**模块B：数据分析（30分）**

**竞赛题目：奥林匹克数据分析**

**背景介绍**

奥林匹克运动会，是国际奥林匹克委员会主办的世界规模最大的综合性运动会，每四年一届，会期不超过16日，是世界上影响力最大的体育盛会。奥林匹克运动会发源于两千多年前的古希腊，因举办地在奥林匹亚而得名。古代奥林匹克运动会停办了1500年之后，法国人顾拜旦于19世纪末提出举办现代奥林匹克运动会的倡议。1894年成立奥委会，1896年举办了首届奥运会。接下来让我们来分析一下各国在奥运史上的表现。

**环境介绍**

在本竞赛模块环境中，已自动启动了 jupyter notebook 环境。环境中已安装 Python 中多种数据分析和可视化库，包括：

* Scipy
* Numpy
* Pandas
* Matplotlib
* Seaborn
* Pyecharts
* 其它

**数据介绍**

数据集包括从1896年雅典奥运会到2016年里约奥运会的比赛。数据集共有两个文件，分别为data/athlete\_events.csv和data/noc\_regions.csv。

athlete\_events.csv文件： 记录着运动员赛事(运动员的名字，参加的比赛，体重等等) 。 该文件包含271116行和15列，每行对应一个运动员参加单个奥林匹克项目的情况。具体信息如下：

| **属性** | **No** | **数据类型** | **字段描述** |
| --- | --- | --- | --- |
| ID | 1 | Integer | 给每个运动员的唯一ID |
| Name | 2 | String | 运动员名字 |
| Sex | 3 | Integer | 性别 |
| Age | 4 | Float | 年龄 |
| Height | 5 | Float | 身高 |
| Weight | 6 | Float | 体重 |
| Team | 7 | String | 所代表的国家队 |
| NOC | 8 | String | 国家奥委会3个字母的代码 |
| Games | 9 | String | 年份与季节 |
| Year | 10 | Integer | 比赛年份 |
| Season | 11 | String | 比赛季节 |
| City | 12 | String | 举办城市 |
| Sport | 13 | String | 运动类别 |
| Event | 14 | String | 比赛项目 |
| Medal | 15 | String | 奖牌 |

noc\_regions.csv文件：记录着国家奥委会地区信息。具体信息如下：

| **No** | **属性** | **数据类型** | **字段描述** |
| --- | --- | --- | --- |
| 1 | NOC | String | 国家奥委会3个字母的代码 |
| 2 | Region | String | 国家 |
| 3 | Notes | String | 地区 |

**任务列表**

1. 数据整合
2. 数据处理
3. 数据分析
4. 数据可视化

**注意事项**

每题完成之后必须保存该题结果，每题的保存代码在该题下的结果保存处，执行即可，禁止修改。

* 每题的最终结果变量名已预先设定，禁止修改
* 若没有保存结果则该题无分
* 若该题有修改需重新运行结果保存代码，否则视为上次提交

**任务启动**

**任务1：数据整合（1分）**

这部分任务为对数据进行导入，然后合并数据构建一个完整统一的数据源。具体任务要求如下：

* 载入 athlete\_events.csv 文件，数据路径为 ../data/athlete\_events.csv
* 载入 noc\_regions.csv 文件，数据路径为 ../data/noc\_regions.csv
* 合并上述两个文件，以data\_merge命名
* 导出合并后的数据集到该../result路径下，命名为data\_out.csv

