|  |  |
| --- | --- |
| 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»**

Отчет по лабораторной работе №1

**«Разведочный анализ данных.**

**Исследование и визуализация данных»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б

Перевощиков Н.Д.

подпись, дата

Проверил:

к.т.н., доц., Гапанюк Ю.Е.

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc159534153)

[2. Задание 3](#_Toc159534154)

[3. Текст программы и результаты 3](#_Toc159534155)

[4. Вывод 18](#_Toc159534163)

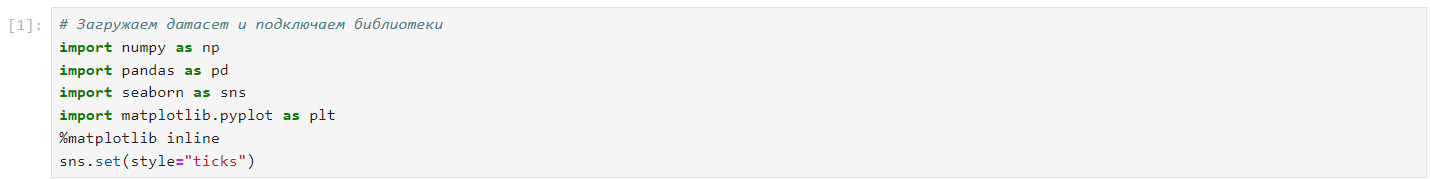
## Цель лабораторной работы

Изучение различных методов визуализация данных.

## Задание

* 1. Выбрать набор данных (датасет).
  2. Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных.
  3. Создать ноутбук, который содержит следующие разделы:
     1. Текстовое описание выбранного Вами набора данных.
     2. Основные характеристики датасета.
     3. Визуальное исследование датасета.
     4. Информация о корреляции признаков.
  4. Сформировать отчет и разместить его в своем репозитории на github.

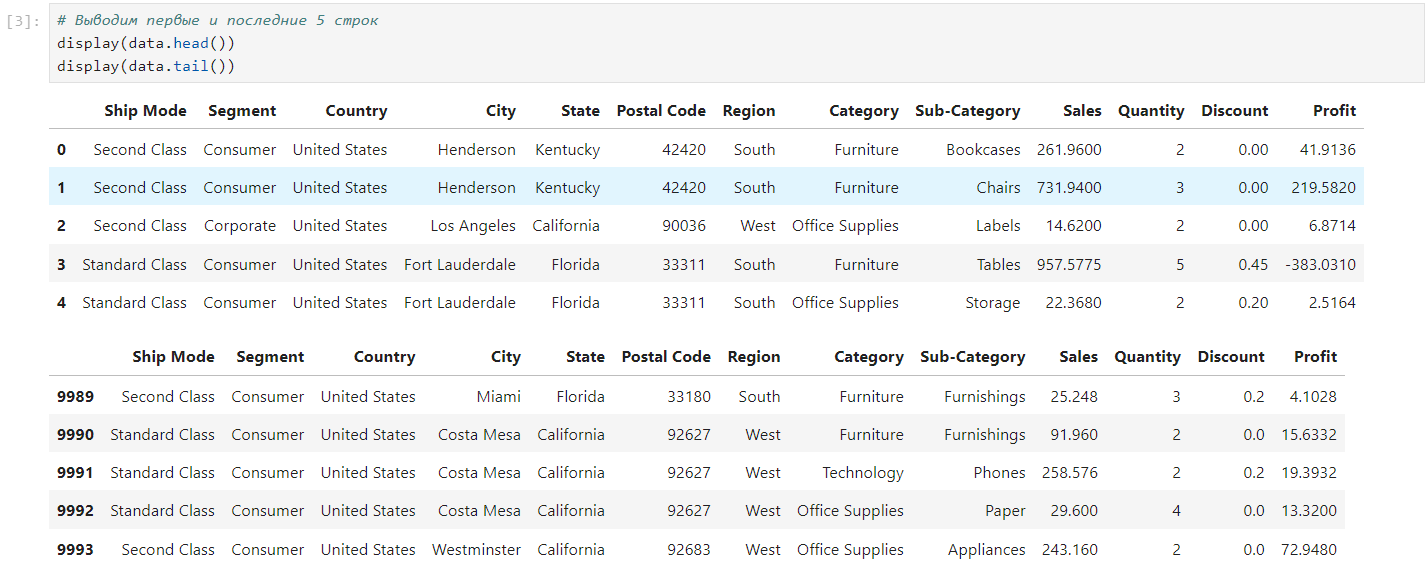
## Текст программы и результаты

Импортируем библиотеки с помощью команды import. Как правило, все команды import размещают в первой ячейке ноутбука, но мы в этом примере будем подключать все библиотеки последовательно, по мере их использования.

Используем библиотеку pandas для чтения данных из CSV-файла 'SampleSuperstore.csv' и загрузки их в объект DataFrame под названием data.



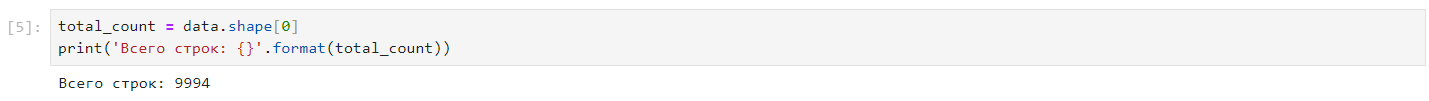
Выводим первые и последние 5 строк.



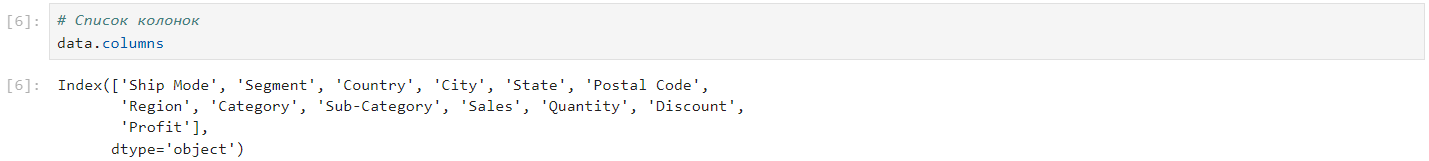
Вводим размер строки и столбца датасета.



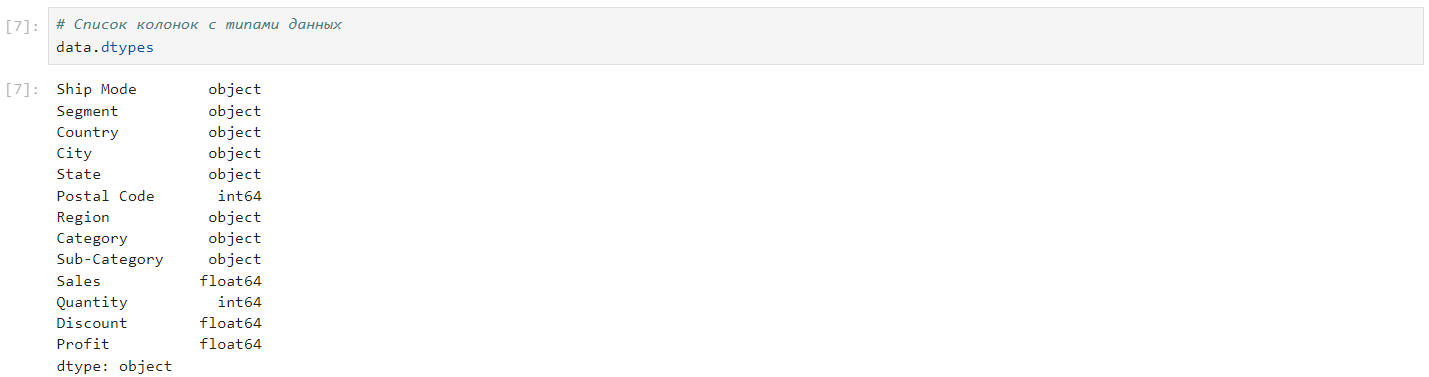
Подсчитаем количество строк.



