|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

по дисциплине «Технологии и инструментарий анализа больших данных»

**Практическая работа № 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Выполнил студент группы ИНБО-01-17 | ИВБО-06-20, Чурилов Антон Константинович | (подпись) | |
| Принял | Парамонов Александр Александрович | (подпись) | |
| Отчет представлен | «21» сентября 2023 г. | |  | |

Москва 2023 г.

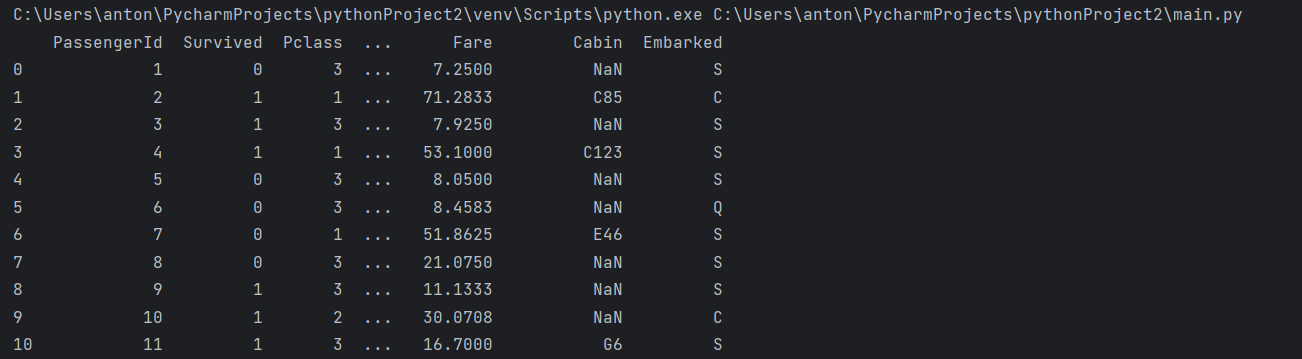
**Ход работы**

**Задание 1**

Найти и выгрузить многомерные данные (с большим количеством признаков – столбцов) с использованием библиотеки pandas. В отчёте описать найденные данные.

****

**Рисунок 1 – Выгрузка датафрейма из csv**

****

**Рисунок 2 – вывод программы**

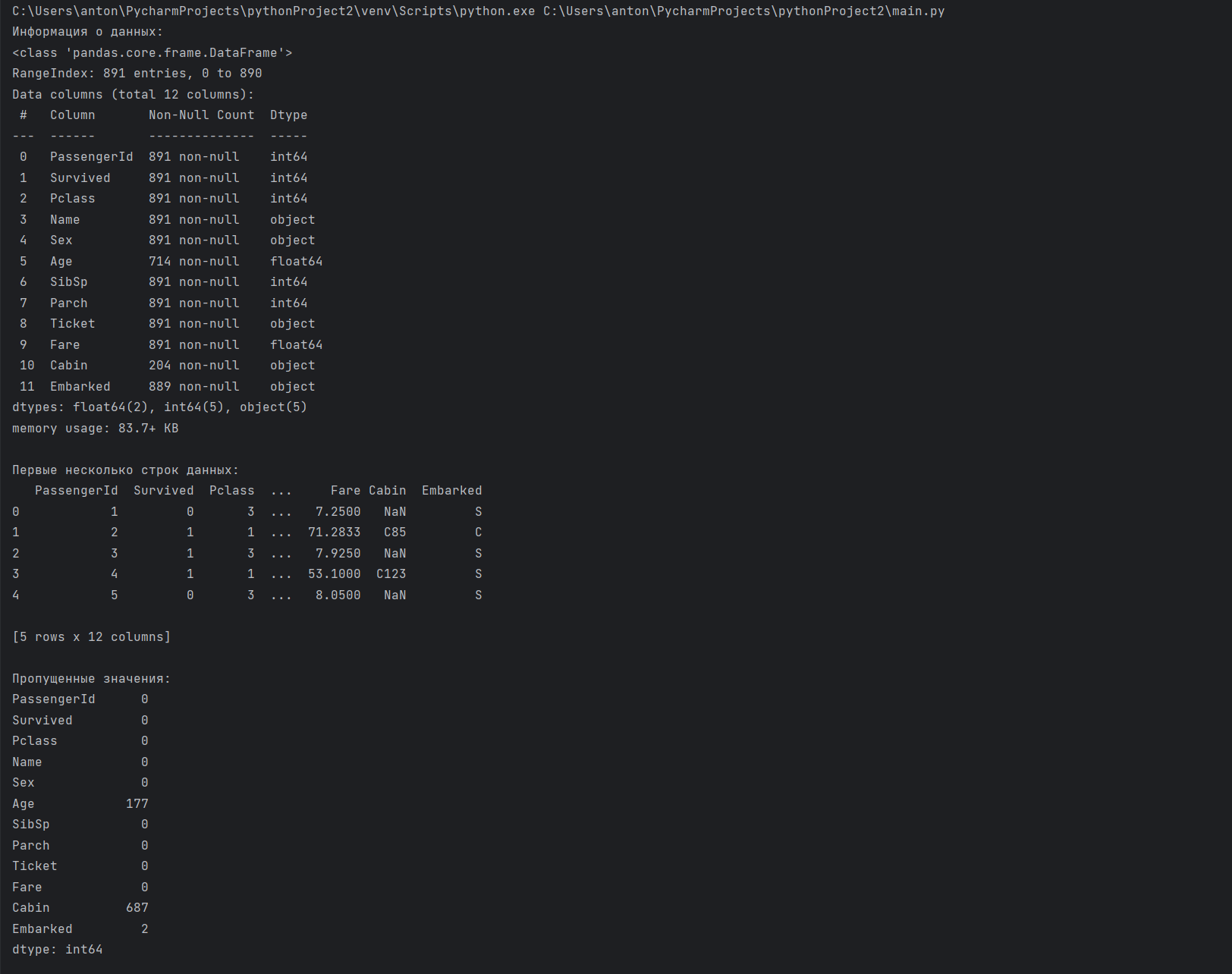
Выгруженный набор данных содержит информацию о пассажирах Титаника их класс билета, номер каюты, возраст, количество братьев и сестер\сестер на борту, количество родителей/детей на борту, номер билета, пассажирский тариф и номер каюты.

