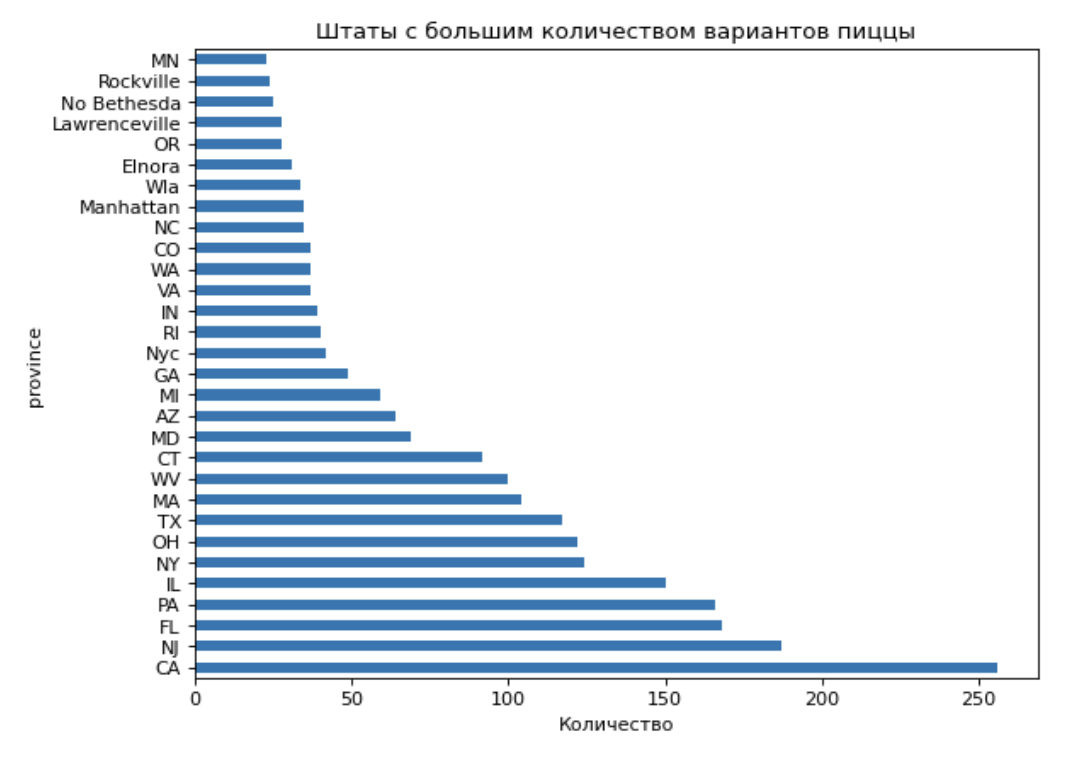


Таким образом, гипотеза о том, что самая дорогая пицца продаётся в Аляске, оказалась верна.

1. **Больше всего видов пиццы продаётся в Калифорнии.**

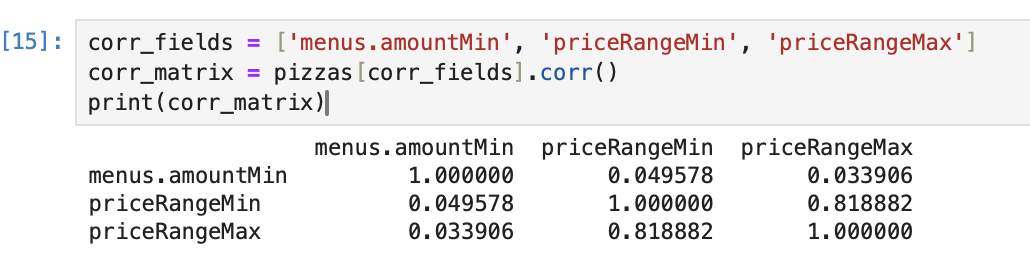
Чтобы проверить эту гипотезу, воспользуемся горизонтальной столбчатой диаграммой.





На диаграмме легко можно заметить, что в Калифорнии, действительно, продаётся больше всего видов пиццы. Значит, гипотеза оказалась верна.

**Корреляционный анализ**



Рассмотрим связки данных:

* 1. Menus.amountMin – priceRangeMin переменные связаны коэффициентом 0,049578. Следовательно, между минимальной ценой в меню ресторана и минимальной ценой пиццы в ресторане существует слабая положительная связь.
  2. Menus.amountMin – priceRangeMax переменные связаны коэффициентом 0,033906. Следовательно, между минимальной ценой в меню ресторана и максимальной ценой пиццы существует слабая положительная связь.
  3. priceRangeMin – priceRangeMax переменные связаны коэффициентом 0.818882. Следовательно, между минимальной ценой пиццы и максимальной ценой пиццы в ресторане существует сильная положительная связь.

В качестве вывода можно сказать, что существенной связью обладают минимальная и максимальная цена пиццы в ресторане: если минимальная цена для пиццы увеличивается, то увеличится и максимальная цена для этой пиццы.

# Заключение

В ходе научно-исследовательской работы был проведен анализ базы данных «Pizza restaurants and the pizza they sell», выявлены важные статистические данные, которые можно использовать при желании поесть пиццу в Америке.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas», «matplotlib.pyplot», «seaborn» и «scipy» на базе языка программирования Python.

# Список используемых источников

1. Лекции по предмету «Оперативный анализ данных», преподаватель кафедры ИУ5 Маслеников К.Ю.
2. Официальная документация «Pandas»
3. «Pandas. Работа с данными» (2020), Автор: Абдрахманов М. И.
4. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi»
5. <https://www.kaggle.com/datasets/datafiniti/pizza-restaurants-and-the-pizza-they-sell/data>