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

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

«Создание "истории о данных".»

Выполнил:
студент группы ИУ5-22М
Лун Сыхань

Москва — 2024г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

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 "Фильмы Marvel".

The dataset includes the following columns:

- Movie Title
- Release Date (USA)
- Phase
- Genre
- Movie Rating
- Lead Role
- Runtime (min)

Production Budget

Box Office (Local)

Box Office (International)

Total Box Office Earnings

Rotten Tomatoes Ratings (%)

IMDb Ratings

3.2. Основные характеристики набора данных

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
from IPython.display import display
```

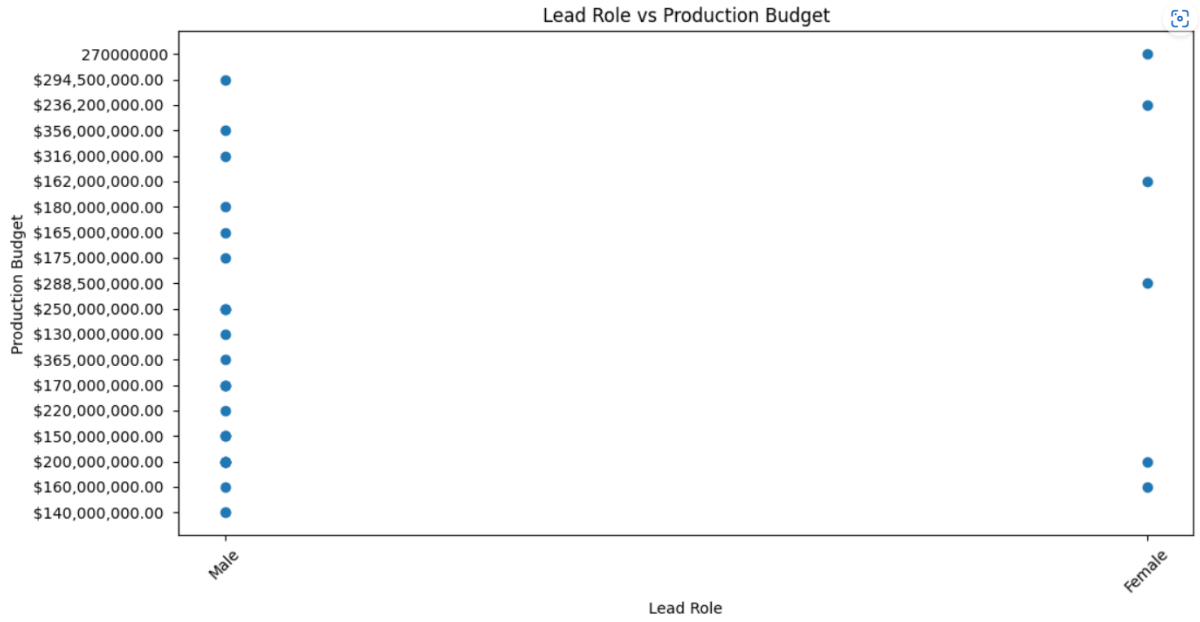