**Задание 2**

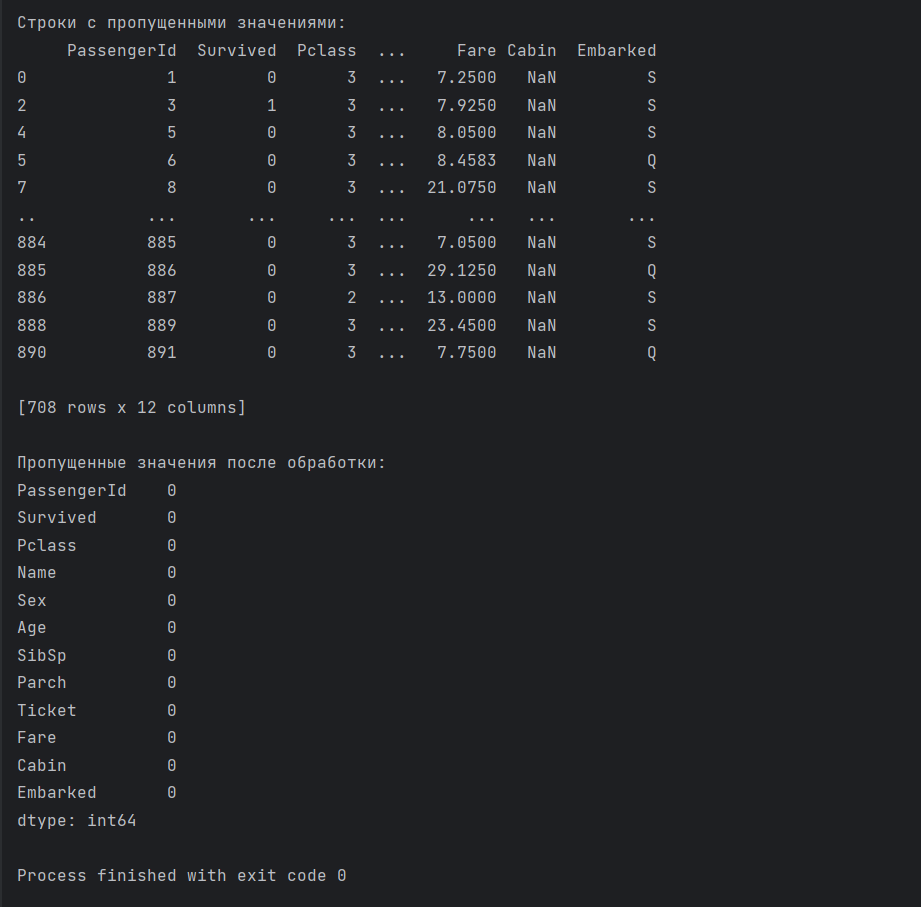
Вывести информацию о данных при помощи методов .info(), .head(). Проверить данные на наличие пустых значений. В случае их наличия удалить данные строки или интерполировать пропущенные значения. При необходимости дополнительно проработать данные для дальнейшей работы с ними.

****

**Рисунок 3 – код программы**

****

**Рисунок 4 – вывод info() и head()**

****

**Рисунок 5 – вывод info() и head()**

**Задание 3**

3. Построить столбчатую диаграмму (.bar) с использованием модуля graph\_objs из библиотеки Plotly со следующими параметрами:

3.1. По оси Х указать дату или название, по оси У указать количественный показатель.

3.2. Сделать так, чтобы столбец принимал цвет в зависимости от значения показателя (marker=dict(color=признак, coloraxis="coloraxis")).

3.3. Сделать так, чтобы границы каждого столбца были выделены чёрной линией с толщиной равной 2.

3.4. Отобразить заголовок диаграммы, разместив его по центру сверху, с 20 размером текста.

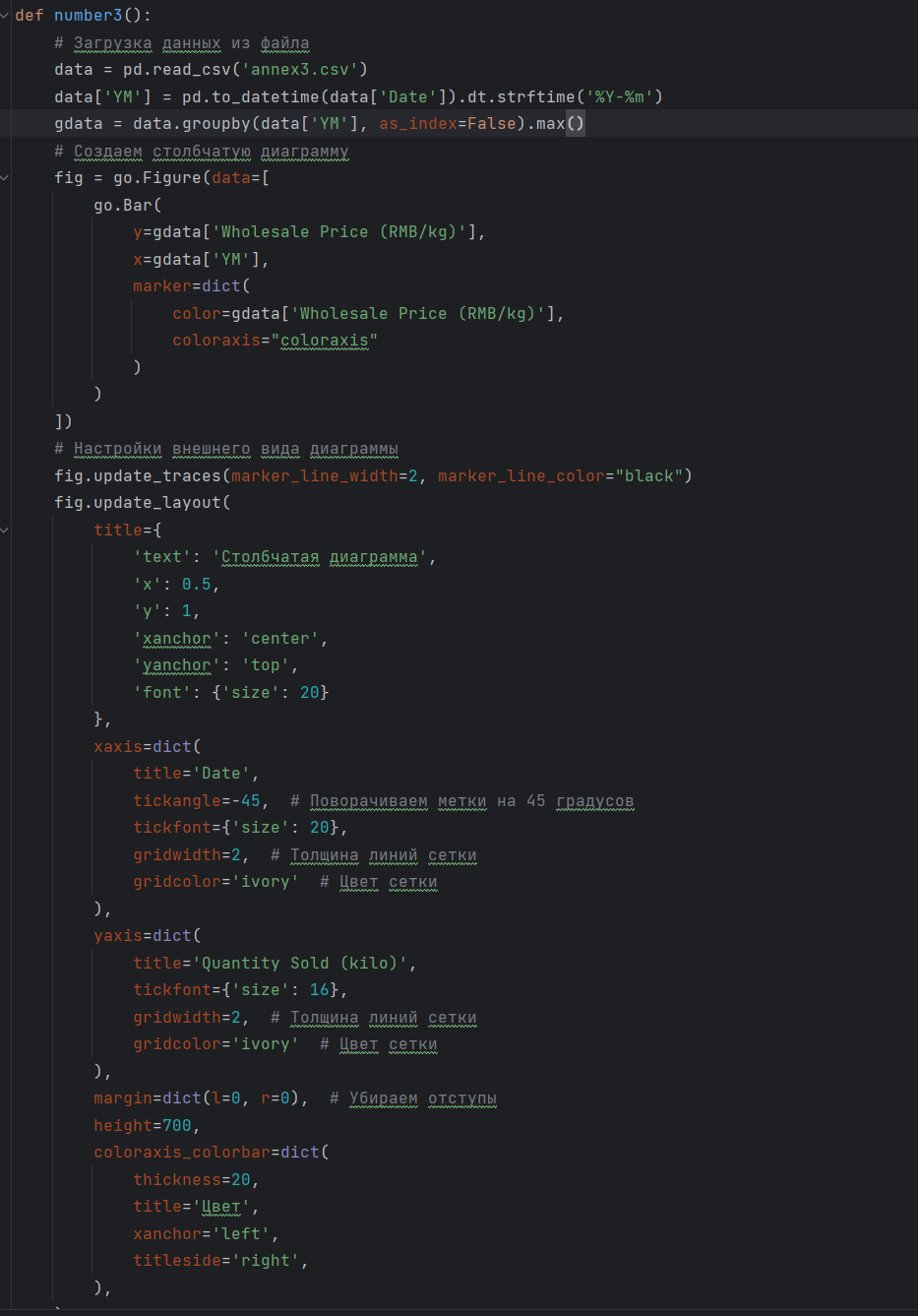
3.5. Добавить подписи для осей X и Y с размером текста, равным 16. Для оси абсцисс развернуть метки так, чтобы они читались под углом, равным 315.