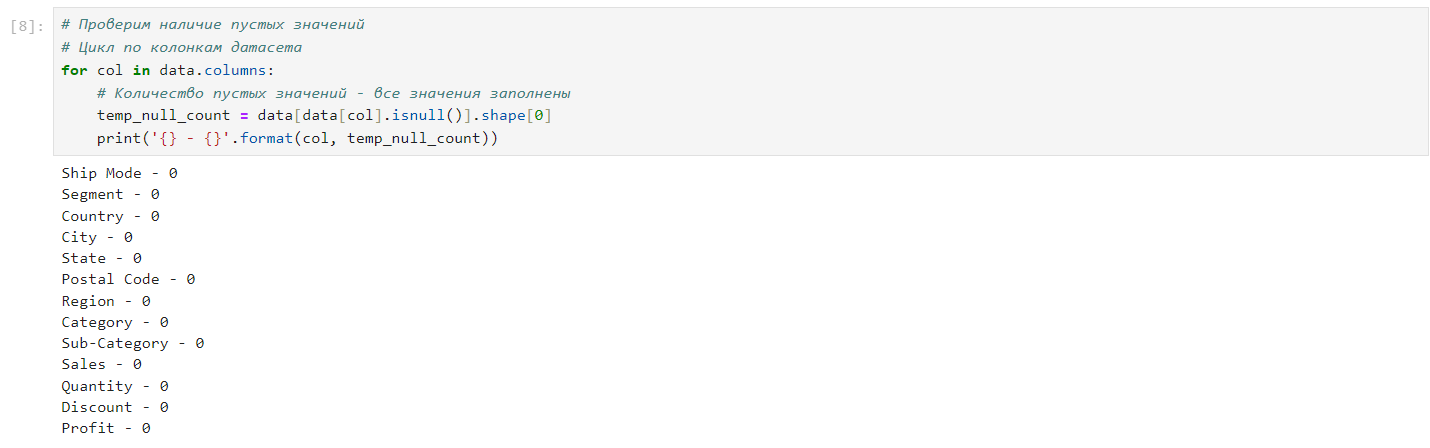
Выводим список колонок.



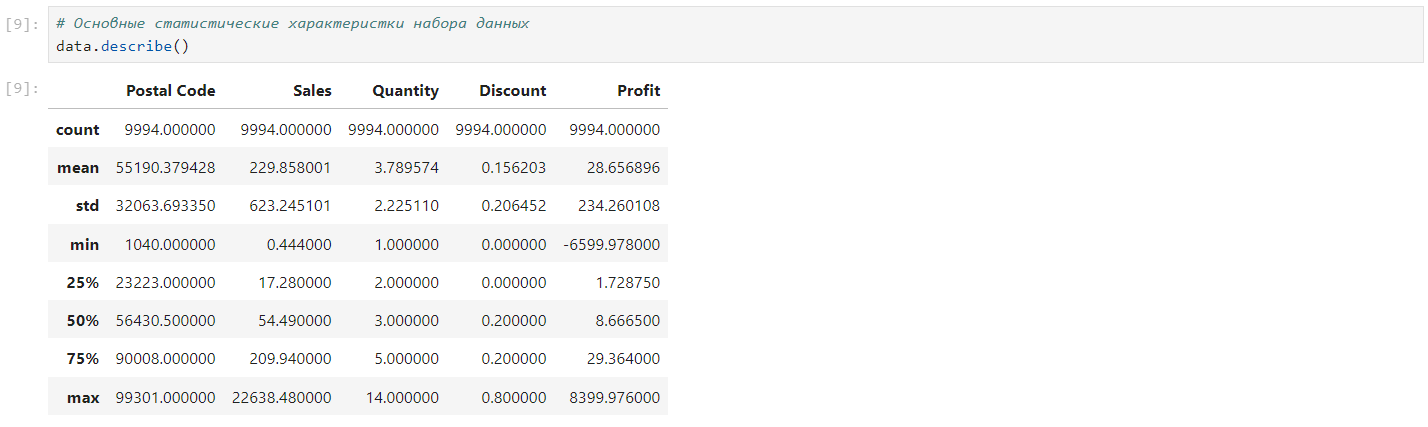
Выводим список колонок с типами данных.



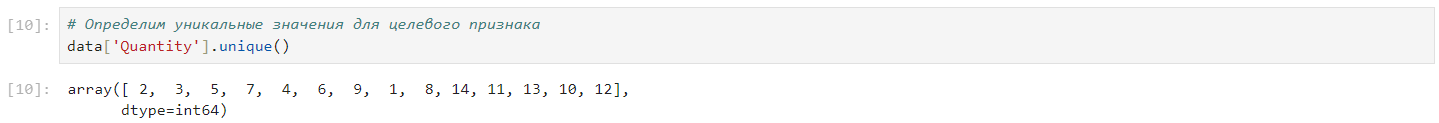
Проверим наличие пустых значений.



Выводим основные статистические характеристики набора данных



Определим уникальные значения для целевого признака



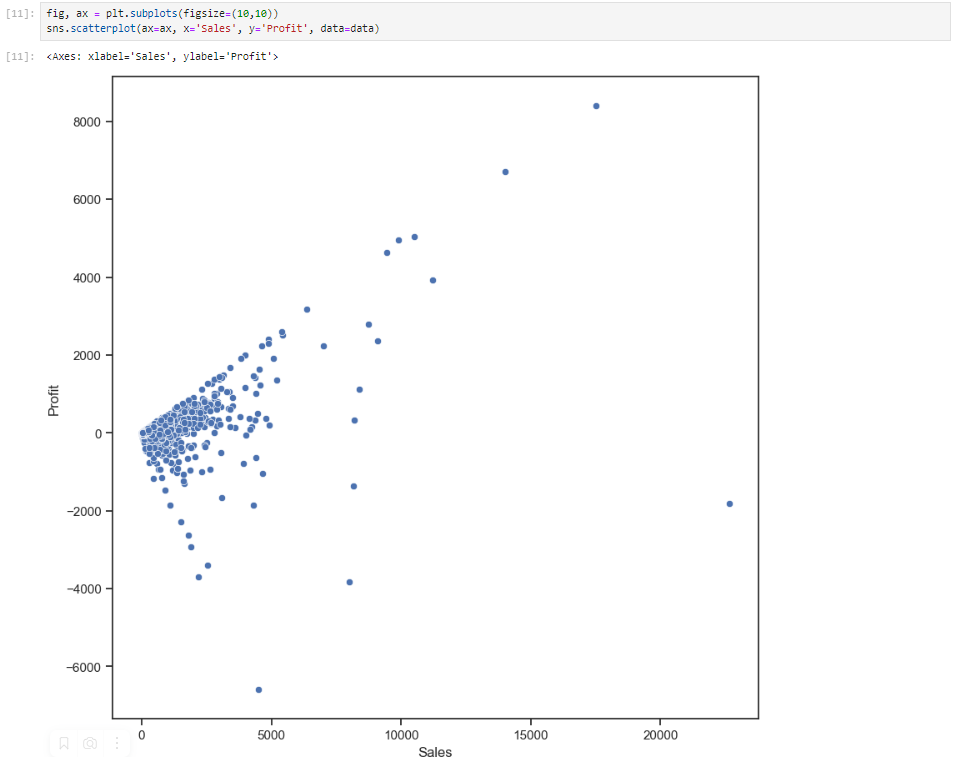
Целевой признак является бинарным и содержит только значения от 2 до 12.

# Визуальное исследование датасета.

Для визуального исследования могут быть использованы различные виды диаграмм, мы построим только некоторые варианты диаграмм, которые используются достаточно часто.

