Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5И-21М
Ли Лююй

Москва — 2024 г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

Требуется выполнить следующие действия [1]:

• Выбрать набор данных (датасет).

• Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

• Сформировать отчет и разместить его в своем репозитории на GitHub

3. Ход выполнения работы

3.1. Текстовое описание набора данных

В этой тетради я буду использовать графики для визуализации взаимосвязи между переменными в наборе данных "ноутбуки".

- **The dataset includes the following columns:**

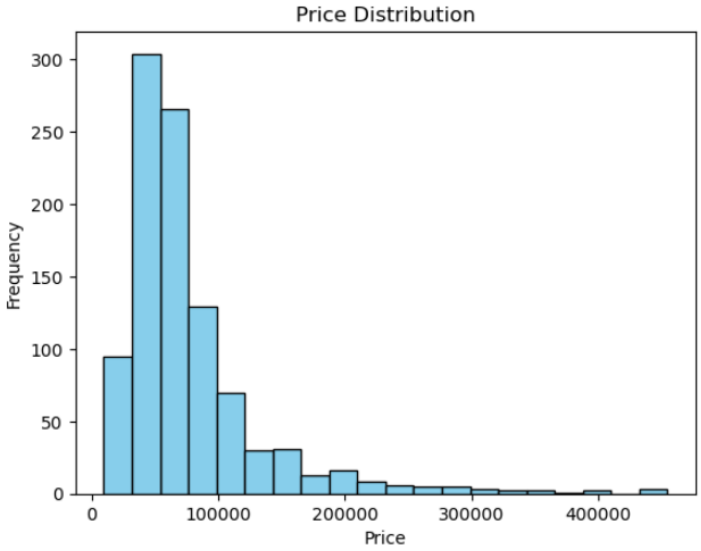
• index	• Индекс
• brand	• Марка
• Model	• Модель
• Price	• Цена,
• Rating	• Рейтинг
• processor	• Марка процессора
• brand	• Уровень процессора
• num_cores	• Количество ядер

• num_threads	• Количество потоков
• ram_memory	• Оперативная память
• primary_storage_type	• Тип основного накопителя
• primary_storage_capacity	• Емкость основного накопителя
• secondary_storage_type	• Тип вторичного накопителя
• secondary_storage_capacity	• Емкость вторичного накопителя
• gpu_brand	• Марка GPU
• gpu_type	• Тип GPU
• is_touch_screen	• Сенсорный экран
• display_size	• ,Размер дисплея
• resolution_width	• Ширина разрешения
• resolution_height	• Высота разрешения
• OS	• ОС
• year_of_warranty	• Год гарантии

•

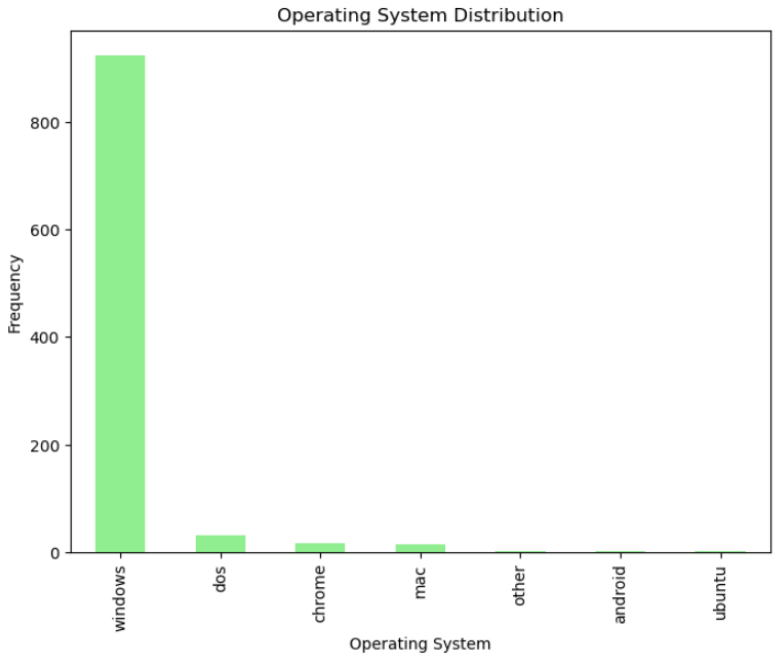
3.2. Основные характеристики набора данных