**任务2：数据处理（2分）**

* 删除notes字段
* 提取Season字段中值为Summer的数据，然后删除Season字段
* 删除整个数据集中的重复值

**任务3：数据分析（8分）**

**任务3.1哪些国家每届奥运会都参加了？（2分）**

1896年—2016年，四年一届的奥运会共应举办31届。由于战争、届间运动会等影响，到今天为止共有29届奥运会成功举办。那么哪些国家每届奥运会都参加了？

* 统计每届奥运会都参加了的国家
* 结果以变量country保存，该变量的数据类型为list

**任务3.2中国参加了多少次奥运会？（1分）**

1896年—2016年间，中国参加了多少次奥运会？

* 统计中国参加了多少次奥运会
* 结果以变量china\_num保存，该变量的数据类型为int

**任务3.3 2008年奥运会中国获奖人数 （2分）**

* 统计2008年奥运会中国获奖人数
* 结果以变量china\_medal\_num\_2008保存，该变量的数据类型为int

**任务3.4 奥运会参与次数最多的选手（1分）**

* 统计奥运会参与次数最多的选手
* 结果以变量num\_max\_name保存，该变量的数据类型为str

**任务3.5 奥运会年龄最大的运动员（2分）**

* 排除掉Art Competitions体育项目，统计年龄最大的运动员
* 结果以变量age\_max\_name保存，该变量的数据类型为list

**任务4：数据可视化（19分）**

**任务4.1 奥运项目词云图（2分）**

针对Sport字段制作词云图

* 成功制作词云图（1分）
* 词云图图形icon为running（1分）

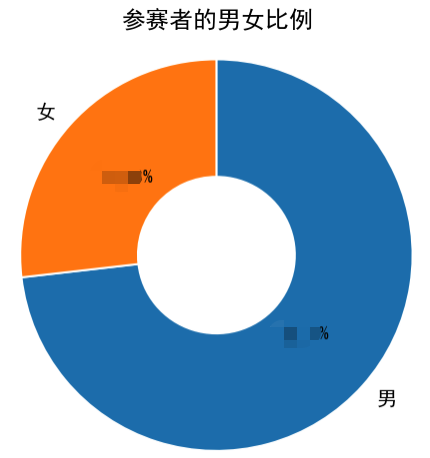
参考图形如下：



**任务4.2 奥运会参赛选手男女比例 - 环形图（2分）**

* 标题为参赛者的男女比例（0.5分）
* 参数选手男女比例保留一位小数（0.5分）
* 数据标签正确（1分）

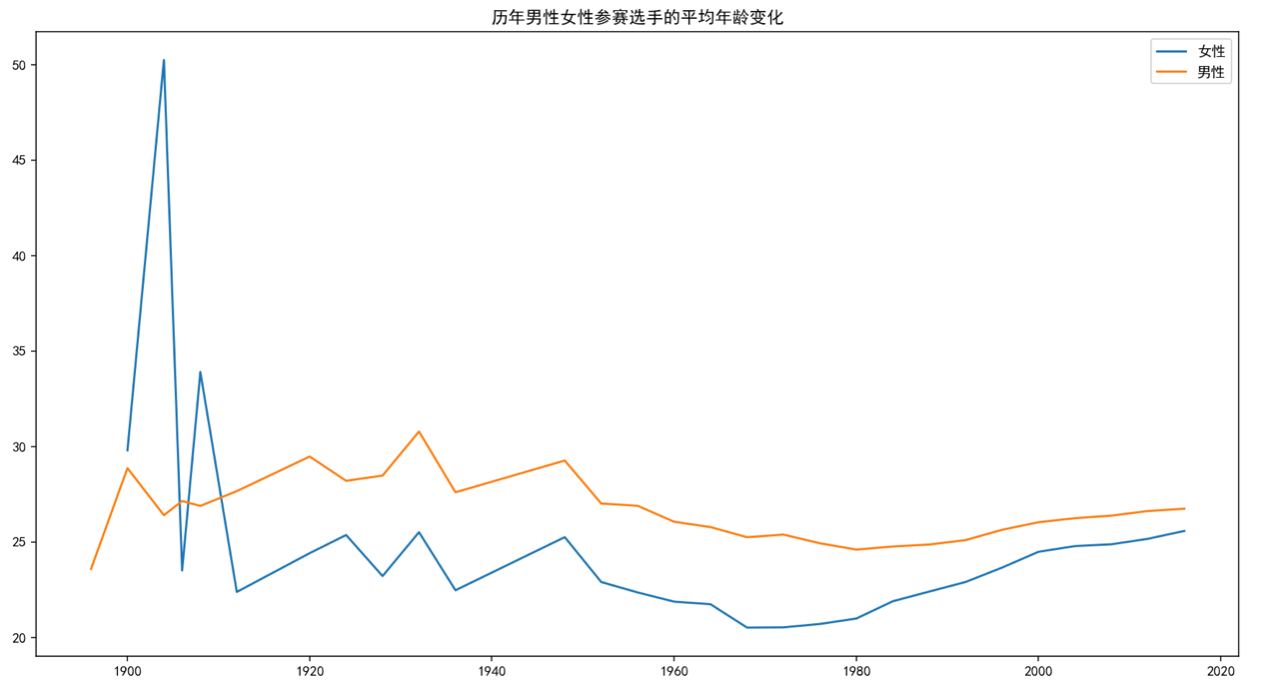
参考图形如下（马赛克里面的数值自行计算）：



**任务4.3 奥运会历年男女参赛者的平均年龄变化 - 折线图（2分）**

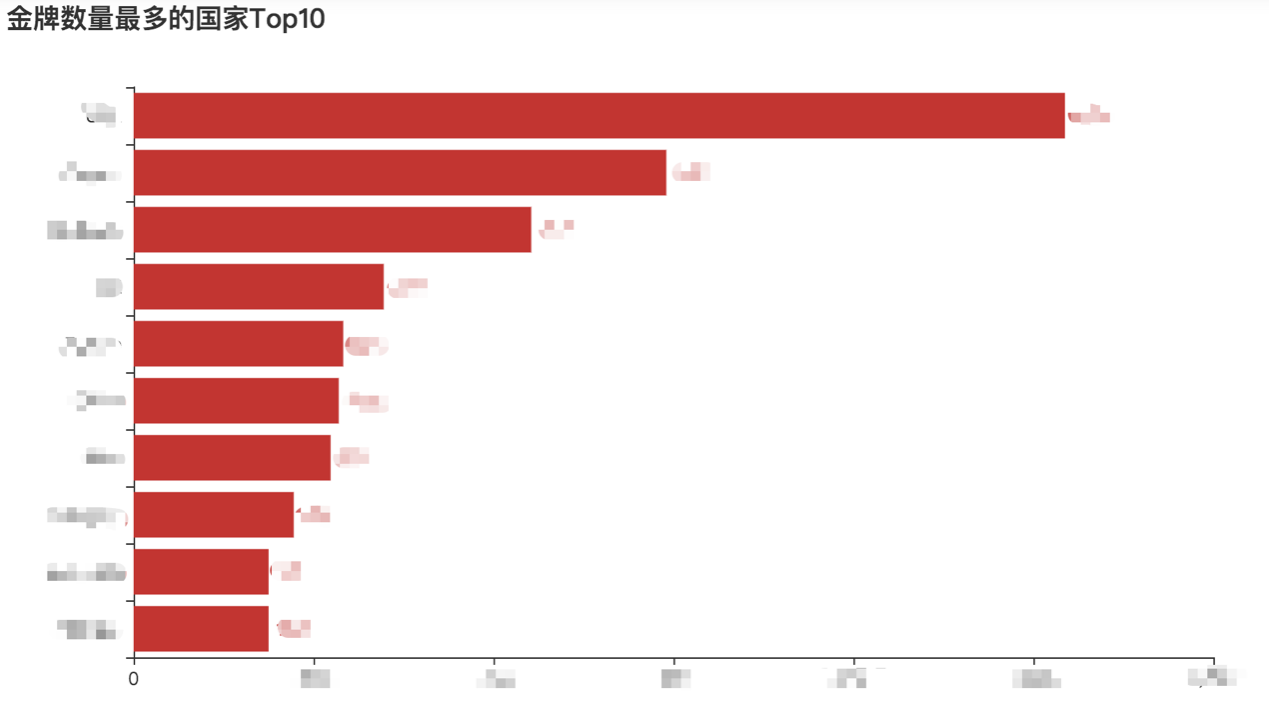
* 标题为历年男性女性参赛选手的平均年龄变化（0.5分）
* 添加图例（0.5分）
* 数据趋势正确（1分）

