一、项目概述

本项目旨在对电影数据进行多维度的分析，并将分析结果以图表形式展示在新生成的 Excel 文件中。主要分析内容包括电影票房总和随年份的变化、电影类型分布、票房排名前五十的电影、电影上映日期与票房的关系、电影票房分类以及电影发布日期分类等，并针对电影发布日期和电影类型的组合情况进行分类统计。

二、运行环境

开发工具：Visual Studio Code。

Python 版本：确保安装了 Python 3.x 版本，本代码基于 Python 的数据分析库进行开发，在运行前请确认已正确安装相关依赖库。

三、依赖库安装

本项目主要依赖 pandas 和 xlsxwriter 库。

Pandas库：用于数据处理与分析，提供了高效的数据结构和数据分析工具，如数据读取、清洗、转换和统计分析等功能。

xlsxwriter 库：用于创建 Excel 文件并写入数据和图表，支持多种图表类型和丰富的格式设置选项。

四、生成表格过程说明

（1）数据读取与预处理

读取指定 Excel 文件所有工作表数据并存储，选取特定工作表，清理票房数据中的逗号并转换为数值类型，以便后续处理。

1. 数据处理与分析
2. 去重并汇总票房，先分组计算各电影总票房后合并数据。
3. 按年份分组算出每年票房总和，

记录每年的票房总和，Year列表示年份，Box Office列表示票房。

1. 统计电影类型分布。

记录不同电影类型的数量，其中Genres列表示电影类型，count列表示数量。

1. 依票房筛选出排名前五十的电影并整理列名，

记录票房排名前五十的电影的相关信息，其中rank列表示排名，movie\_name列表示电影名称，box\_office列表示票房，release\_date列表示上映日期，year列表示年份。

1. 分析电影上映日期与票房关系，按上映日期分组求和。

记录票房排名前五十的电影的相关信息，其中rank列表示排名，movie\_name列表示电影名称，box\_office列表示票房，release\_date列表示上映日期，year列表示年份。

1. 依据票房数值划分电影票房类别并统计各类电影数量。

Box Office Class列表示票房分类，count列表示数量。

1. 按季度对电影发布日期分类并统计，还对发布日期与电影类型组合分类统计。

Release Date Class列表示发布季节分类，count列表示数量。关于电影发布日期和电影类型的分类，Release Date Class列表示发布日期分类，Genres列表示电影类型，数量列表示数量。

（3）数据输出与图表创建

创建 Excel 文件并将各分析结果写入对应工作表。在该文件中新建图表工作表，依次创建每年票房总和折线图、电影类型分布饼图、电影票房分类饼图、电影发布日期分类柱状图、电影发布日期与类型分类柱状图，分别设置图表数据来源、标题、标签等信息并插入指定位置。

1. 保存 Excel 文件

关闭 ExcelWriter，将数据与图表存入指定 Excel 文件。

代码部分：

import pandas as pd

# 读取Excel文件，获取所有工作表的数据

dfs = pd.read\_excel('E:\\学习\\大三上\\python\\top\_movies.xlsx', sheet\_name=None)

# 获取所有表名

sheet\_names = list(dfs.keys())

sheet\_names

df = dfs['Sheet1']

# 去除票房数据中的逗号

df['Box Office'] = df['Box Office'].str.replace(',', '')

# 将票房数据转换为数值类型

df['Box Office'] = pd.to\_numeric(df['Box Office'])

# 读取Sheet1数据

data = df

print('数据基本信息：')

data.info()

print('数据的前几行：')

print(data.head())

# 进行去重，得到去重数据data

# 对Title列进行分组，计算每组中Box Office列的总和

group\_data = data.groupby('Title')['Box Office'].sum().reset\_index()

# 合并

data = pd.merge(data.drop\_duplicates(subset='Title'), group\_data, on='Title')

# 按照上映年份进行分组，并计算每年的票房总和

yearly\_box\_office = df.groupby('Year')['Box Office'].sum().reset\_index()