3.6. Размер текста меток осей сделать равным 14.

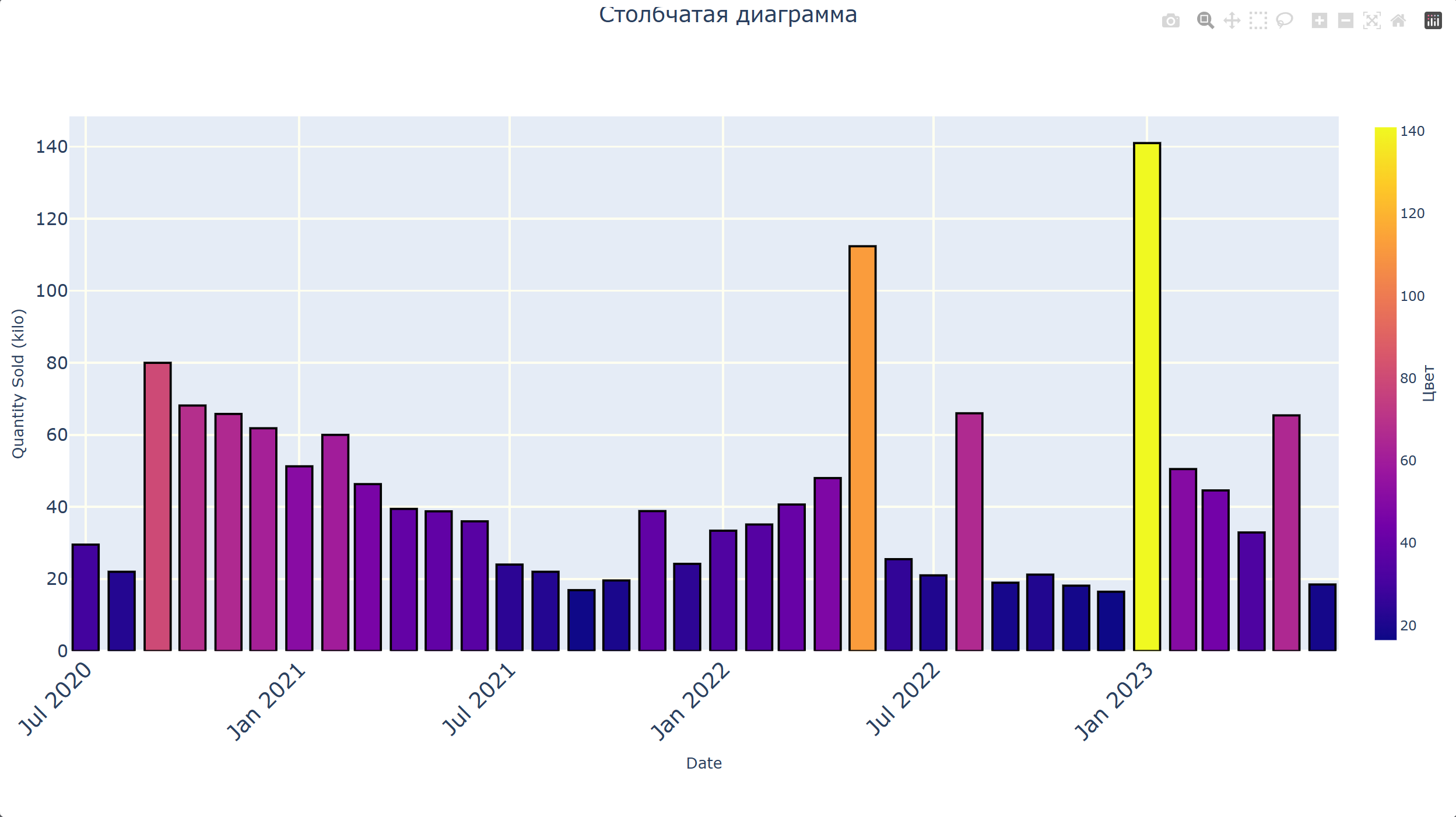
3.7. Расположить график во всю ширину рабочей области и присвоить высоту, равную 700 пикселей.

3.8. Добавить сетку на график, сделать её цвет 'ivory' и толщину равную 2. (Можно сделать это при настройке осей с помощью gridwidth=2, gridcolor='ivory')

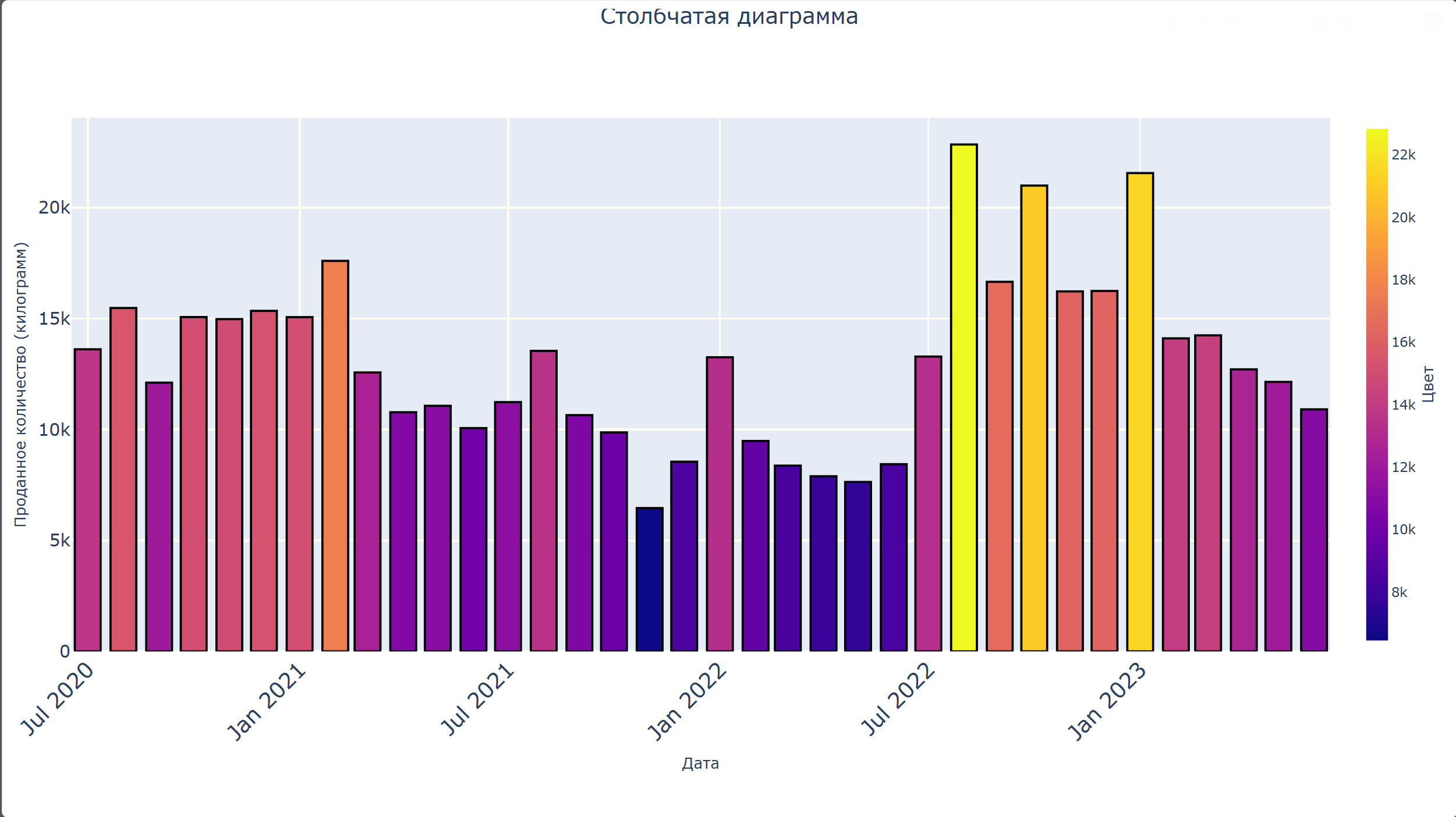
3.9. Убрать лишние отступы по краям.



**Рисунок 6 – Код задания 3**

****

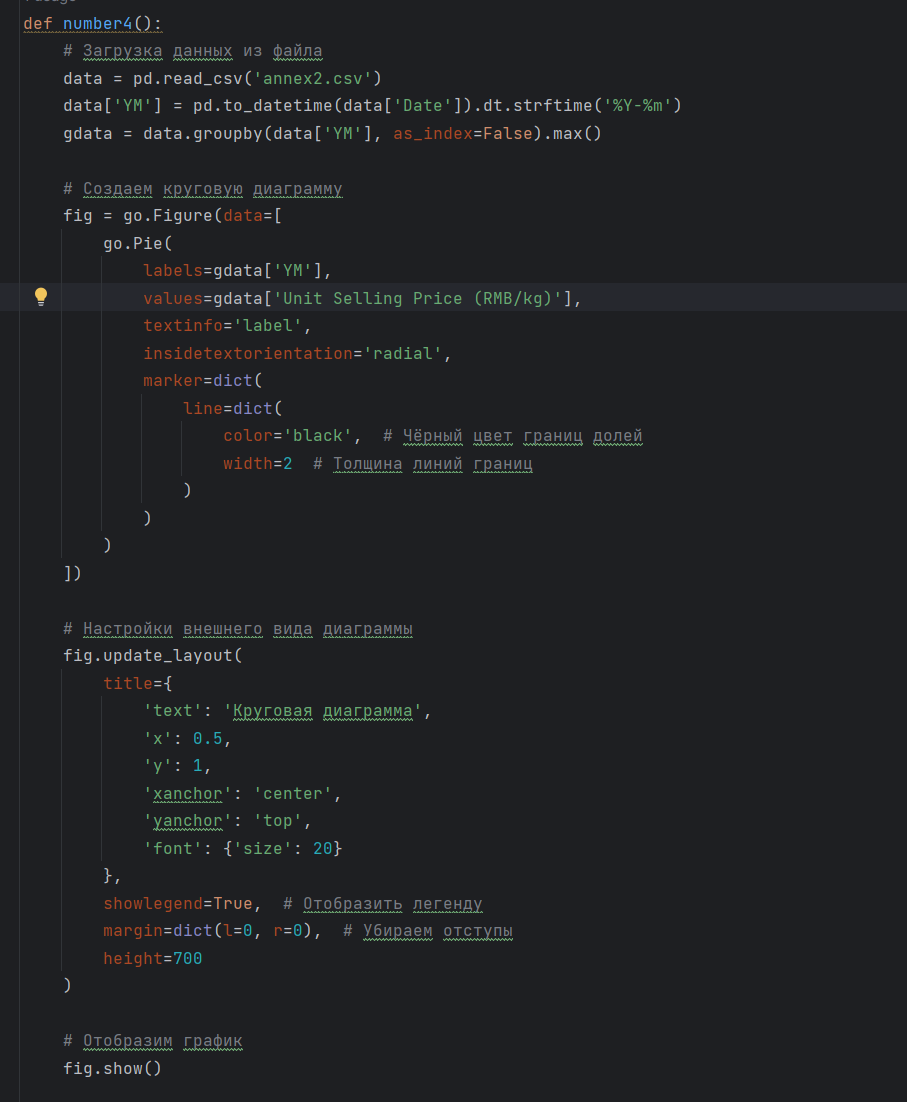
**Рисунок 7.1 – Столбчатая диаграмма библиотеки Plotly(год/месяц –макс. Оптовая цена)**

****

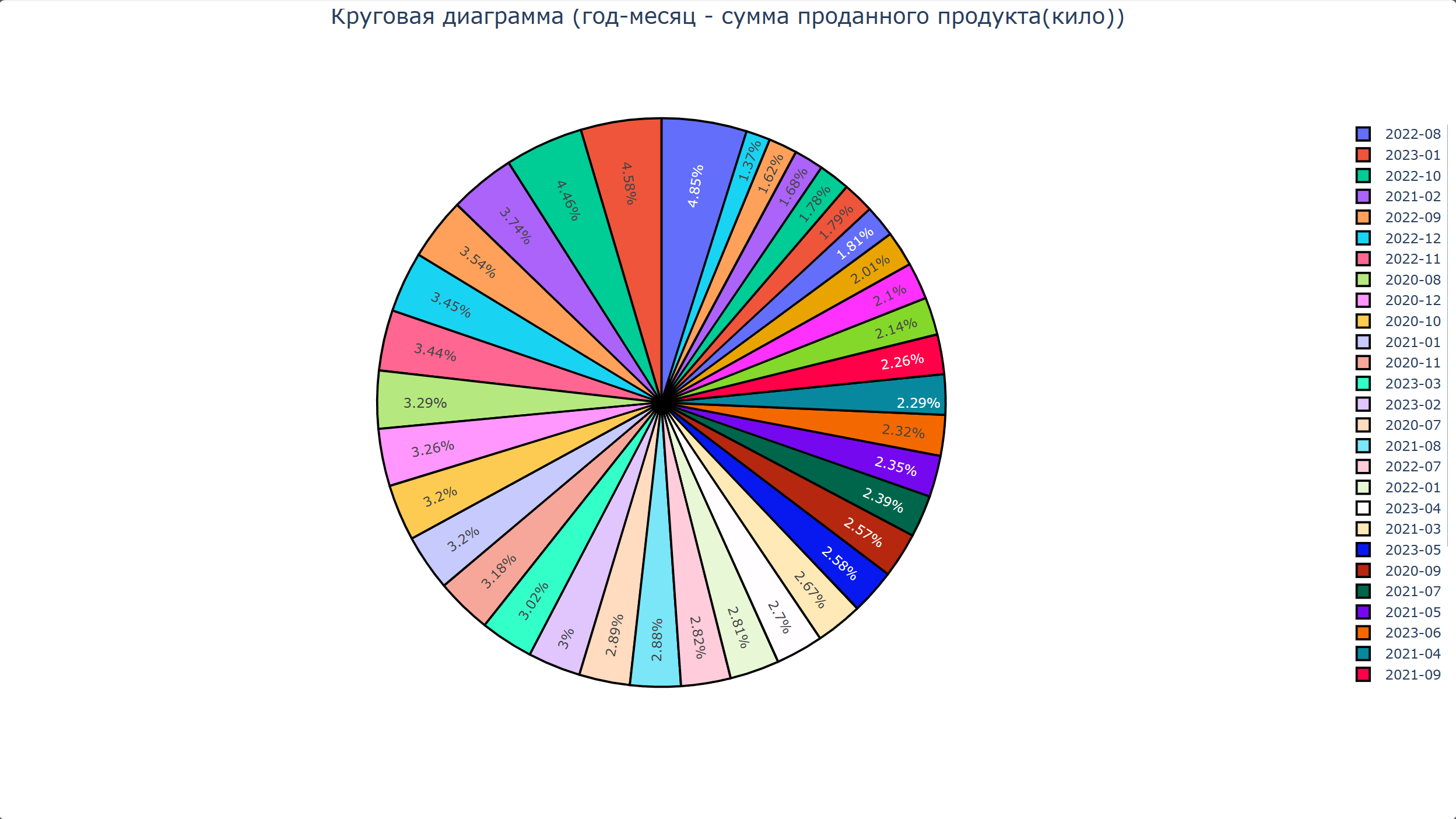
**Рисунок 7.2 – Столбчатая диаграмма библиотеки Plotly(год/месяц – сумма проданного товара(кило))**

**Задание 4**

Построить круговую диаграмму (go.Pie), использовав данные и стиль оформления из предыдущего графика. Сделать так, чтобы границы каждой доли были выделены чёрной линией с толщиной, равной 2 и категории круговой диаграммы были читаемы (к примеру, объединить часть объектов)



**Рисунок 8 – Код задания 4**



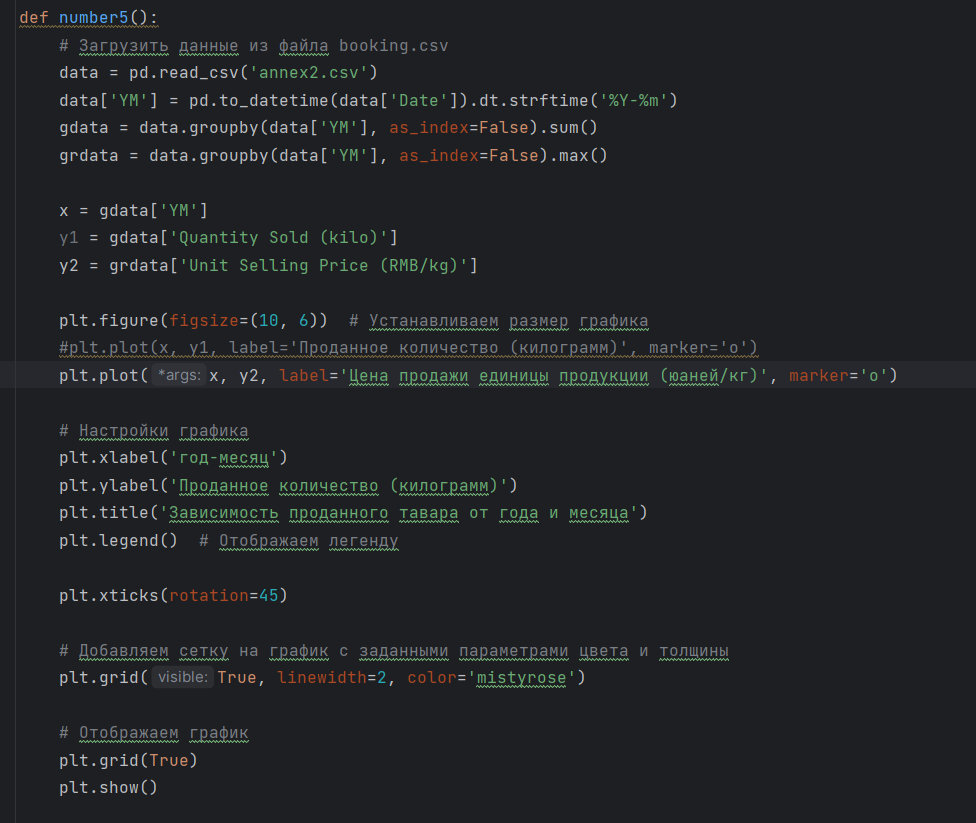
**Рисунок 9 – Круговая диаграмма библиотеки Plotly**