参考图形如下：



**任务4.4 奥运会获得金牌最多的国家 - 条形图（2分）**

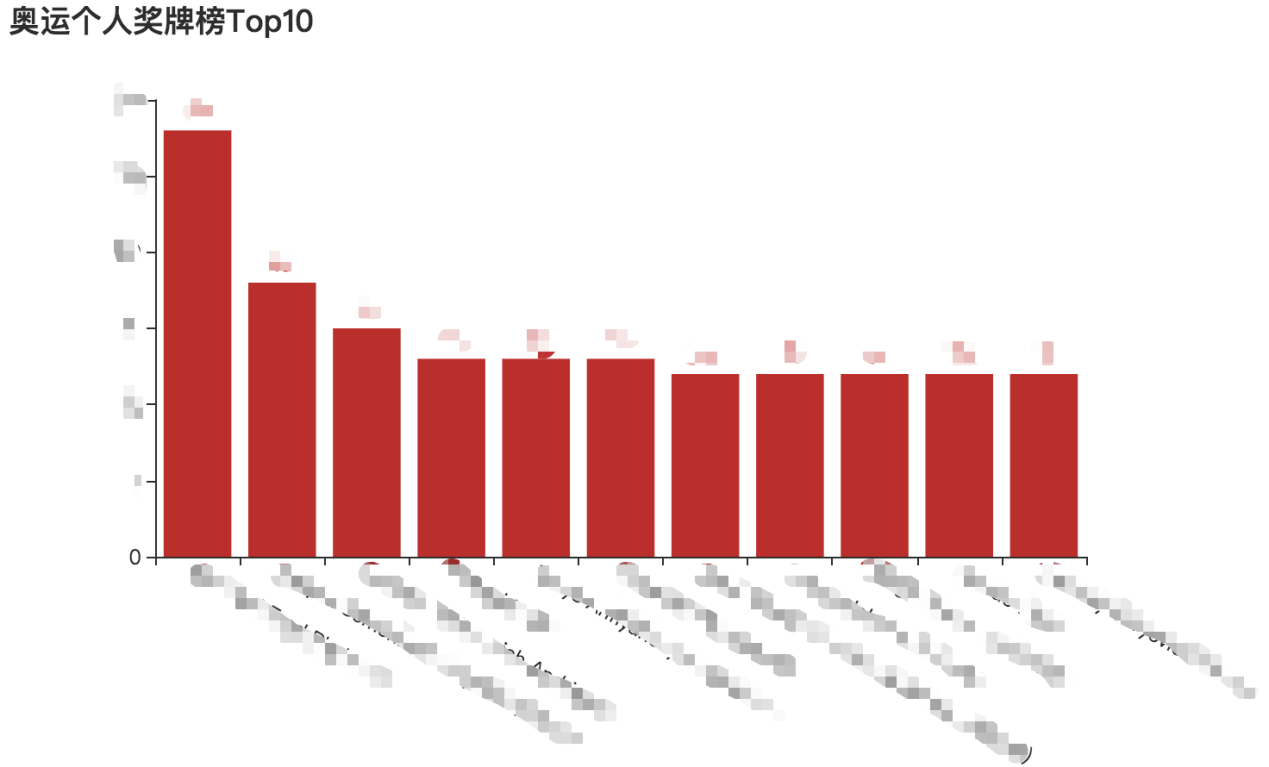
* 标题为金牌数量最多的国家Top10（0.5分）
* 纵坐标标签正确（0.5分）
* 数据标签正确（1分）