[2]:
```
# 读取CSV文件
df = pd.read_csv('MCU Movies.csv')
```

[3]:
```
# 查看数据
display(df.head())
```

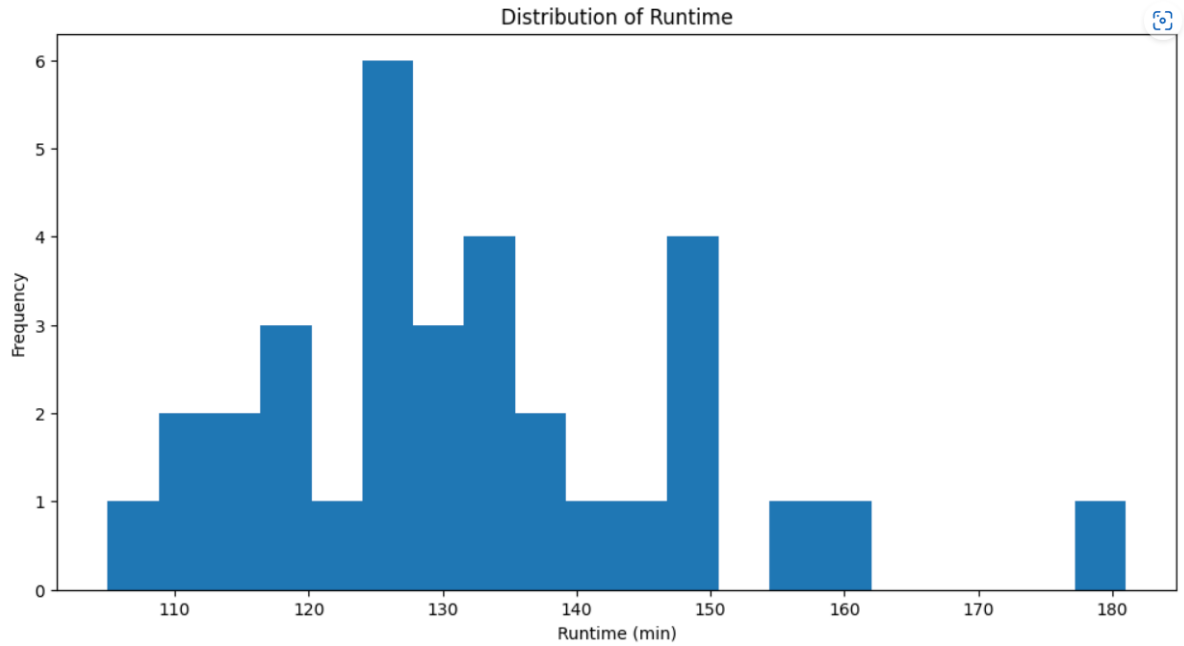
	Movie Title	Release Date (USA)	Phase	Genre	Movie Rating	Lead Role	Runtime (min)	Production Budget	Box Office (Local)	Box Office (International)	Total Box Office Earnings	Rotten Tomatoes Ratings (%)	IMDb Ratings
0	Captain America: The First Avenger	22-Jul-2011	1	Action | Adventure | Sci-Fi	PG-13	Male	124	$140,000,000.00	$176,654,505.00	$193,915,269.00	$370,569,774.00	80	6.9
1	Captain Marvel	8-Mar-2019	3	Action | Adventure | Sci-Fi	PG-13	Female	125	$160,000,000.00	$426,829,839.00	$704,586,607.00	$1,131,416,446.00	79	6.8
2	Iron Man	2-May-2008	1	Action | Adventure | Sci-Fi	PG-13	Male	126	$140,000,000.00	$319,034,126.00	$266,762,121.00	$585,796,247.00	94	7.9