**Задание 5**

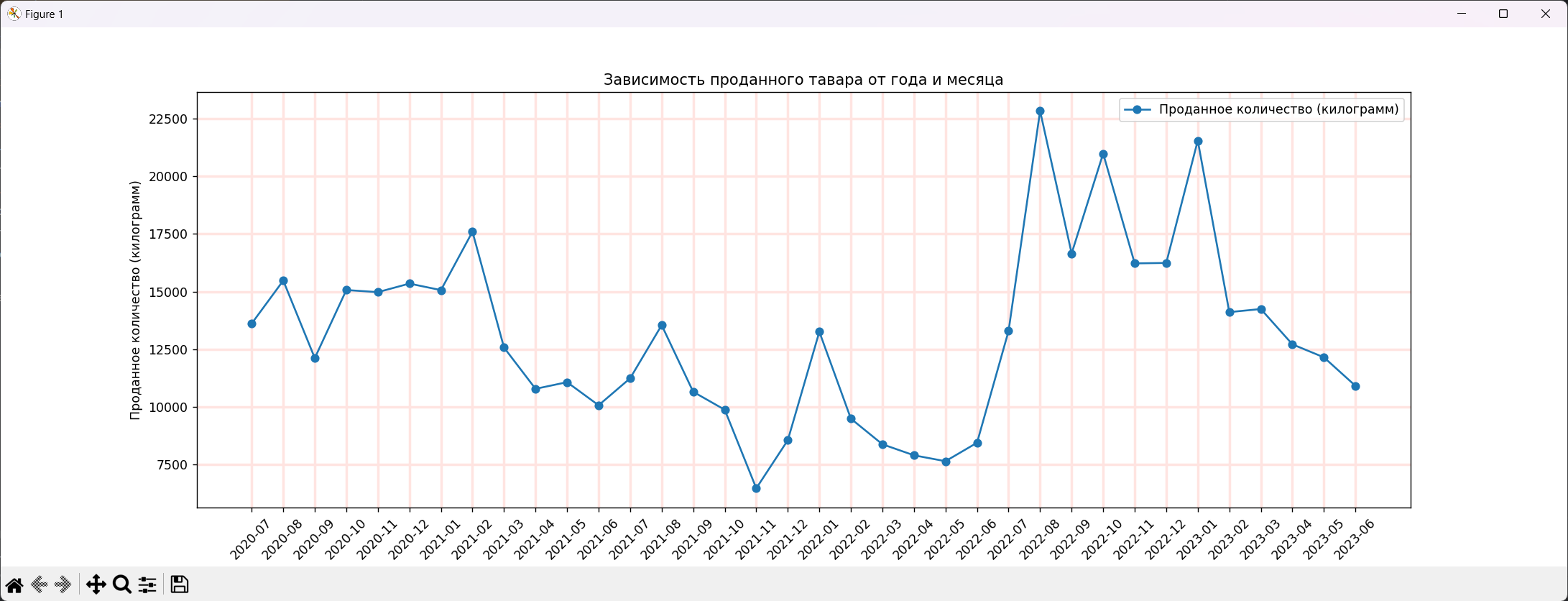
Построить линейные графики, взять один из параметров и определить зависимость между другими несколькими (от 2 до 5) показателями с использованием библиотеки matplotlib. Сделать вывод.

Сделать график с линиями и маркерами, цвет линии 'crimson', цвет точек 'white', цвет границ точек 'black', толщина границ точек равна 2.

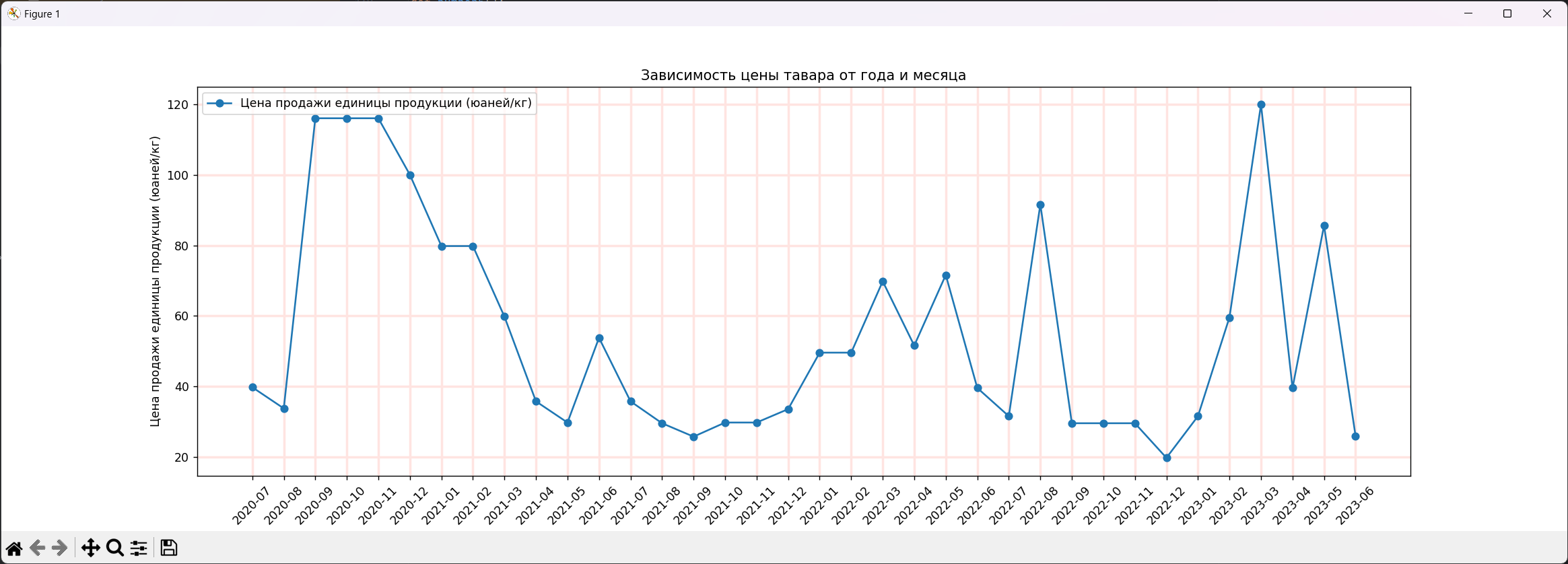
Добавить сетку на график, сделать её цвет 'mistyrose' и толщину равную 2. (Можно сделать это при настройке осей с помощью linewidth=2, color='mistyrose').



**Рисунок 10 – Код задания 5**



**Рисунок 11.1 – Линейный график библиотеки Plotly**

****

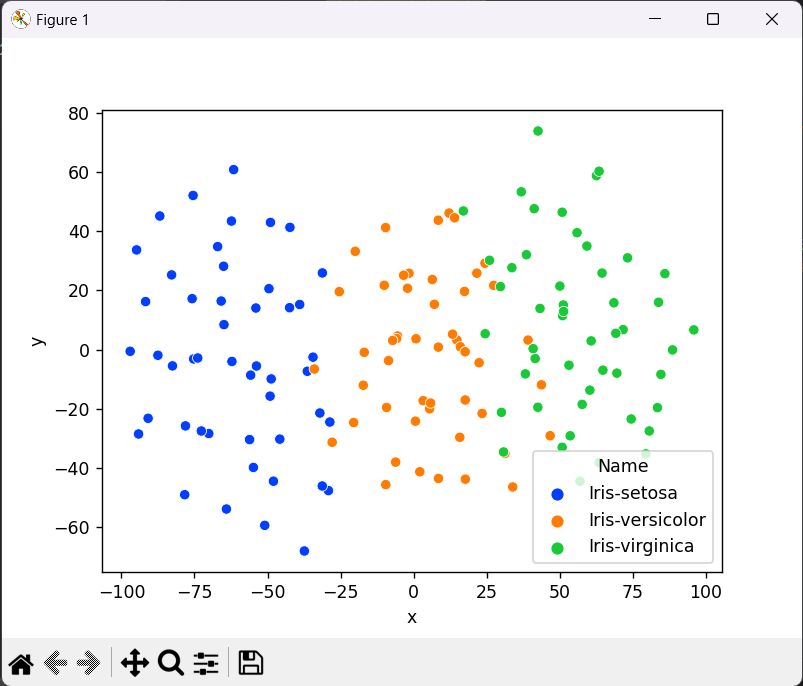
**Рисунок 11.2 – Линейный график библиотеки Plotly**