参考图形如下（马赛克里面的数值自行计算）：

**任务4.5 奥运会获得金牌最多的选手 - 柱状图（2分）**

* 标题为奥运个人奖牌榜Top10 （0.5分）
* 横轴标签旋转30度（0.5分）
* 横轴标签显示完整（0.5分）
* 数据标签数值正确（0.5分）

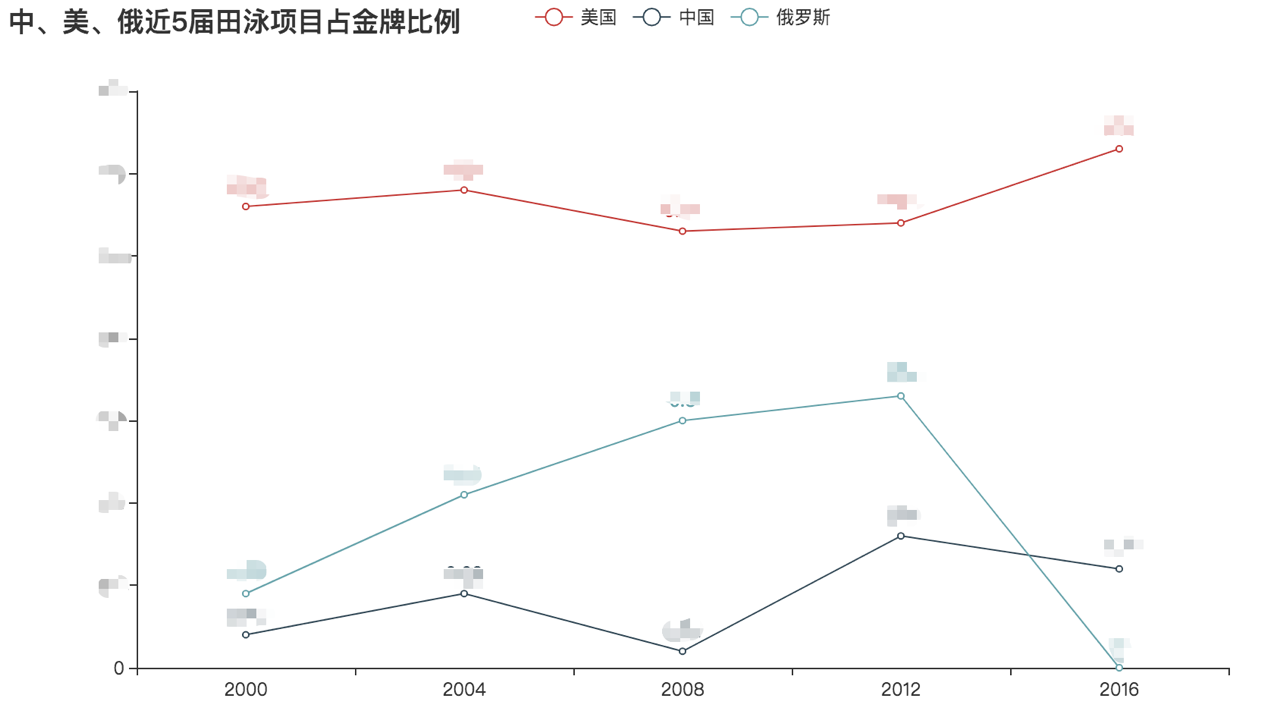
参考图形如下（马赛克里面的数值自行计算）：



**任务4.6 中、美、俄近5届田泳项目占金牌比例 - 折线图（2分）**

* 标题为中、美、俄近5届田泳项目占金牌比例（0.5分）
* 横坐标标签正确（0.5分）
* 数据标签正确（1分）

参考图形如下（马赛克里面的数值自行计算）：