### [Диаграмма рассеяния](https://en.wikipedia.org/wiki/Scatter_plot)

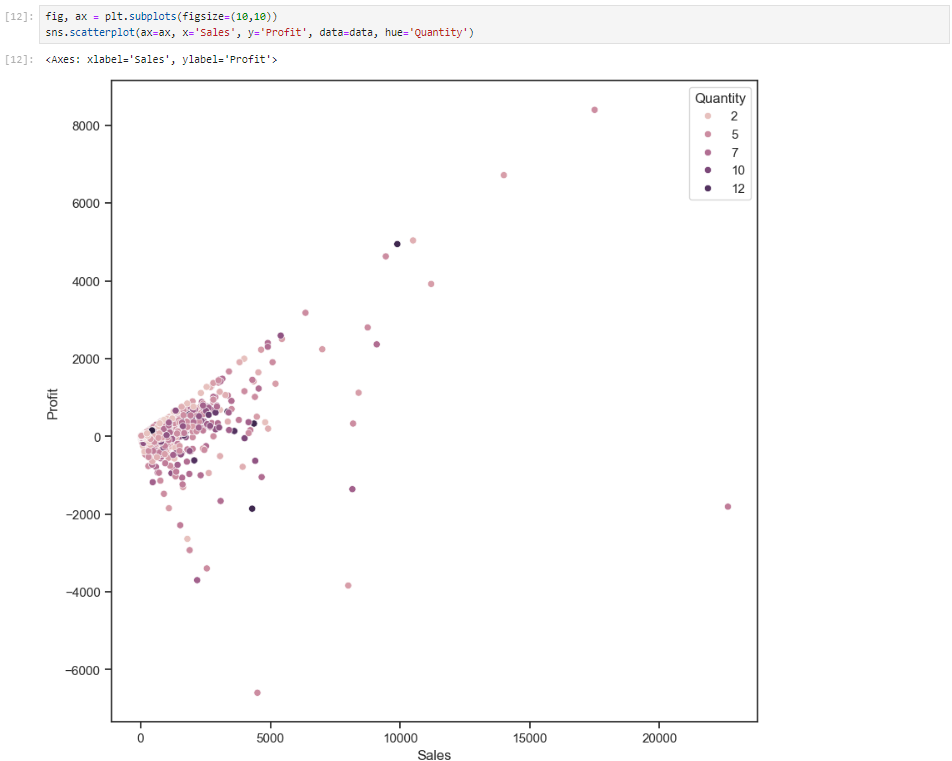
Позволяет построить распределение двух колонок данных и визуально обнаружить наличие зависимости. Не предполагается, что значения упорядочены.



Можно видеть что между полями Sales и Profit пристутствует не совсем линейная зависимость.

Посмотрим насколько на эту зависимость влияет целевой признак.

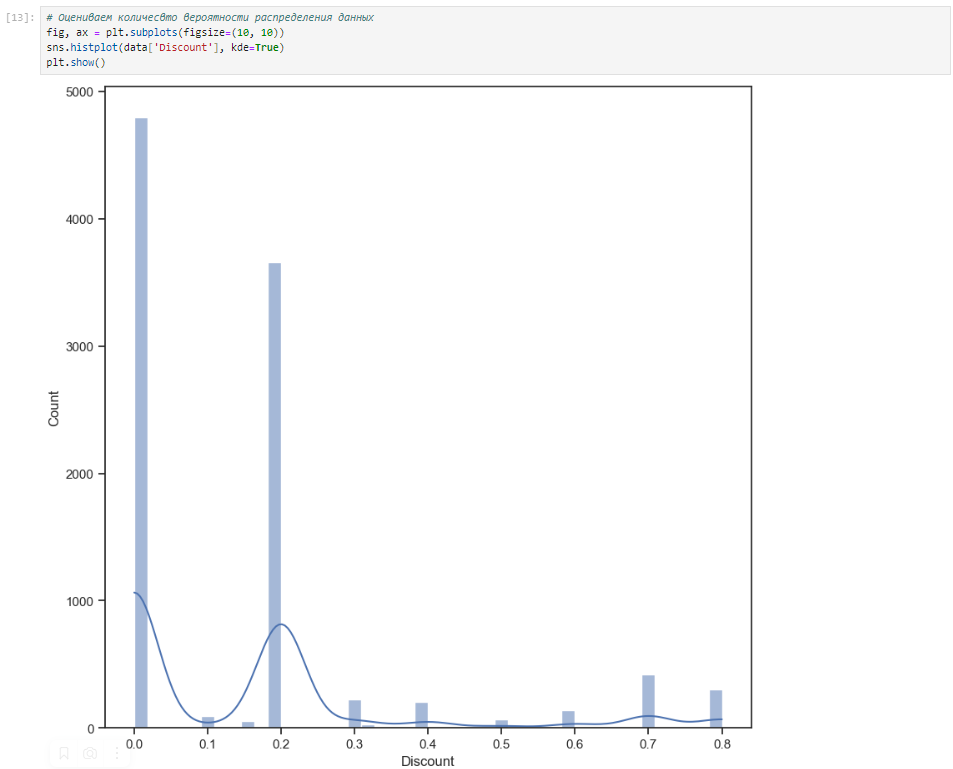
Диаграмма рассеяния, где каждая точка представляет собой товар с определенными значениями продаж, прибыли и количества проданных товаров. Цвет каждой точки показывает количество проданных товаров.



## [Гистограмма](https://en.wikipedia.org/wiki/Histogram)

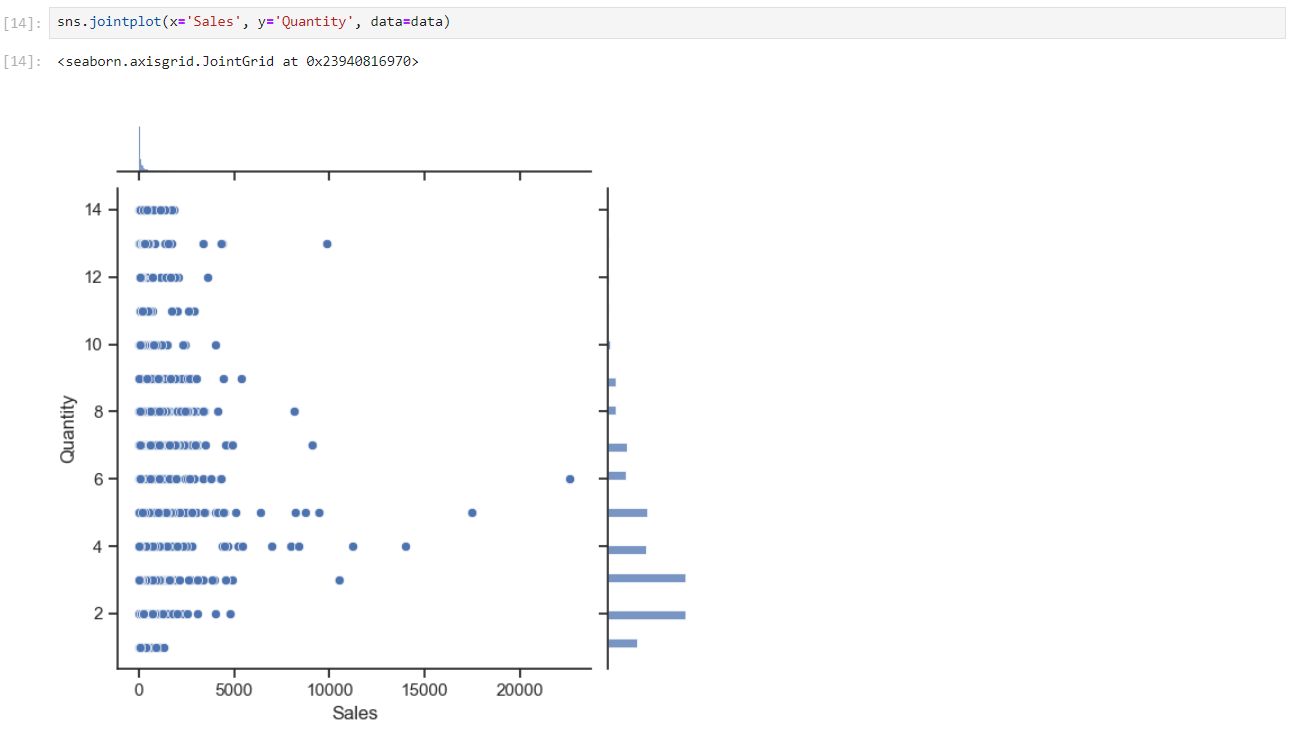
Позволяет оценить плотность вероятности распределения данных.

