#### 个人工作报告

在本次电影票房数据分析项目中，我主要负责数据筛选工作，以下是我对个人工作的详细报告。

**一、工作内容概述**

我的主要任务是对从 The number 电影网站获取的原始票房数据进行处理与分析，确保数据的准确性、完整性和有效性，为后续的深入研究提供坚实的数据基础。具体工作涵盖了数据读取与初步处理、数据去重与整合操作、数据分析与统计分析以及数据可视化与结果输出等多个环节。

**二、工作完成过程**

（一）数据处理流程设计

1、数据读取与初步处理

利用 pandas 库的 read\_excel 函数读取存储数据的 top\_movies.xlsx 文件，获取所有工作表的数据。

针对 Box Office 列，先去除其中的逗号，再将其转换为数值类型，确保后续计算和分析的准确性。

2、数据去重与整合

考虑到多部电影在三年票房数据中多次出现，以 Title 列为分组依据，计算每组 Box Office 列的总和，得到每个电影的总票房，从而保证数据唯一性。

将去重后的原始数据与分组求和后的数据合并，保留准确的票房汇总信息，形成更干净、完整的数据集 data。

（二）数据去重与整合操作

1、按年份统计票房总和

计算每年票房总和，得到 yearly\_box\_office 数据，用于观察电影市场年度票房表现，为分析票房年度变化趋势提供基础。

2、电影类型分布分析

明确各类电影在市场中的占比，洞察观众对不同电影类型的喜好倾向。

3、筛选票房排名前五十电影

对数据集 data 降序排序，选取前五十部电影并添加排名信息，生成 top\_50\_movies 数据集，为研究高票房电影特征和市场表现提供样本。

4、探究电影上映日期与票房关系

以 “电影上映日期” 分组，对 “票房” 求和，得到 release\_date\_box\_office 数据，探索两者关联，为电影发行策略研究提供参考。

5、电影票房与发布日期分类统计

根据票房数值和上映日期月份（分季度）分别对电影进行分类统计，得到 box\_office\_classification 和 release\_date\_classification 数据，从不同角度分析电影市场结构和分布特点。

6、电影发布日期与电影类型联合分类统计

按Release Date Class 和Genres 两列分组统计数量，得到 release\_date\_genre\_classification 数据，深入分析电影市场细分特征和规律。

（三）结果展示设计

创建新的 Excel 文件 top\_movies\_analysis\_charts.xlsx，使用 pandas 的 ExcelWriter 功能将各类分析数据写入相应工作表。同时利用 xlsxwriter 引擎在该文件中创建 “图表” 工作表，绘制多种可视化图表，如折线图、饼图、柱状图等，直观展示数据处理和分析结果。

1. **工作成果与收获总结**

（一）工作成果

通过以上一系列的数据筛选工作，我将原始一个关于2020-2022三年内票房前100的电影数据转化为了清晰、有价值的数据集和可视化图表。这些成果为后续的数据分析，结合 AI 模型，提供了高质量的数据基础，有助于更深入地洞察电影市场的趋势和规律；同时，数据可视化图表也使得数据结果更加直观易懂，方便非技术人员理解和解读电影市场的相关信息。

1. 收获总结

熟练掌握了 pandas 库，在数据处理过程中，深入学习并熟练运用 pandas 库的各种函数和方法，如数据读取（read\_excel）、数据清洗（去除逗号、转换数据类型）、数据分组（groupby）、计算汇总（sum）等操作，大大提高了数据处理效率和准确性。

数据可视化技能增强，学习使用 xlsxwriter 引擎在 Excel 中创建多种可视化图表，了解不同图表类型（折线图、饼图、柱状图等）的适用场景，能够根据数据特点选择合适的图表进行展示，使数据更直观、易懂。