3	Iron Man 2	7-May-2010	1	Action | Sci-Fi	PG-13	Male	124	$200,000,000.00	$312,433,331.00	$311,500,000.00	$623,933,331.00	72	6.9
4	The Incredible Hulk	13-Jun-2008	1	Action | Adventure | Sci-Fi	PG-13	Male	112	$150,000,000.00	$134,806,913.00	$129,964,083.00	$264,770,996.00	67	6.6

[9]:
```
# 数据清洗：移除美元符号和逗号，并转换为数值
df['Box Office (Local)'] = df['Box Office (Local)'].replace('[\$,]', '', regex=True).astype(float)
```

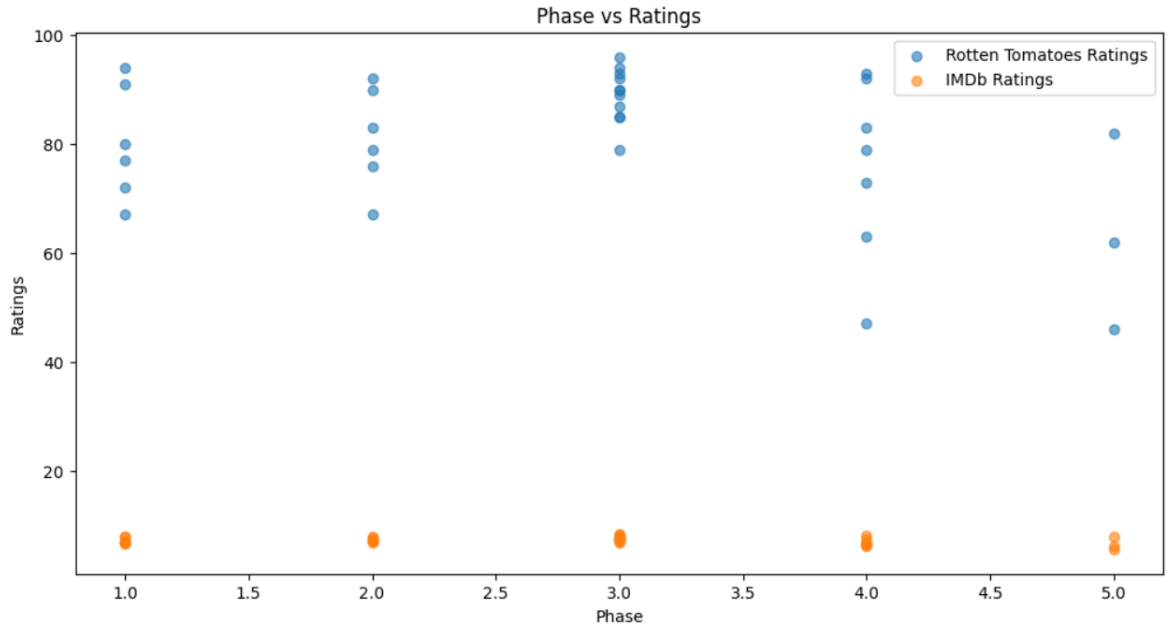
```python
# 1. 分析Lead Role与Production Budget的关系
plt.figure(figsize=(12, 6))
plt.scatter(df['Lead Role'], df['Production Budget'])
plt.title('Lead Role vs Production Budget')
plt.xticks(rotation=45)
plt.xlabel('Lead Role')
plt.ylabel('Production Budget')
plt.show()
```



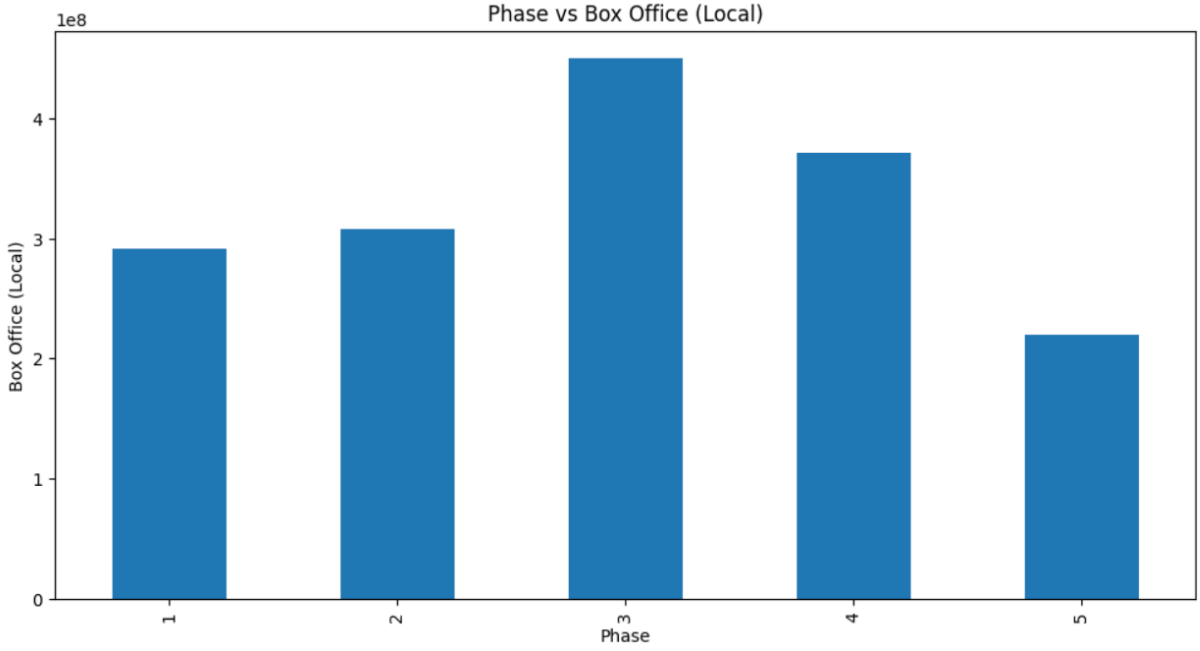