Оцениваем количество вероятности распределения данных

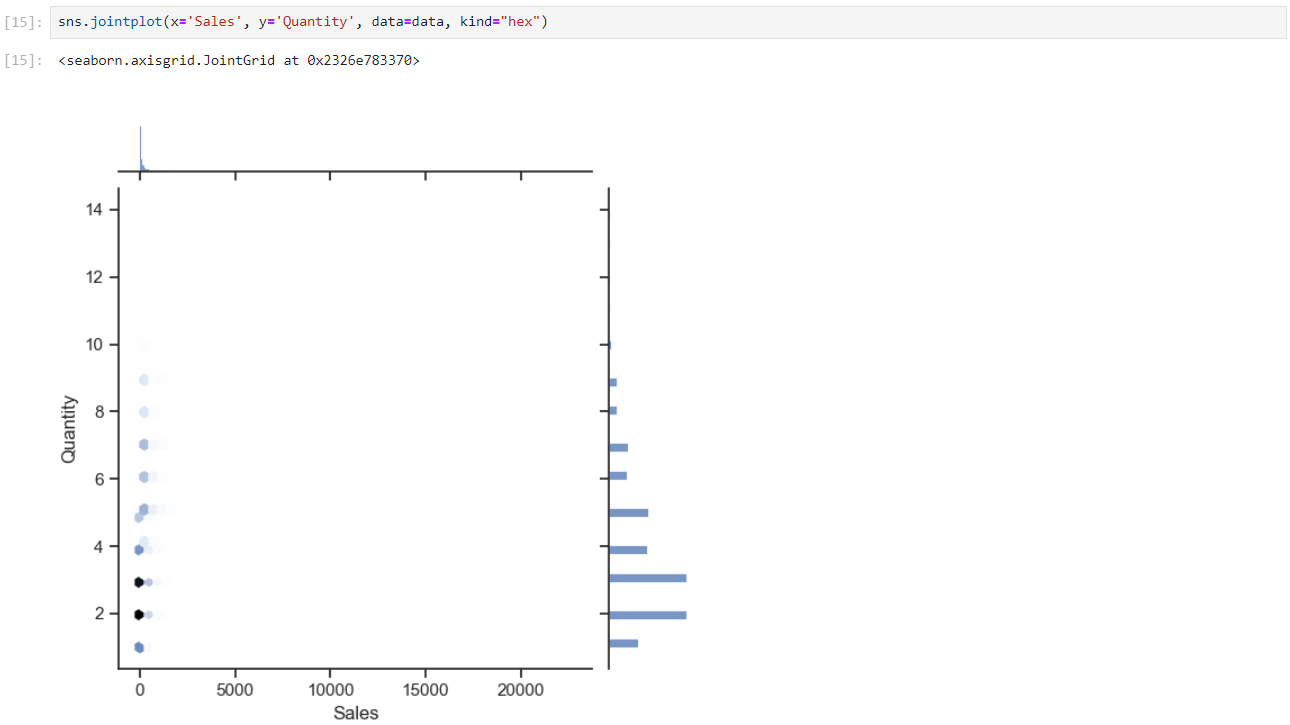


**Jointplot**

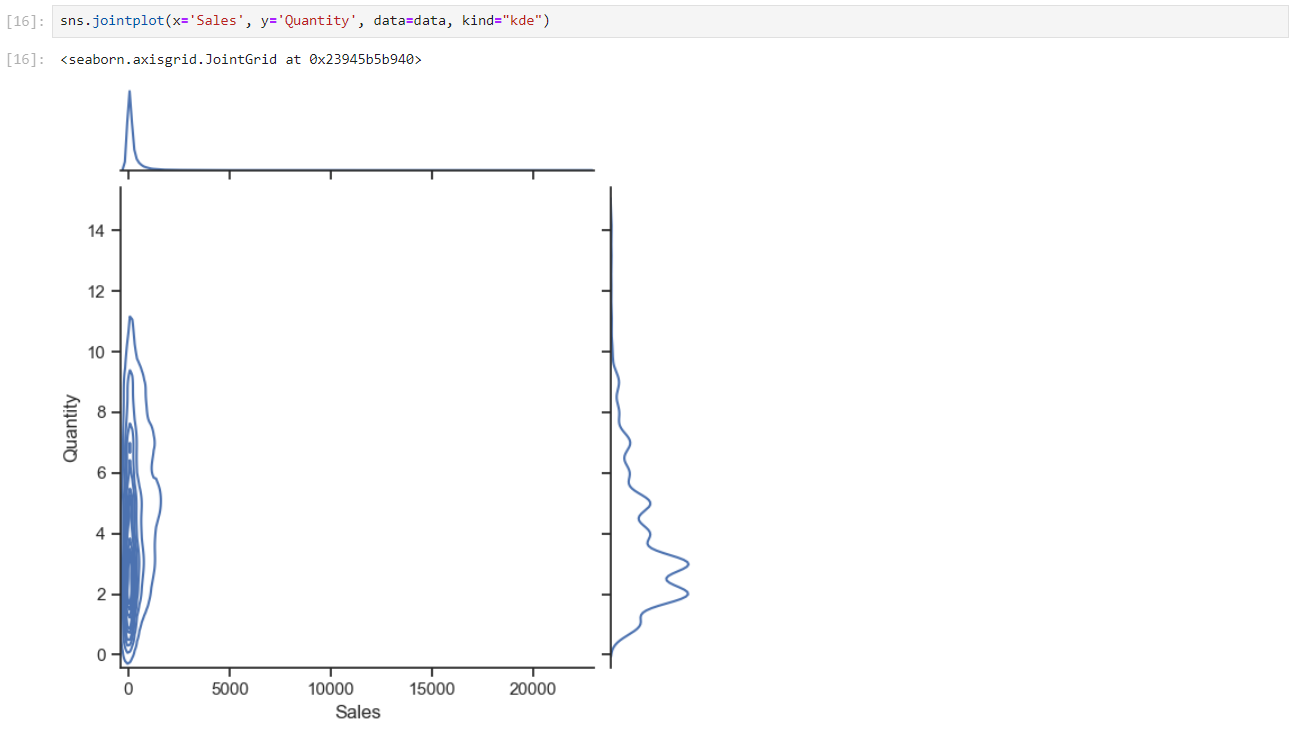
Комбинация гистограмм и диаграмм рассеивания.



Строим совместный график для двух переменных 'Sales' и 'Quantity'. Ось x представляет значения 'Sales', ось y - значения 'Quantity'. В данном случае, kind="hex" указывает использовать шестиугольные ячейки (hexbin) для представления данных, что полезно для визуализации плотности точек.



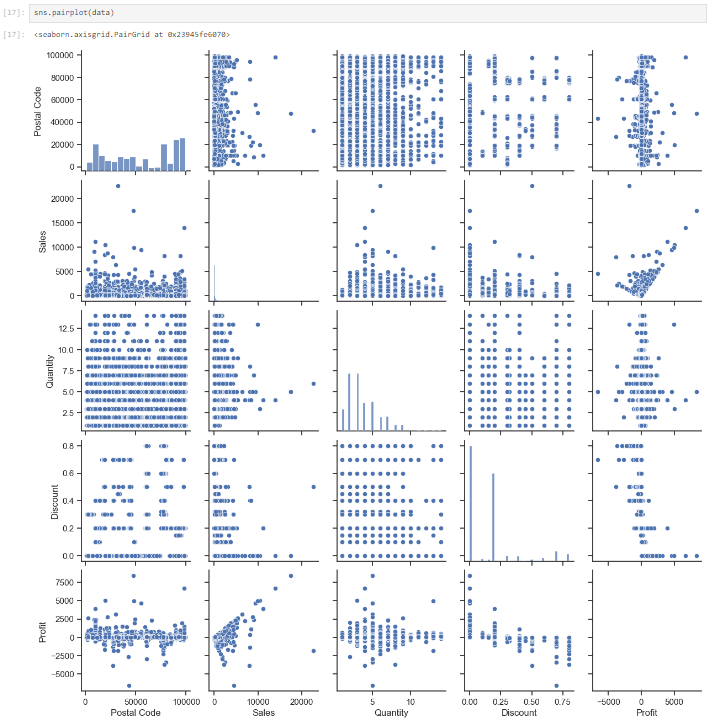
Совместные графики с ядерной оценкой плотности полезны для визуализации распределения и взаимосвязи между двумя переменными. График KDE создает контурные линии, показывающие уровни плотности в различных областях графика. Это позволяет лучше оценить форму распределения и сильные стороны взаимосвязи между 'Sales' и 'Quantity'.