In [6]:
```
# Price distribution
plt.hist(data['Price'], bins=20, color='skyblue', edgecolor='black')
plt.title('Price Distribution')
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.show()
```



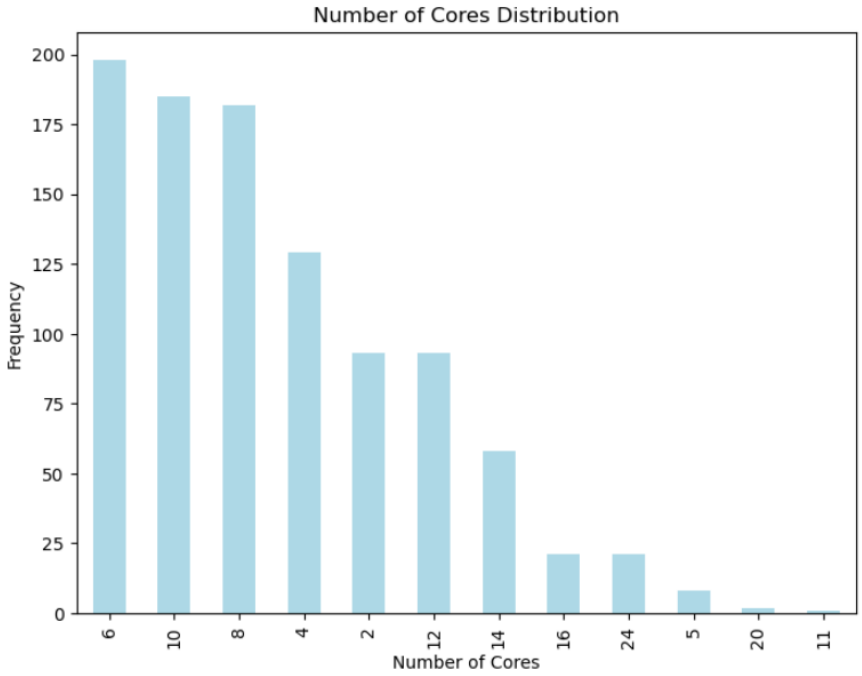
In [7]:
```
# Relationship between Price and Rating
sns.scatterplot(x='Price', y='Rating', data=data)
plt.title('Relationship between Price and Rating')
plt.show()
```

In [8]:
```
# Operating System Distribution
plt.figure(figsize=(8, 6))
data['OS'].value_counts().plot(kind='bar', color='lightgreen')
plt.title('Operating System Distribution')
plt.xlabel('Operating System')
plt.ylabel('Frequency')
plt.show()
```




```
# Group by year and calculate mean prices
mean_prices = data.groupby('year_of_warranty')['Price'].mean().reset_index()

# Relationship between Year and Mean Price
plt.plot(mean_prices['year_of_warranty'], mean_prices['Price'], marker='o', color='orange')
plt.title('Relationship between Year and Mean Price')
plt.xlabel('Year')
plt.ylabel('Mean Price')
plt.grid(True)
plt.show()
```



Conclusion:
Based on the analysis and visualizations of the dataset, we can conclude the following:
- The price distribution shows that most prices are concentrated towards the lower end.
- There is a positive relationship between price and rating, indicating that more expensive computers tend to have higher ratings.
- Most computers in the dataset use the Windows operating system.
- The relationship between the year of warranty and mean price suggests a decreasing trend in average prices over time.
- The distribution of the number of cores among computers varies, with some having higher numbers of cores.
These findings provide valuable insights for understanding the characteristics of computers in the dataset.

---

1. Most computer prices are concentrated towards the lower end, which is common in the computer market.
2. There is a positive correlation between computer price and rating, but there are also some outliers.
3. The Windows operating system predominates among our computer samples.
4. The average price of computers tends to decrease over time.
5. Analyzing these data can help in making decisions regarding pricing, product assortment, and marketing strategies.

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets