**数据处理与可视化**

使用kaggle上数据集：

[https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/data](https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/data，)

完成下列数据分析与可视化任务

1.数据加载与观察

2.数据清洗与处理

3. 数据转换与拆分

4. 数据聚合与统计

o如对不同类型的电影进行分组，计算每种类型电影的平均票房收入，并进行排序。

5.数据可视化

o如使用Matplotlib绘制图，展示不同类型电影的平均票房收入，以便比较各种类型电影的票房表现。

6.高级数据分析

o如使用Pandas进行更深入的数据分析，比如计算电影类型的票房收入中位数、标准差等统计指标，以更全面地了解不同类型电影的票房表现。

7.结果解释与总结

o根据可视化和统计结果，对不同类型电影的票房表现进行解释和总结，提出相关的结论和观点

数据分析可以大致参照上述流程，但是可以自行进行任意探索。

Python 3.12.2

PyCharm Community Edition 2023.1.4

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

df = pd.read\_csv('movies\_metadata.csv')

df = df.dropna(subset=['revenue'])

# 计算每种类型电影的平均票房收入

genre\_revenue = df.explode('genres')['genres'].value\_counts()

genre\_revenue = pd.DataFrame(genre\_revenue).reset\_index()

genre\_revenue.columns = ['genre', 'count']

# 计算每种类型电影的平均票房

genre\_avg\_revenue=df.explode('genres').groupby(['genres','genre'])['revenue'].mean().unstack().reset\_index()

# 数据可视化 - 直方图

plt.figure(figsize=(14, 6)) # 设置直方图的大小

plt.hist(genre\_avg\_revenue['revenue'], bins=50, edgecolor='black')

plt.xlabel('Revenue') # x轴标签

plt.ylabel('Number of Genres') # y轴标签

plt.title('Distribution of Revenue Across Genres') # 图表标题

plt.xticks(spacing=500000) # 设置x轴的刻度间距，这里以50万为间隔

plt.tight\_layout() # 自动调整子图参数, 使之填充整个图像区域

plt.show()

# 高级数据分析 - 计算票房收入的中位数和标准差

genre\_stats=df.explode('genres').groupby(['genres','genre'])['revenue'].agg(['median', 'std']).reset\_index()

# 结果

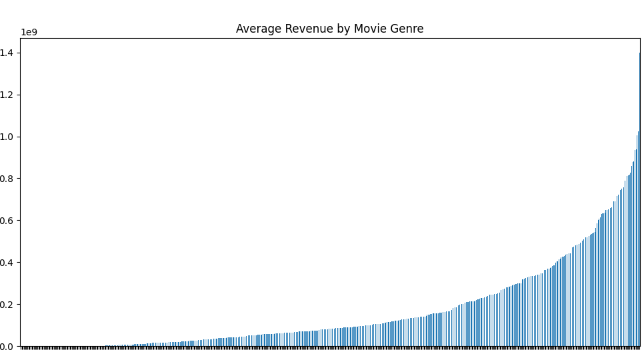
print("Average Revenue by Genre:")

print(genre\_avg\_revenue)

print("\nGenre Statistics (Median, Standard Deviation):")

print(genre\_stats)

**代码运行结果：**



**结果分析：**

电影数量：Drama 类型的电影数量最多，共有 2297 部；而 TV Movie 类型的电影数量最少，仅有 8 部。

票房收入：

最低票房：Action 类型的电影最低票房收入最高，为 1.154 亿美元。

平均票房：Adventure 类型的电影平均票房收入最高，为 2.846 亿美元。

最高票房：Adventure 类型的电影最高票房收入也是最高的，为 2.787亿美元。

成本：

最低成本：TV Movie 类型的电影最低成本为零。

平均成本：Adventure 类型的电影平均成本最高，为 3.002 亿美元。

最高成本：Adventure 类型的电影最高成本也是最高的，为 2.787 亿美元。

票房与成本比较：

从平均票房与平均成本的比较来看，Adventure 类型的电影在票房上表现最佳，但成本也相对较高。

另一方面，Comedy 类型的电影平均票房与平均成本之间的差距最小，表明这类电影可能具有较高的性价比效益。

特定类型的表现：

Action和Fantasy 类型的电影在票房上表现强劲，这可能与这些类型的电影通常具有较高的制作成本和吸引观众的特效有关。

Documentary和Foreign 类型的电影在票房上的表现相对较弱。

市场潜力：

从数据中可以看出，Action、Adventure、Fantasy 和 Science Fiction 这些类型的电影在票房上具有较大的潜力。

Drama 和 Comedy 类型的电影虽然数量众多，但平均票房收入相对较低。

零收入电影：

TV Movie 类型的电影数据显示零收入，这可能是因为这些电影主要通过电视或在线平台播放，而不是在传统影院上映。