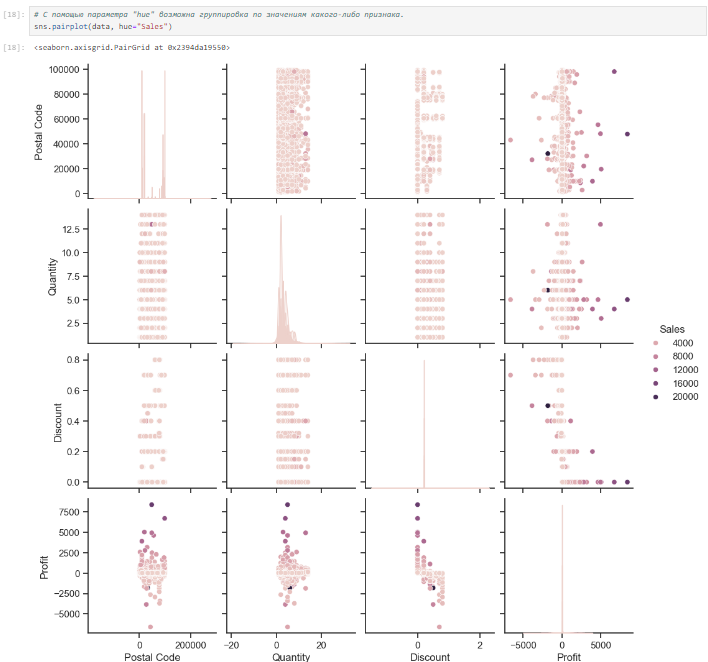
### "Парные диаграммы"

Комбинация гистограмм и диаграмм рассеивания для всего набора данных.

Выводится матрица графиков. На пересечении строки и столбца, которые соответстуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.

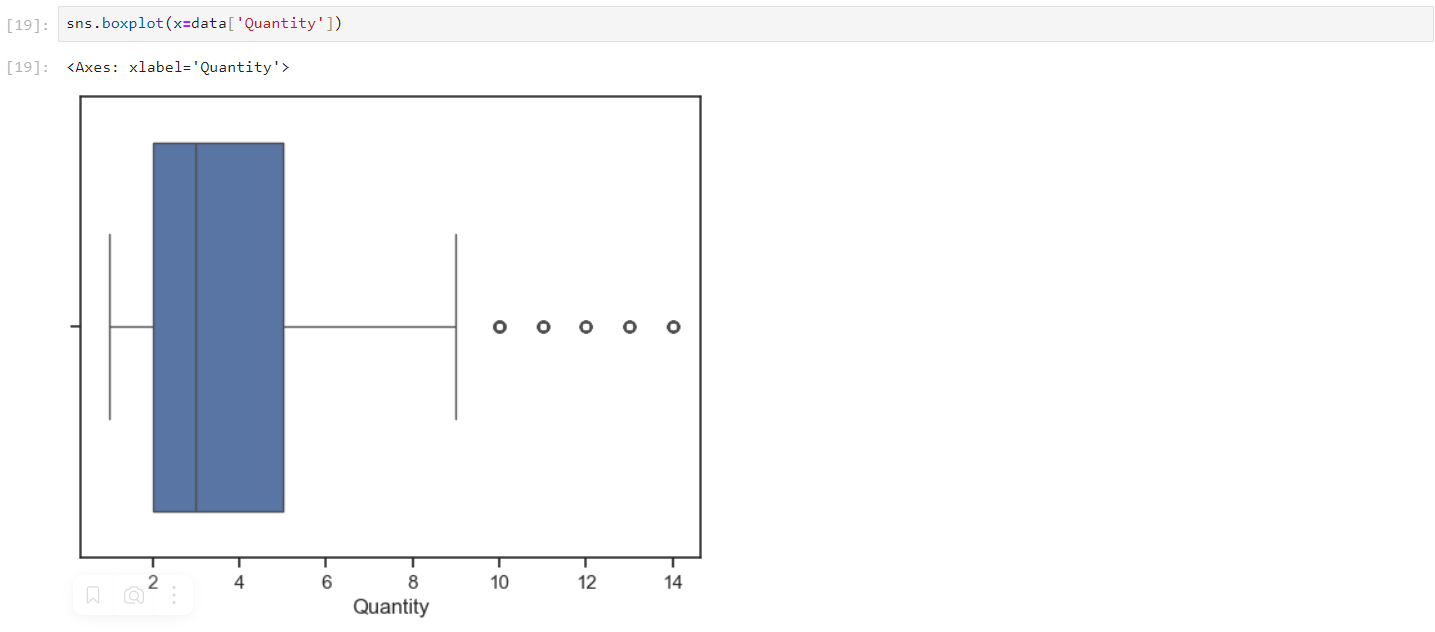
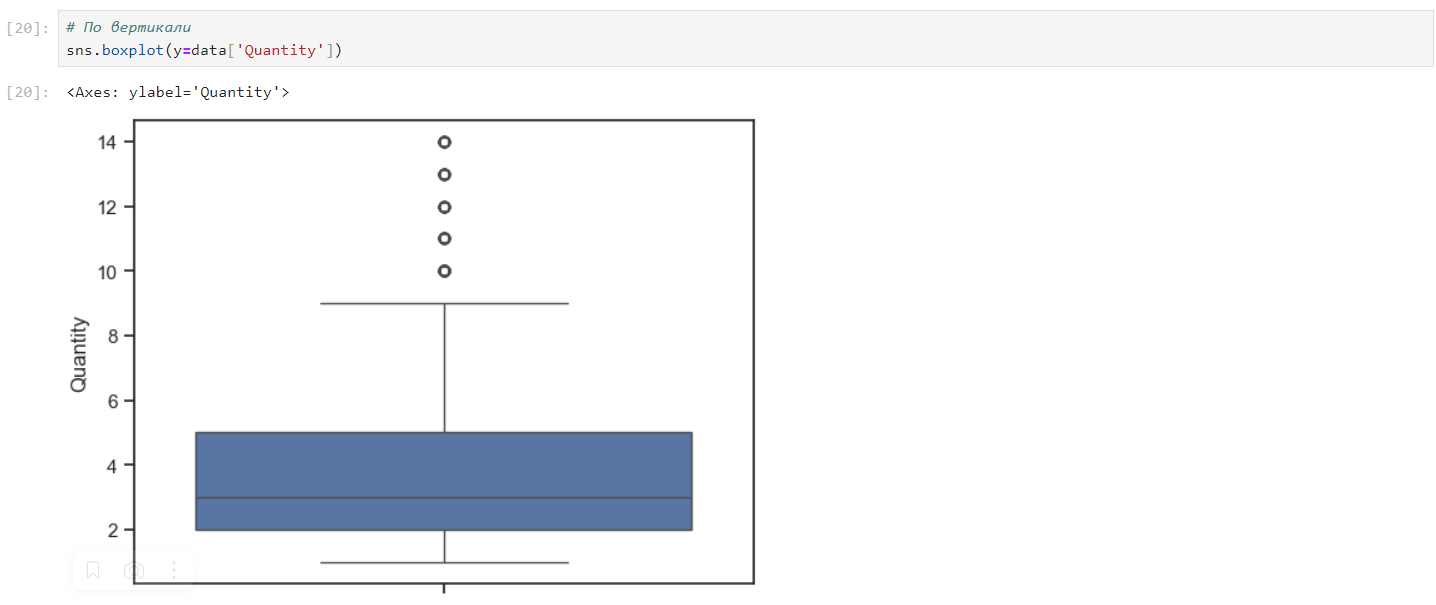


С помощью параметра "hue" возможна группировка по значениям какого-либо признака.



### [Ящик с усами](https://en.wikipedia.org/wiki/Box_plot)

Отображает одномерное распределение вероятности.

### [Violin plot](https://en.wikipedia.org/wiki/Violin_plot)

Похоже на предыдущую диаграмму, но по краям отображаются распределения плотности.

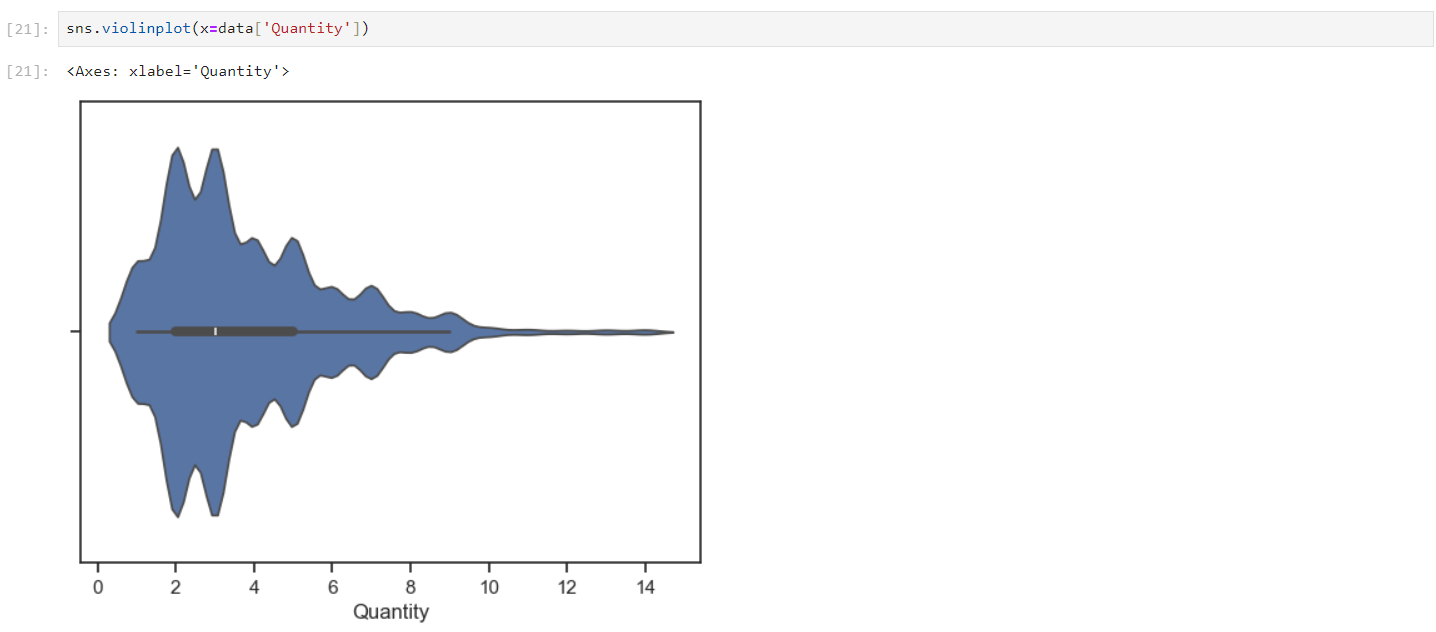
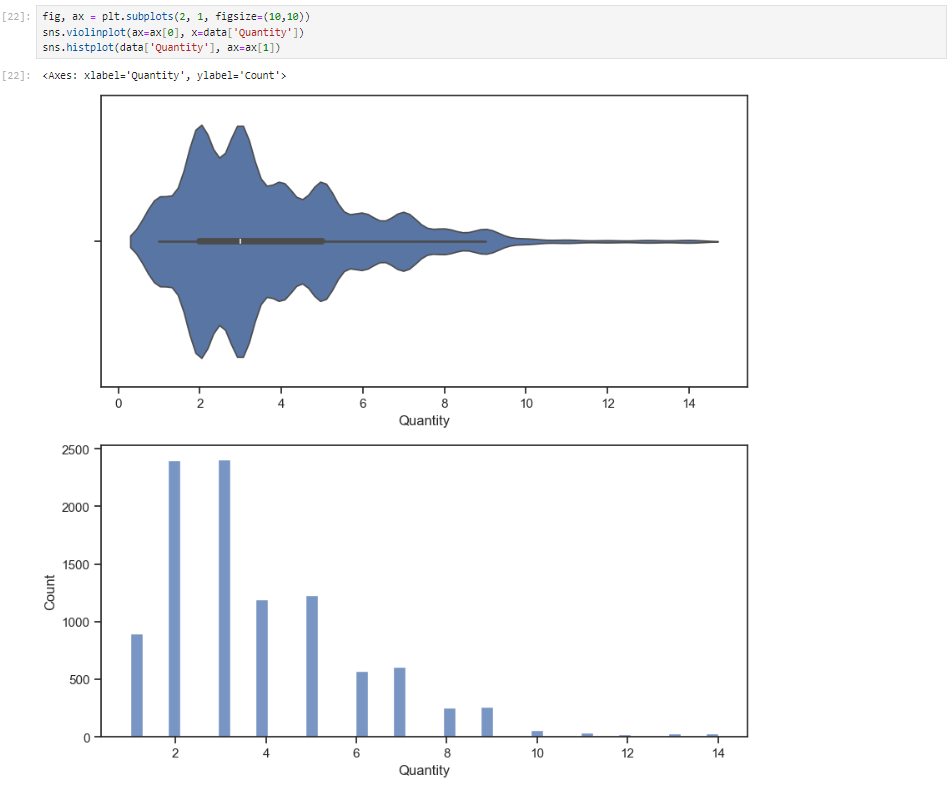


График скрипичного графика в верхнем подграфике и гистограммы в нижнем подграфике для переменной 'Quantity'. Это позволяет визуально оценить распределение и характеристики переменной 'Quantity'.



# Информация о корреляции признаков

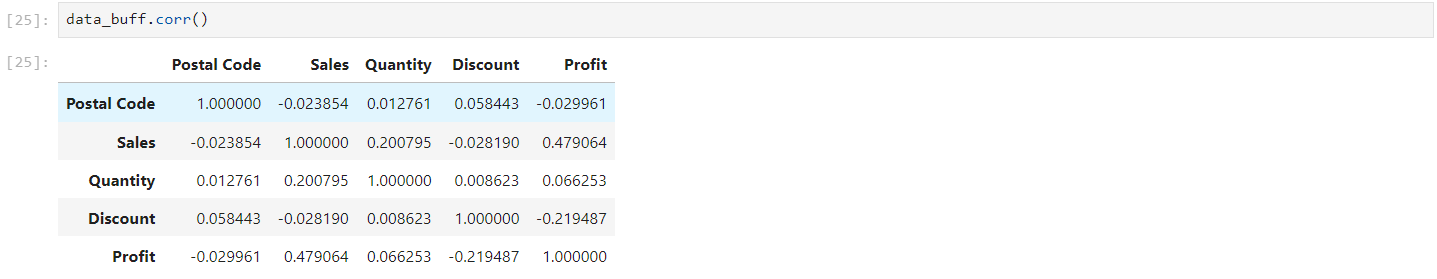
Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком. Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

Получим информацию и преобразуем типы.



Используем для вычисления корреляционной матрицы для числовых столбцов в DataFrame data\_buff. Каждый элемент этой матрицы представляет собой коэффициент корреляции между соответствующими парами столбцов.



