实训一：分析学生考试成绩特征的分布与分散情况

1. 导入所需库并读取数据：

源程序：

```
import pandas as pd
import matplotlib.pyplot as plt

# 读取数据
data = pd.read_excel('student_grade.xlsx')

# 查看数据
display(data)
```

过程性结果：

1. 导入所需库并读取数据：

[11]:

```
import pandas as pd
import matplotlib.pyplot as plt

# 读取数据
data = pd.read_excel('student_grade.xlsx')

# 查看数据
display(data)
```

	性别	文化成绩	完成情况	阅读成绩	写作成绩	总成绩
0	女	中	未完成	72	74	218
1	女	中	完成	69	90	247
2	女	高	未完成	72	73	188
3	女	高	完成	91	96	276
4	男	中	未完成	47	57	148
5	男	中	完成	76	78	229

2. 将学生考试总成绩分为 4 个区间，计算各区间下的学生人数，并绘制分布图：

源程序：

```
plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False #用来正常显示负号

# 定义总成绩区间
bins = [0, 150, 200, 250, 300]
labels = ['不及格', '及格', '良好', '优秀']

# 将总成绩分区间
data['总成绩区间'] = pd.cut(data['总成绩'], bins=bins, labels=labels)

# 计算各区间下的学生人数
score_distribution = data['总成绩区间'].value_counts().sort_index()

# 绘制分布图
plt.figure(figsize=(10, 6))
score_distribution.plot(kind='bar', color='skyblue')
plt.xlabel('总成绩区间')
plt.ylabel('学生人数')
plt.title('学生考试总成绩分布')
plt.show()
```

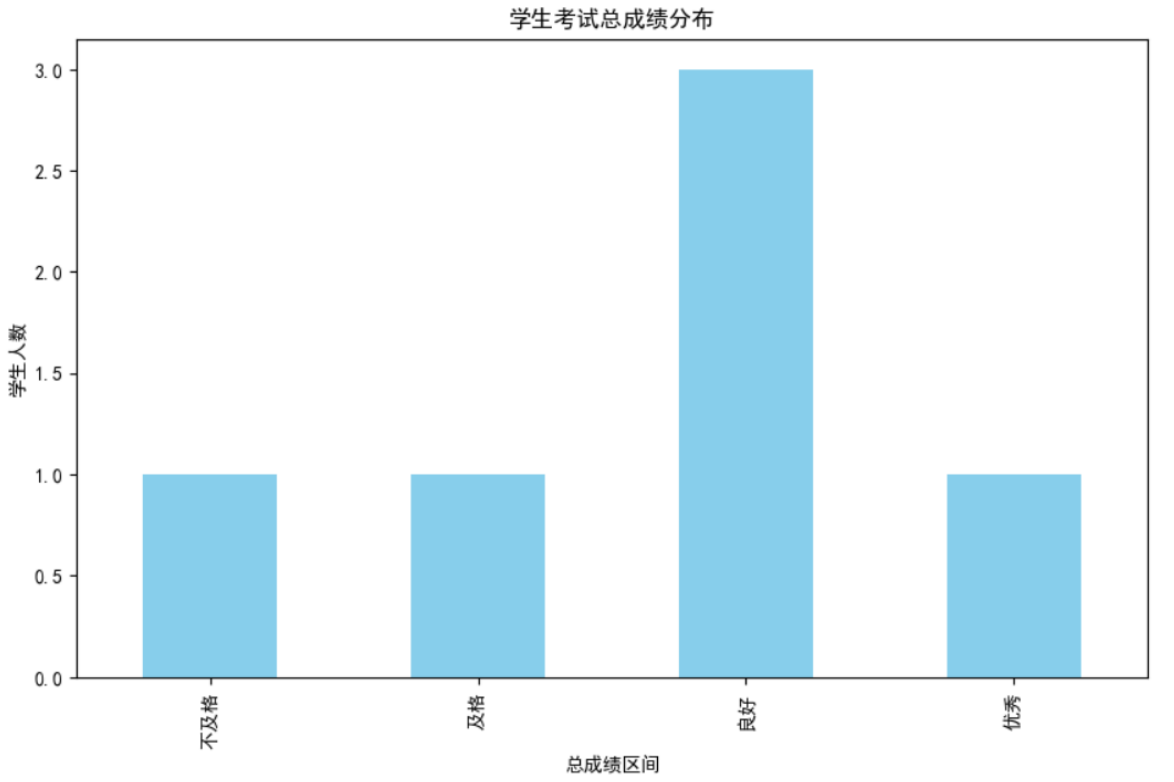
过程性结果：

```
# 定义总成绩区间
bins = [0, 150, 200, 250, 300]
labels = ['不及格', '及格', '良好', '优秀']

# 将总成绩分区间
data['总成绩区间'] = pd.cut(data['总成绩'], bins=bins, labels=labels)

# 计算各区间下的学生人数
score_distribution = data['总成绩区间'].value_counts().sort_index()

# 绘制分布图
plt.figure(figsize=(10, 6))
score_distribution.plot(kind='bar', color='skyblue')
plt.xlabel('总成绩区间')
plt.ylabel('学生人数')
plt.title('学生考试总成绩分布')
plt.show()
```



结论：总体上学生考试总成绩分布中良好的最多，其他的半斤八两。

3. 获取学生 3 项单科成绩的数据，绘制分数分散情况箱线图：

源程序：

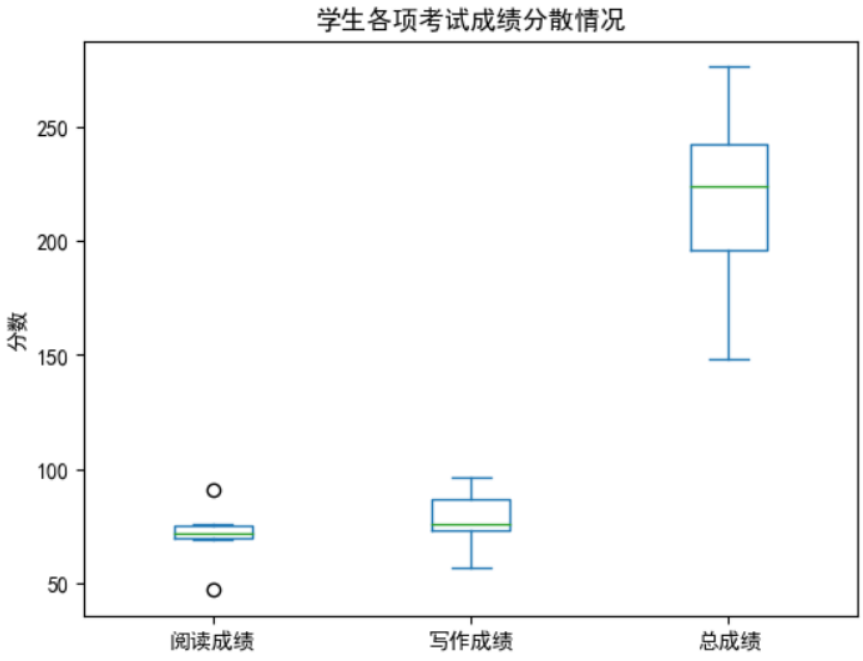
```
# 绘制单科成绩的箱线图
plt.figure(figsize=(10, 6))
data[['阅读成绩', '写作成绩', '总成绩']].plot(kind='box')
plt.ylabel('分数')
plt.title('学生各项考试成绩分散情况')
plt.show()
```

过程性结果：

3. 获取学生3项单科成绩的数据，绘制分数分散情况箱线图：

[13]:
```
# 绘制单科成绩的箱线图
plt.figure(figsize=(10, 6))
data[['阅读成绩', '写作成绩', '总成绩']].plot(kind='box')
plt.ylabel('分数')
plt.title('学生各项考试成绩分散情况')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



结论：总成绩分散程度最高。其他的还好。

实训二：分析学生考试成绩与各个特征之间的关系

1. 导入所需库并读取数据：

源程序：

import pandas as pd

import matplotlib.pyplot as plt

重新读取数据

data = pd.read_excel('student_grade.xlsx')

过程性结果：

实训二：分析学生考试成绩与各个特征之间的关系

1. 导入所需库并读取数据：

[19]:
```
import pandas as pd
import matplotlib.pyplot as plt

# 重新读取数据
data = pd.read_excel('student_grade.xlsx')
display(data)
```

	性别	文化成绩	完成情况	阅读成绩	写作成绩	总成绩
0	女	中	未完成	72	74	218
1	女	中	完成	69	90	247
2	女	高	未完成	72	73	188
3	女	高	完成	91	96	276
4	男	中	未完成	47	57	148
5	男	中	完成	76	78	229

结论：成功读取重置了数据。

2. 计算不同特征下学生总成绩的均值：

源程序：

计算不同特征下的总成绩均值

mean_scores = data.groupby(['性别', '文化成绩', '完成情况'])['总成绩'].mean().unstack()

查看结果

display(mean_scores)

过程性结果：

[22]:
```
# 计算不同特征下的总成绩均值
mean_scores = data.groupby(['性别', '文化成绩', '完成情况'])['总成绩'].mean().unstack()

# 查看结果
display(mean_scores)
```

	完成情况	完成	未完成
性别	文化成绩		
女	中	247.0	218.0
	高	276.0	188.0
男	中	229.0	148.0

结论：可以看到女生不管完没完成、文化成绩怎么样，总成绩都相对较高。

3. 绘制对应内容的柱形图：

源程序：

绘制柱形图

mean_scores.plot(kind='bar', figsize=(12, 8))

plt.ylabel('平均总成绩')

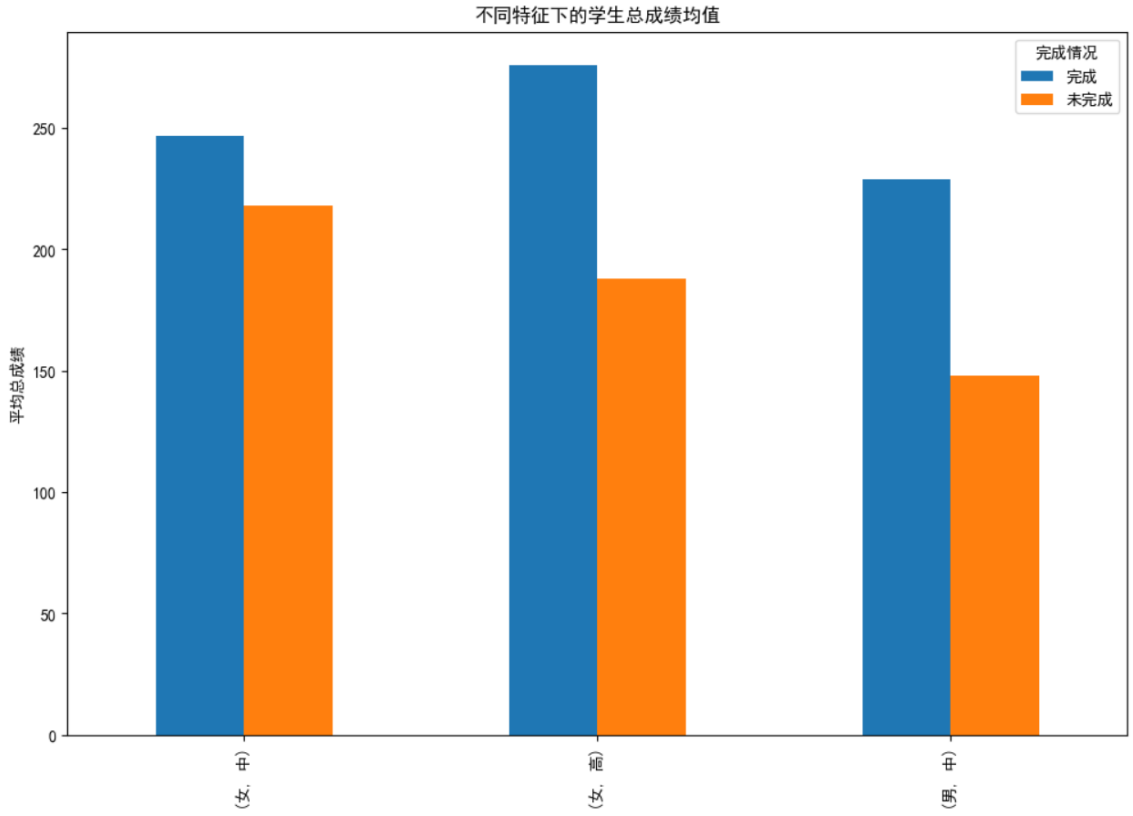
plt.title('不同特征下的学生总成绩均值')

plt.show()

过程性结果：

3. 绘制对应内容的柱形图：

[23]:
```
# 绘制柱形图
mean_scores.plot(kind='bar', figsize=(12, 8))
plt.ylabel('平均总成绩')
plt.title('不同特征下的学生总成绩均值')
plt.show()
```



总结论：

1. 性别与总成绩的关系：

从图中可以看到，无论是完成情况如何，女性学生（女）在文化成绩为“高”的情况下，总成绩的平均值最高。

对于男性学生（男），文化成绩为“中”的情况下，完成情况对总成绩也有显著影响。

2. 文化成绩与总成绩的关系：

对于女性学生（女），文化成绩为“高”的情况下，不论完成情况如何，总成绩都显著高于文化成绩为“中”的情况下的总成绩。

对于男性学生（男），文化成绩为“中”的情况下，完成情况对总成绩有显著影响，完成的学生总成绩显著高于未完成的学生。

3. 完成情况与总成绩的关系：

无论男女，完成任务的学生总成绩都明显高于未完成任务的学生。这表明完成任务的情况对总成绩有明显的影响。

结论：

文化成绩和完成情况对学生的总成绩有显著影响，文化成绩越高，完成情况越好，总成绩越高。

完成情况对男性学生的影响更加显著，而对女性学生的影响相对较小。

性别在文化成绩为“高”的情况下，对总成绩的影响不明显。