# 分析电影类型的分布情况

genre\_distribution = data['Genres'].value\_counts()

# 找出票房排名前五十的电影（根据处理后的票房重新排名）

# 这里将排序的列名改为合并后正确的列名 'Box Office\_y'

top\_50\_movies = data.sort\_values('Box Office\_y', ascending=False).head(50).reset\_index()

top\_50\_movies['rank'] = range(1, 51)

top\_50\_movies = top\_50\_movies[['rank', 'Title', 'Box Office\_y', 'Release Date', 'Year']]

top\_50\_movies.columns = ['rank', 'movie\_name', 'box\_office', 'release\_date', 'year']

# 分析电影上映日期与票房的关系

release\_date\_box\_office = data.groupby('Release Date')['Box Office\_y'].sum()

# 对电影票房进行分类

def classify\_box\_office(box\_office):

if box\_office >= 10000000:

return '一千万以上'

elif box\_office >= 5000000:

return '五百万以上'

elif box\_office >= 1000000:

return '一百万以上'

else:

return '一百万以下'

data['Box Office Class'] = data['Box Office\_y'].apply(classify\_box\_office)

box\_office\_classification = data['Box Office Class'].value\_counts()

# 对电影发布日期分类

def classify\_release\_date(release\_date):

month = pd.to\_datetime(release\_date).month

if month in [1, 2, 3]:

return '第一季度'

elif month in [4, 5, 6]:

return '第二季度'

elif month in [7, 8, 9]:

return '第三季度'

else:

return '第四季度'

data['Release Date Class'] = data['Release Date'].apply(classify\_release\_date)

release\_date\_classification = data['Release Date Class'].value\_counts()

# 对电影发布日期和电影类型进行分类

release\_date\_genre\_classification = data.groupby(['Release Date Class', 'Genres']).size().reset\_index(name='数量')

# 创建一个新的Excel文件

wb = pd.ExcelWriter('E:\\学习\\大三上\\python\\top\_movies\_analysis\_charts.xlsx', engine='xlsxwriter')

# 将数据写入相应的工作表

yearly\_box\_office.to\_excel(wb, sheet\_name='每年票房总和', index=False)

genre\_distribution.to\_excel(wb, sheet\_name='电影类型分布情况', index=True)

top\_50\_movies.to\_excel(wb, sheet\_name='票房排名前五十的电影', index=False)

release\_date\_box\_office.to\_excel(wb, sheet\_name='电影上映日期与票房的关系', index=True)

box\_office\_classification.to\_excel(wb, sheet\_name='电影票房分类', index=True)

release\_date\_classification.to\_excel(wb, sheet\_name='电影发布日期分类', index=True)

release\_date\_genre\_classification.to\_excel(wb, sheet\_name='电影发布日期和电影类型分类', index=False)

# 获取工作簿对象

workbook = wb.book

# 创建一个新的图表工作表

chart\_sheet = workbook.add\_worksheet('图表')

# 创建每年票房总和折线图

chart\_yearly\_box\_office = workbook.add\_chart({'type': 'line'})

chart\_yearly\_box\_office.add\_series({

'name': '票房总和',

'categories': '=每年票房总和!$A$2:$A$' + str(yearly\_box\_office.shape[0] + 1),

'values': '=每年票房总和!$B$2:$B$' + str(yearly\_box\_office.shape[0] + 1)

})

chart\_yearly\_box\_office.set\_title({'name': '每年票房总和'})

chart\_yearly\_box\_office.set\_x\_axis({'name': '年份'})

chart\_yearly\_box\_office.set\_y\_axis({'name': '票房总和'})

chart\_sheet.insert\_chart('A1', chart\_yearly\_box\_office)

# 创建电影类型分布情况饼图

chart\_genre\_distribution = workbook.add\_chart({'type': 'pie'})

chart\_genre\_distribution.add\_series({

'name': '电影类型分布情况',

'categories': '=电影类型分布情况!$A$2:$A$' + str(genre\_distribution.shape[0] + 1),

'values': '=电影类型分布情况!$B$2:$B$' + str(genre\_distribution.shape[0] + 1)

})

chart\_genre\_distribution.set\_title({'name': '电影类型分布情况'})

# 设置饼图的数据标签格式，显示类别名称、值和百分比

data\_labels\_format = {

'num\_format': '0.00%', # 设置百分比格式，保留两位小数

'font': {'name': '宋体', 'size': 9}, # 设置字体为宋体，大小为9

'border': {'color': 'black'}, # 设置边框颜色为黑色

'fill': {'color': 'white'}, # 设置背景颜色为白色

}

chart\_genre\_distribution.set\_legend({'value': True, 'category': True, 'percentage': True, 'data\_labels\_format': data\_labels\_format})

chart\_sheet.insert\_chart('A20', chart\_genre\_distribution)

# 创建电影票房分类饼图

chart\_box\_office\_classification = workbook.add\_chart({'type': 'pie'})

chart\_box\_office\_classification.add\_series({

'name': '电影票房分类',

'categories': '=电影票房分类!$A$2:$A$' + str(box\_office\_classification.shape[0] + 1),

'values': '=电影票房分类!$B$2:$B$' + str(box\_office\_classification.shape[0] + 1)

})

chart\_box\_office\_classification.set\_title({'name': '电影票房分类'})

# 设置饼图的数据标签格式，显示类别名称、值和百分比

data\_labels\_format = {

'num\_format': '0.00%', # 设置百分比格式，保留两位小数

'font': {'name': '宋体', 'size': 9}, # 设置字体为宋体，大小为9

'border': {'color': 'black'}, # 设置边框颜色为黑色

'fill': {'color': 'white'}, # 设置背景颜色为白色

}

chart\_box\_office\_classification.set\_legend({'value': True, 'category': True, 'percentage': True, 'data\_labels\_format': data\_labels\_format})

chart\_sheet.insert\_chart('A40', chart\_box\_office\_classification)

# 创建电影发布日期分类柱状图

chart\_release\_date\_classification = workbook.add\_chart({'type': 'bar'})

chart\_release\_date\_classification.add\_series({

'name': '电影数量',

'categories': '=电影发布日期分类!$A$2:$A$' + str(release\_date\_classification.shape[0] + 1),

'values': '=电影发布日期分类!$B$2:$B$' + str(release\_date\_classification.shape[0] + 1)

})

chart\_release\_date\_classification.set\_title({'name': '电影发布日期分类'})

chart\_release\_date\_classification.set\_x\_axis({'name': '季度'})

chart\_release\_date\_classification.set\_y\_axis({'name': '电影数量'})

chart\_sheet.insert\_chart('A60', chart\_release\_date\_classification)

# 创建电影发布日期和电影类型分类柱状图

chart\_release\_date\_genre\_classification = workbook.add\_chart({'type': 'bar'})

for genre in release\_date\_genre\_classification['Genres'].unique():

genre\_data = release\_date\_genre\_classification[release\_date\_genre\_classification['Genres'] == genre]

chart\_release\_date\_genre\_classification.add\_series({

'name': genre,

'categories': '=电影发布日期和电影类型分类!$A$2:$A$' + str(genre\_data.shape[0] + 1),

'values': '=电影发布日期和电影类型分类!$C$2:$C$' + str(genre\_data.shape[0] + 1)

})

chart\_release\_date\_genre\_classification.set\_title({'name': '电影发布日期和电影类型分类'})

chart\_release\_date\_genre\_classification.set\_x\_axis({'name': '季度'})

chart\_release\_date\_genre\_classification.set\_y\_axis({'name': '电影数量'})

chart\_release\_date\_genre\_classification.set\_legend({'position': 'right'})

chart\_sheet.insert\_chart('A80', chart\_release\_date\_genre\_classification)

# 保存Excel文件

wb.close()