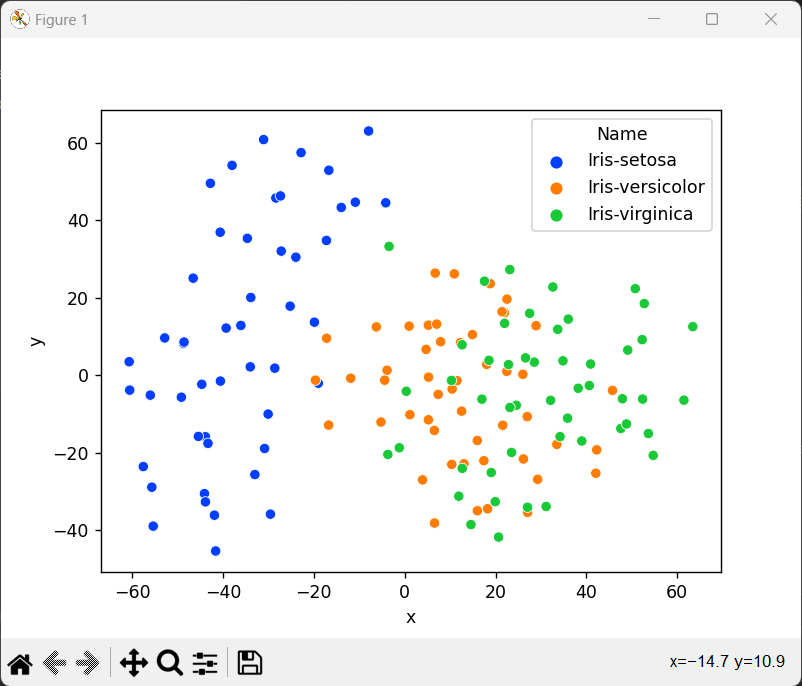
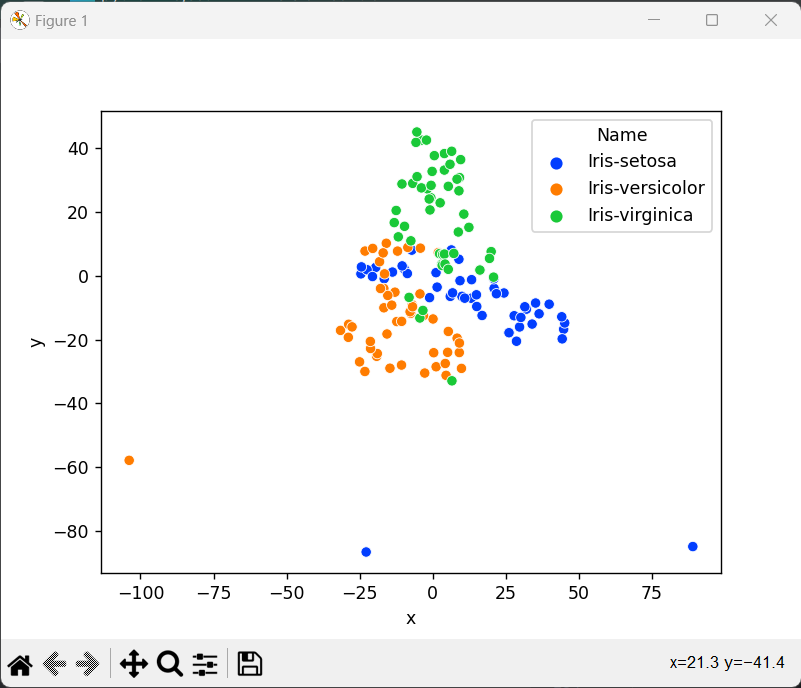
**Задание 6**

Выполнить визуализацию многомерных данных, используя t-SNE. Необходимо использовать набор данных MNIST или fashion MNIST (можно использовать и другие готовые наборы данных, где можно наблюдать разделение объектов по кластерам). Рассмотреть результаты визуализации для разных значений перплексии.