**任务4.7 主场优势真的存在吗 - 折线矩阵图（2分）**

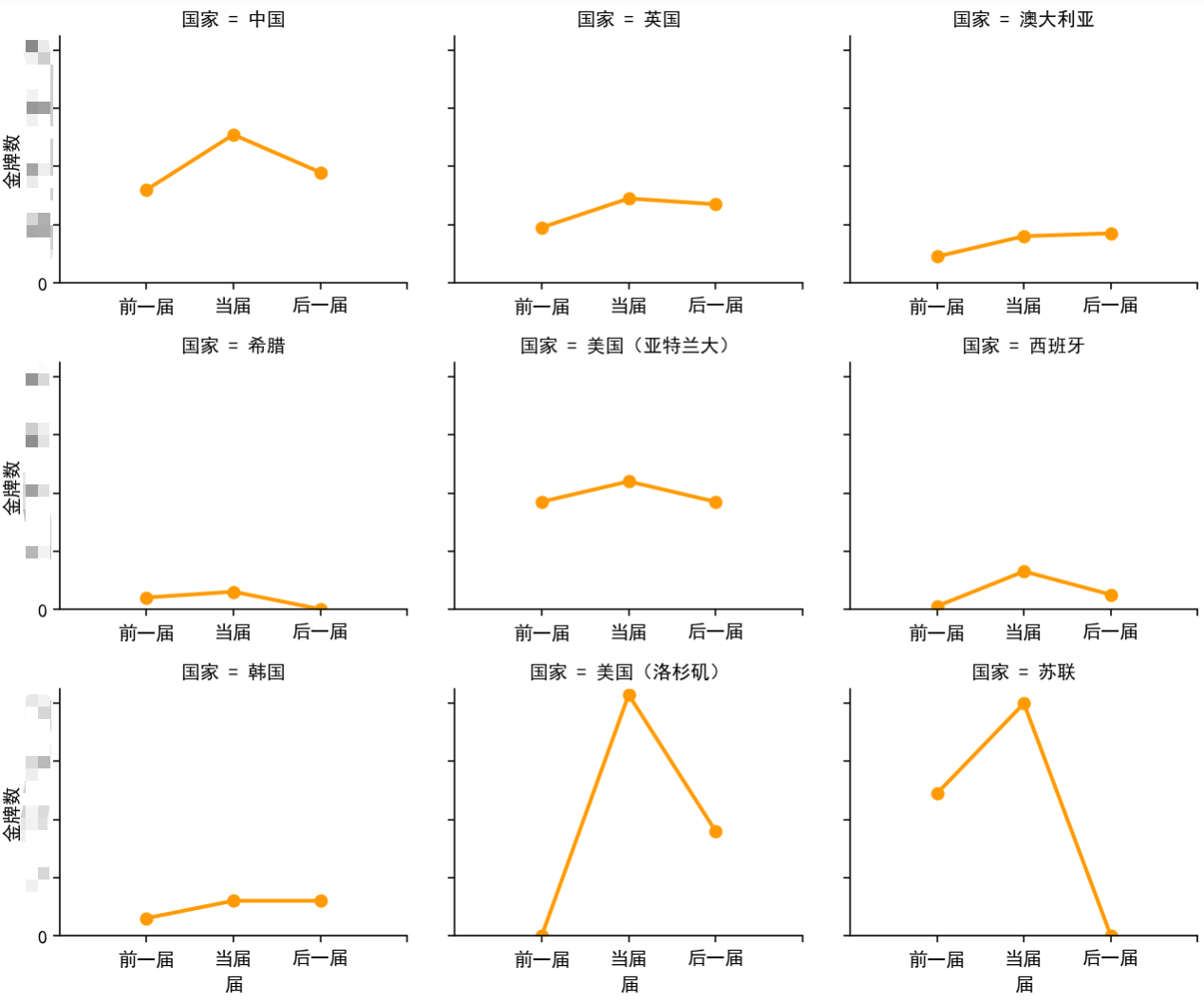
下表为近9届奥运会主办国的主办届及前后两届的金牌数量。

| **国家** | **前一届** | **当届** | **后一届** |
| --- | --- | --- | --- |
| 中国 | 32 | 51 | 38 |
| 英国 | 19 | 29 | 27 |
| 澳大利亚 | 9 | 16 | 17 |
| 希腊 | 4 | 6 | 0 |
| 美国（亚特兰大） | 37 | 44 | 37 |
| 西班牙 | 1 | 13 | 5 |
| 韩国 | 6 | 12 | 12 |
| 美国（洛杉矶） | 0 | 83 | 36 |
| 苏联 | 49 | 80 | 0 |

根据上面的数据制作折线矩阵图（3\*3），观察主场优势是否存在。

* 成功制作折线矩阵图（1分）
* 每个子图要有对应的国家标示（1分）

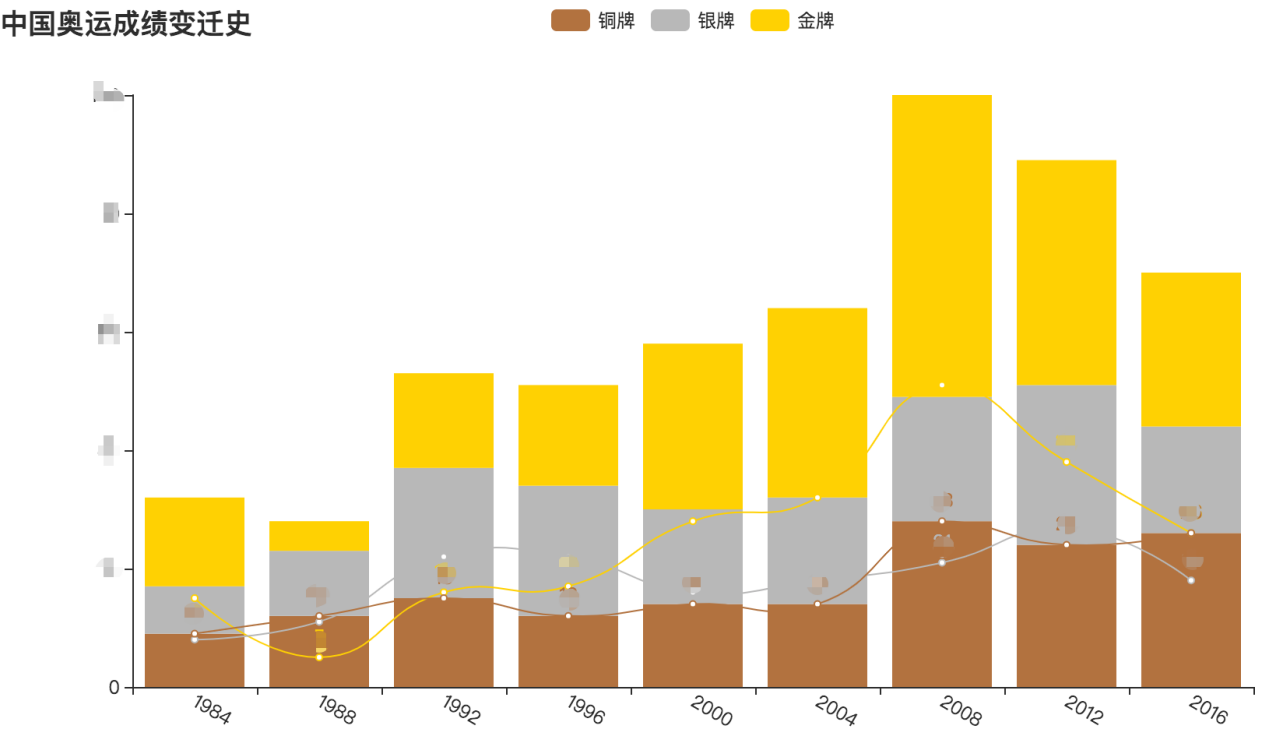
参考图形如下（马赛克里面的数值自行补充）：



**任务4.8中国奥运成绩变迁史 - 柱状图&折线图（2分）**

* 标题为中国奥运成绩变迁史（0.5分）
* 图形为柱形图和折线图结合，折线图显示在柱形图上方，不能被柱形图覆盖（0.5分）
* 折线图设置为顺滑模式（0.5分）
* 金银铜牌颜色分别为gold、silver、#ba7e48（0.5分）

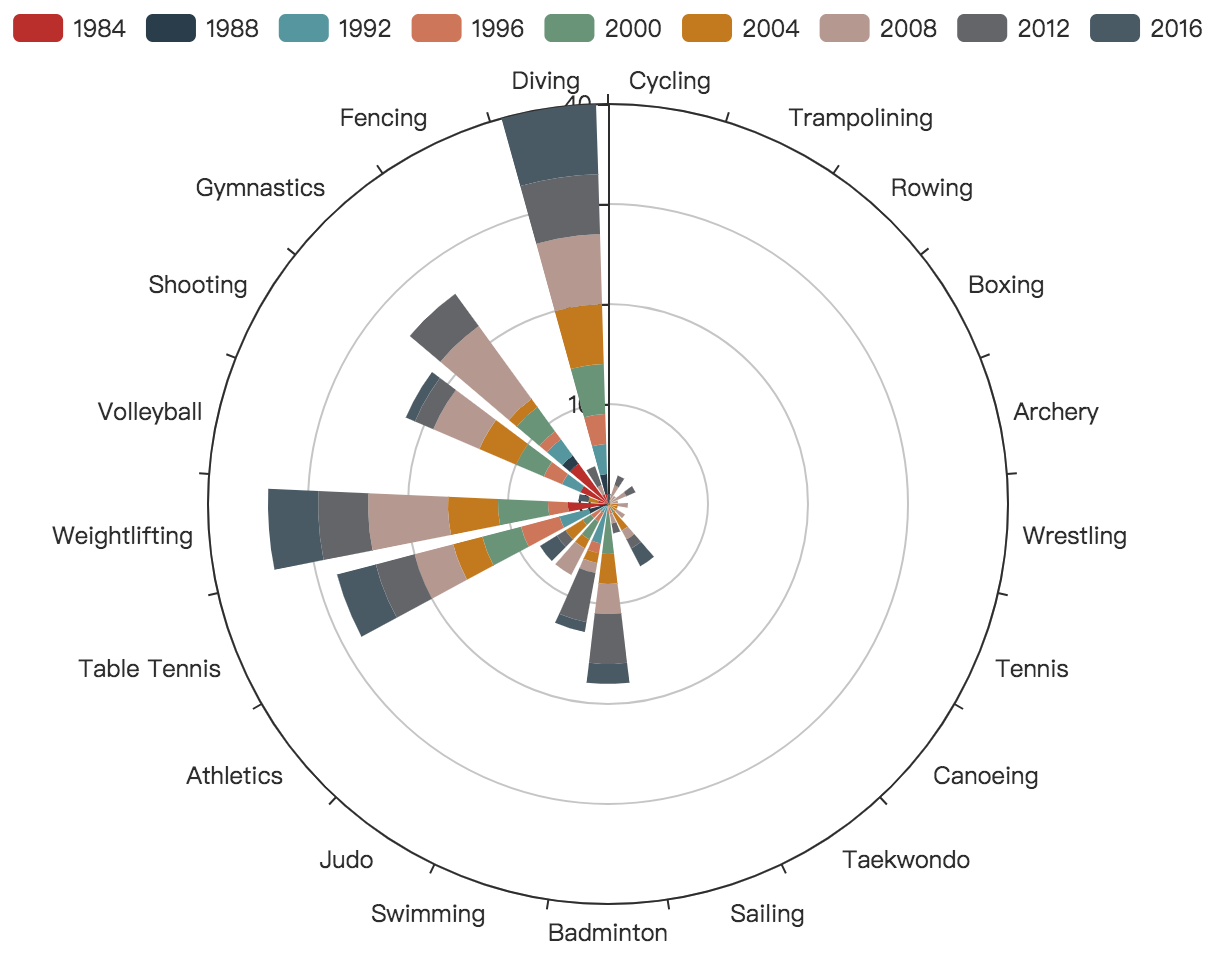
参考图形如下（马赛克里面的数值自行计算）：



**任务4.9 中国有哪些强势项目(金牌) - 蓝丁格尔图（3分）**

* 统计中国参加的奥运会中，每个项目获得的金牌数量，以蓝丁格尔图展示
* 成功绘制蓝丁格尔图（1.5分）
* 图例正确（0.5分）
* 坐标标签正确（0.5分）
* 数据趋势正确（0.5分）

参考图形如下：



**任务结束**

恭喜完成本次比赛，祝您取得好成绩。