```python
# 2. 分析Runtime (min)的分布
plt.figure(figsize=(12, 6))
plt.hist(df['Runtime (min)'], bins=20)
plt.title('Distribution of Runtime')
plt.xlabel('Runtime (min)')
plt.ylabel('Frequency')
plt.show()
```



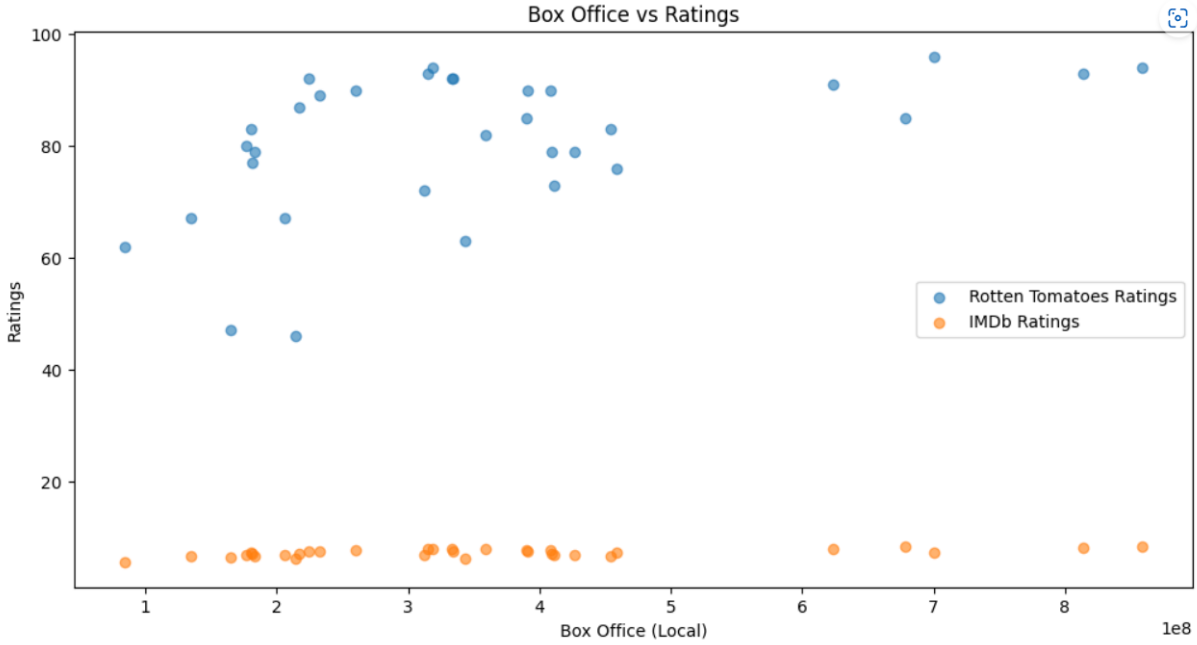
```python
# 3. 分析Phase与Rotten Tomatoes Ratings (%)以及IMDb Ratings的关系
plt.figure(figsize=(12, 6))
plt.scatter(df['Phase'], df['Rotten Tomatoes Ratings (%)'], label='Rotten Tomatoes Ratings', alpha=0.6)
plt.scatter(df['Phase'], df['IMDb Ratings'], label='IMDb Ratings', alpha=0.6)
plt.title('Phase vs Ratings')
plt.xlabel('Phase')
plt.ylabel('Ratings')
plt.legend()
plt.show()
```




```python
# 4. 分析Phase与Box Office的关系
plt.figure(figsize=(12, 6))
df.groupby('Phase')['Box Office (Local)'].mean().plot(kind='bar')
plt.title('Phase vs Box Office (Local)')
plt.xlabel('Phase')
plt.ylabel('Box Office (Local)')
plt.show()
```



```python
# 5. 分析Box Office与Rotten Tomatoes Ratings (%)以及IMDb Ratings的关系
plt.figure(figsize=(12, 6))
plt.scatter(df['Box Office (Local)'], df['Rotten Tomatoes Ratings (%)'], label='Rotten Tomatoes Ratings', alpha=0.6)
plt.scatter(df['Box Office (Local)'], df['IMDb Ratings'], label='IMDb Ratings', alpha=0.6)
plt.title('Box Office vs Ratings')
plt.xlabel('Box Office (Local)')
plt.ylabel('Ratings')
plt.legend()
plt.show()
```



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