На основе корреляционной матрицы можно сделать следующие выводы:

1. Postal Code и другие переменные:

* Postal Code не имеет сильной линейной корреляции с Sales, Quantity, Discount и Profit. Коэффициент корреляции близок к нулю в каждом случае.

1. Sales и другие переменные:

* Sales имеет положительную корреляцию с Profit (коэффициент 0.479), что может указывать на то, что увеличение продаж связано с увеличением прибыли. Это может быть ожидаемым результатом.
* Sales также имеет слабую положительную корреляцию с Quantity (коэффициент 0.200), что может означать, что более высокие объемы продаж могут соответствовать большему количеству проданных товаров.

1. Quantity и другие переменные:

* Quantity имеет слабую положительную корреляцию с Sales (коэффициент 0.200), что подтверждает наблюдение о том, что более высокие объемы продаж могут соответствовать большему количеству проданных товаров.

1. Discount и другие переменные:

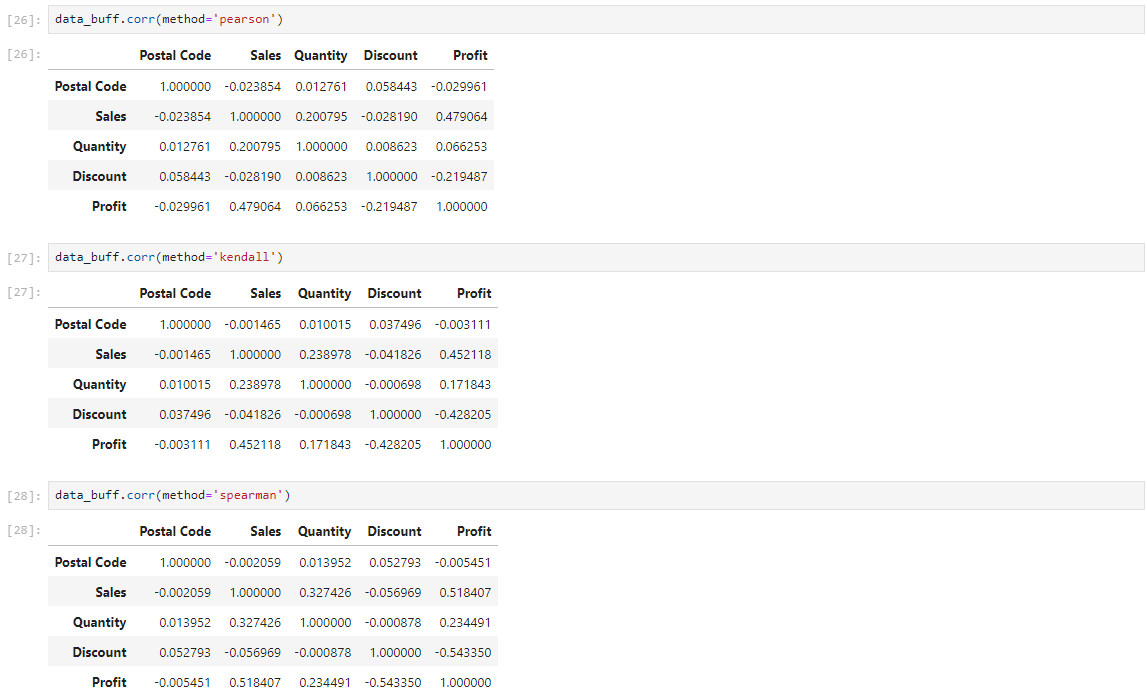
* Discount имеет небольшую отрицательную корреляцию с Profit (коэффициент -0.219), что может свидетельствовать о том, что предоставление скидок может быть связано с уменьшением прибыли.

1. Profit и другие переменные:

* Profit имеет положительную корреляцию с Sales (коэффициент 0.479), что было отмечено выше.
* Profit также имеет слабую положительную корреляцию с Quantity (коэффициент 0.066), что может указывать на то, что более высокая прибыль может быть связана с большим количеством проданных товаров.

Общий вывод: Некоторые переменные, такие как Sales и Quantity, могут иметь положительную корреляцию с прибылью, в то время как другие, такие как Discount, могут иметь отрицательное влияние на прибыль. Эти корреляции могут помочь в формулировании гипотез и дальнейшем анализе данных. Однако, корреляция не всегда означает причинно-следственную связь, и дополнительные исследования могут потребоваться для полного понимания взаимосвязей.

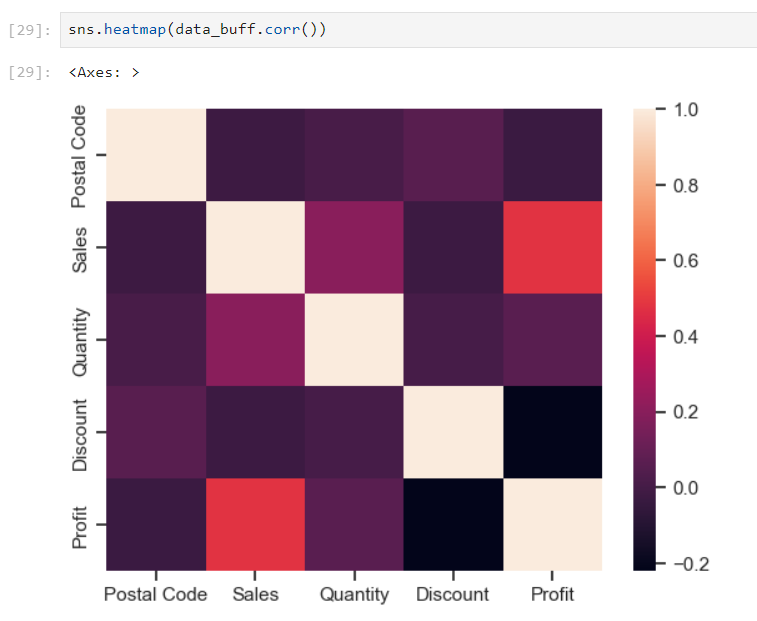
По умолчанию при построении матрицы используется коэффициент корреляции Пирсона. Возможно также построить корреляционную матрицу на основе коэффициентов корреляции Кендалла и Спирмена. На практике три метода редко дают значимые различия.



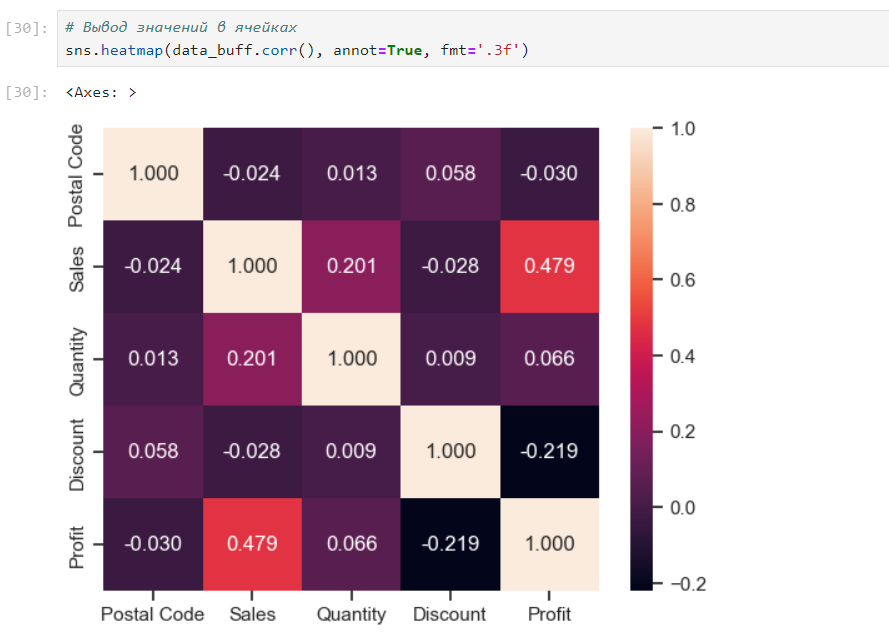
В случае большого количества признаков анализ числовой корреляционной матрицы становится неудобен.

Для визуализации корреляционной матрицы будем использовать "тепловую карту" heatmap которая показывает степень корреляции различными цветами.

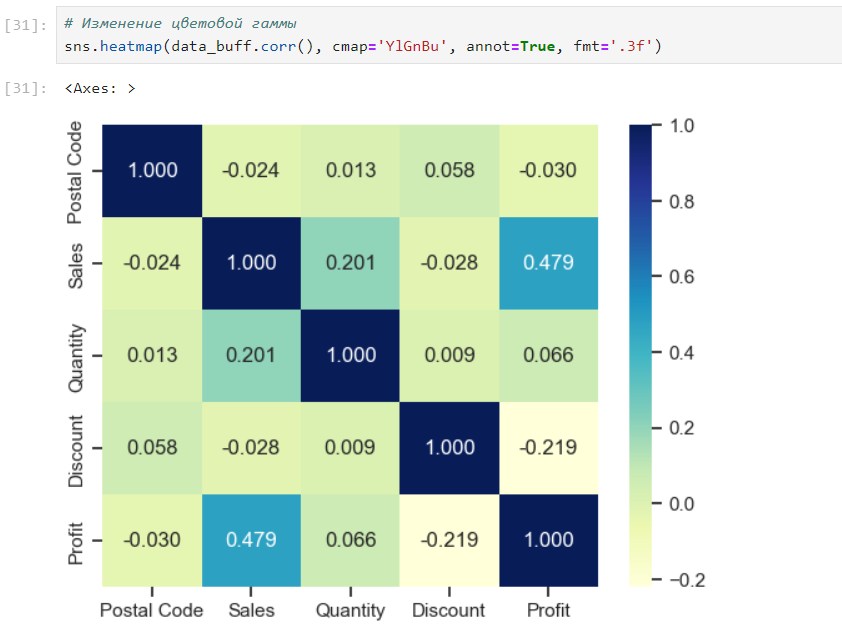
Используем метод heatmap библиотеки seaborn.



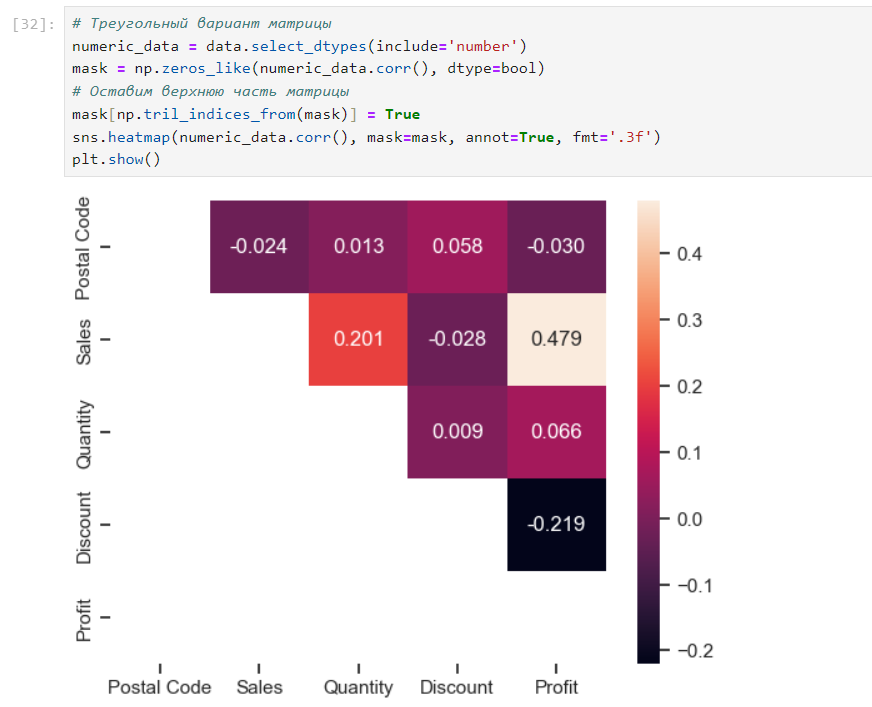
Вывод значений в ячейках



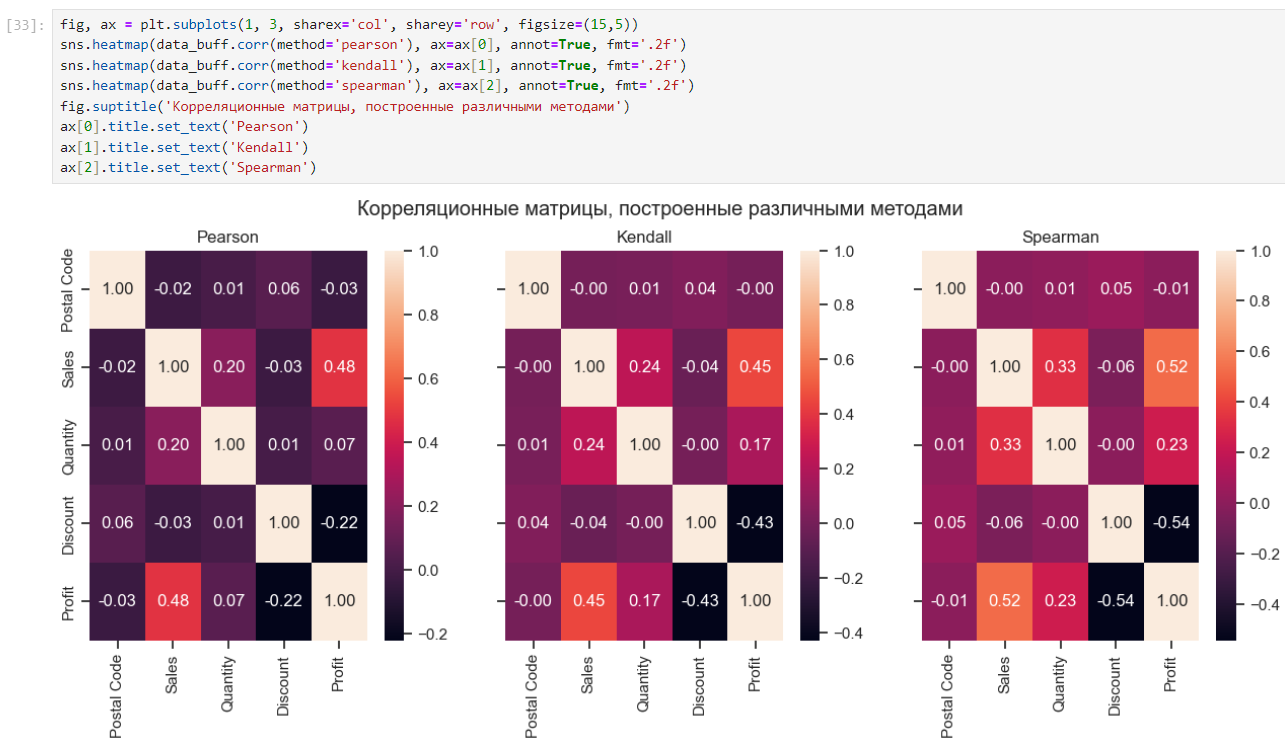
Изменение цветовой гаммы



Треугольный вариант матрицы



Корреляционные матрицы, построенные различными методами



Необходимо отметить, что тепловая карта не очень хорошо подходит для определения корреляции нецелевых признаков между собой.

В примере тепловая карта помогает определить значимую корреляцию между признаками Sales и Quantity, следовательно только один из этих признаков можно включать в модель.

Но в реальной модели могут быть сотни признаков и коррелирующие признаки могут образовывать группы, состояшие более чем из двух признаков. Увидеть такие группы с помощью тепловой карты сложно.

## Вывод

В ходе выполнения лабораторной работы изучили различные методы визуализация данных.