**Рисунок 12 – Код задания 6**

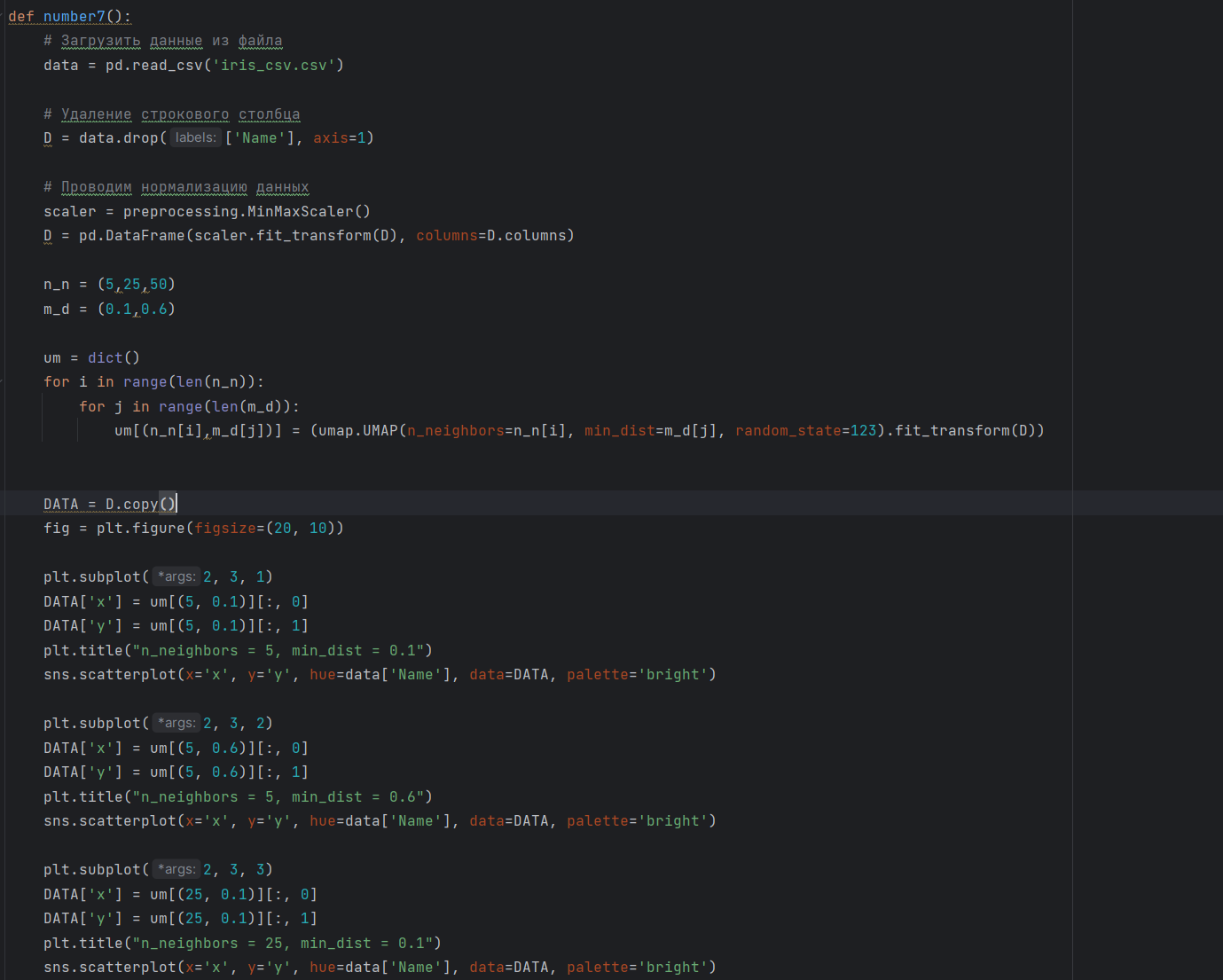
**** perplexity = 50

**** perplexity = 25**** perplexity = 5

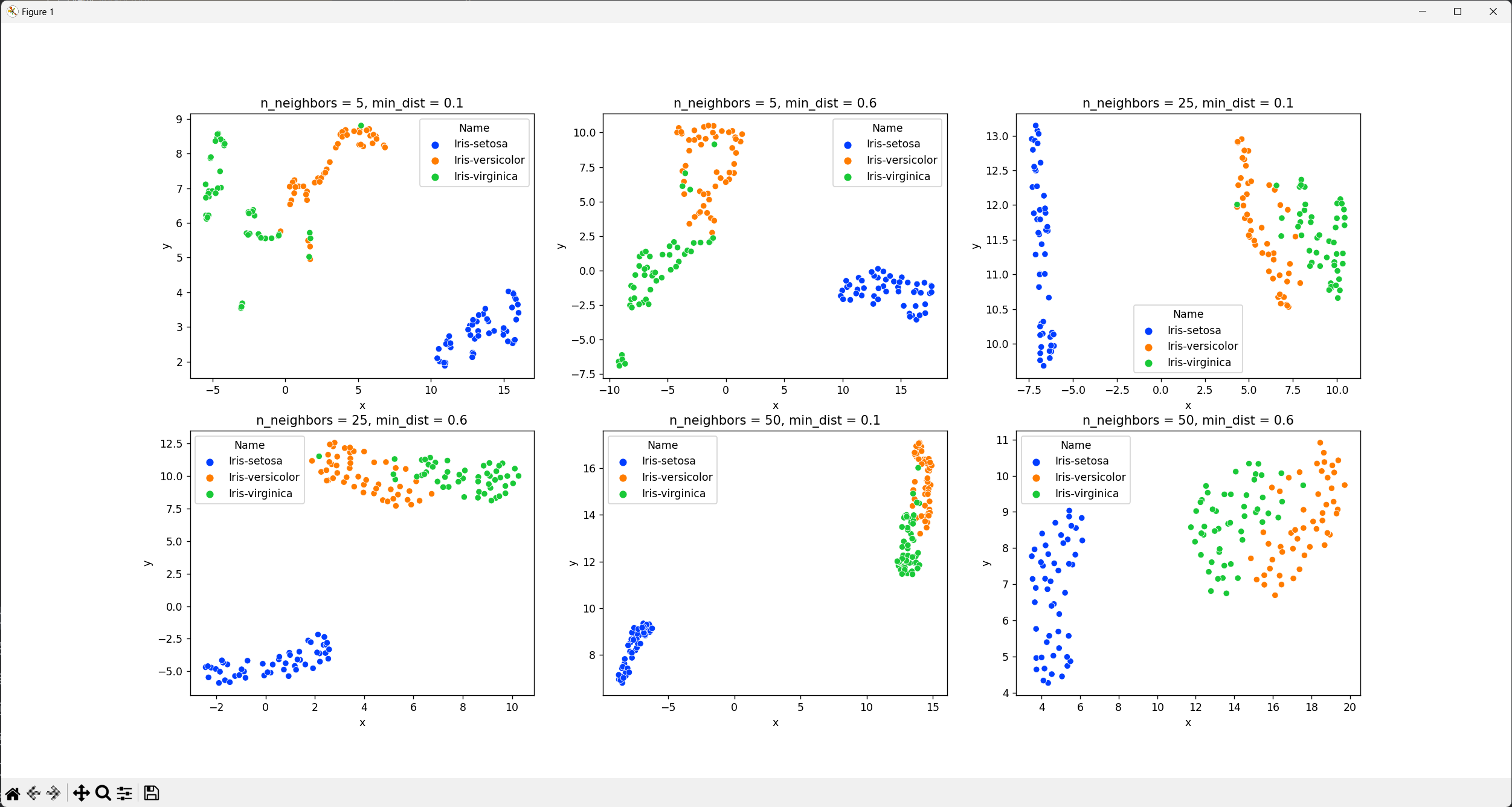
**Рисунок 13 – распределение TSNE**

**Задание 7**

Выполнить визуализацию многомерных данных, используя UMAP с различными параметрами n\_neighbors и min\_dist. Рассчитать время работы алгоритма с помощью библиотеки time и сравнить его с временем работы t-SNE.



**Рисунок 11 – Код задания 7**



**Рисунок 12 – виртуализация данных с помощью библиотеки UMAP**

**Вывод**

В ходе выполнения данной практической работы мы познакомились c библиотеками для визуализации данных, а именно – Plotly, Matplotlib, t-SNE и UMAP. Помимо этого, мы изучили их функционал и применили полученные знания для построения графиков по загруженному датафрейму.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ**

1. Using Python on Windows [Электронный ресурс] – Режим доступа: <https://docs.python.org/3/using/windows.html>, свободный. Языка – английский. Дата обращения: 18.09.2023.
2. Jupyter Notebooks in VS Code [Электронный ресурс] – Режим доступа: <https://code.visualstudio.com/docs/datascience/jupyter-notebooks>, свободный. Языка – английский. Дата обращения: 18.09.2023.
3. Plotly 5.9.0 Documentation [Электронный ресурс] – Режим доступа: <https://plotly.com/python-api-reference/>, свободный. Язык – английский. Дата обращения: 18.09.2023.
4. Matplotlib 3.6.0 Documentation [Электронный ресурс] – Режим доступа: <https://matplotlib.org/stable/index.html>, свободный. Язык – английский. Дата обращения: 18.09.2023.