



廣東工業大學

《数据可视化技术》期末大作业

课程名称 数据可视化技术

题目名称 题目 1

学生学院 计算机学院

专业班级 计算机科学与技术 20(3)

学 号 3120005043

学生姓名 张俊鸿

指导教师 林志毅

2022 年 11 月 24 日

目录

- 一 题目.....4
- 二 数据.....4
 - 2.1 数据选取.....4
 - 2.2 数据内容.....5
- 三 可视化工具.....6
- 四 数据预处理.....6
 - 4.1 空值数据清洗.....6
 - 4.2 国家名统一.....6
 - 4.3 特殊数值处理.....7
- 五 可视化方案以及部分可视化过程.....7
 - 5.1 可视化方案.....7
 - 5.2 部分可视化操作.....8
- 六 可视化结果.....11
 - 6.1 WorldCups 和 WorldCupMatches 热力图相关性分析.....11
 - 6.2 历届冠军及四强队伍分析.....12
 - 6.3 各届世界杯参与人数、参赛队伍、比赛数目分析.....14
 - 6.4 比赛进球数队伍数散点图.....15
 - 6.5 主客队得分对比及结果比较.....16
 - 6.6 主客场得分分布分析.....18
 - 6.7 著名比赛及体育场分析.....20
 - 6.8 著名足球强国、球员、教练分析.....23
 - 6.9 历史获奖队伍地理可视化分析.....26
- 七 总结与体会.....27

一 题目

大作业是对前面学过的数据可视化技术的一个总结、回顾和实践，因此，开始设计前学生一定要先回顾以前所学的内容，明确本次作业设计所要用到的技术点。

从网络上下载一组数据（自行获取），选择一种可视化工具（Excel、Tableau、Matlab、Echarts 等），设计一种可视化方案实现该数据的可视化，并做适当的数据分析（或挖掘）。

PS：可使用数据包括(不限于以下数据)：

（1）数据来源多样化（如爬虫、其他网站下载的相关数据、自己整理的相关数据等）；

（2）报告中包含除了基本数据集外的其他相关资料（网页、期刊、报纸、中英文资料等），以作为报告的支撑材料。

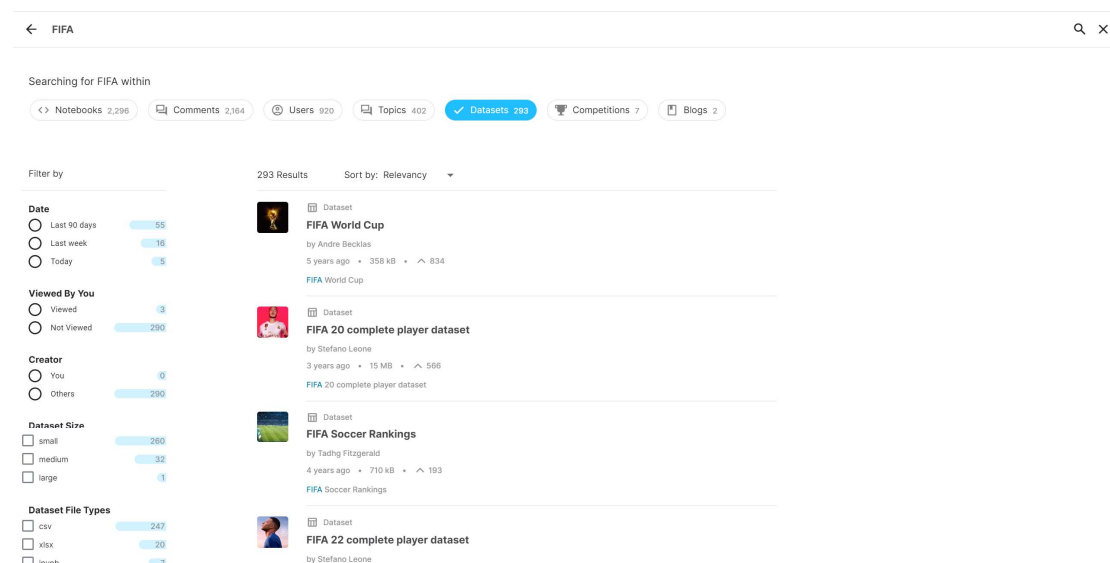
二 数据

2.1 数据选取

在开始着手数据可视化大作业选题的期间，正值 2022 年卡塔尔世界杯开幕。为了更好帮助我了解世界杯历史上的球员以及队伍，以及将我在课堂与课下学习到的数据可视化知识加以实践运用，我选择以“世界杯数据可视化”为主题，开展本次的数据可视化课程设计。

在确定主题之后，我就开始着手寻找合适的的数据源，随后在一番检索抉择后，我在 Kaggle 上找到了名为《FIFA World Cup》的数据集。

该数据集的网址：<https://www.kaggle.com/datasets/abecklas/fifa-world-cup>






2.2 数据内容

该数据集内容为 1930 年至 2014 年 FIFA 世界杯的全部数据（注：受第二次世界大战影响 1942 和 1946 年没有举办），总共 3 张表

大三作业 > 数据可视化 > 数据可视化课设 > data

在 data 中搜索

名称	修改日期	类型	大小
 WorldCupMatches.csv	2019-09-29 3:11	Microsoft Excel 逗...	23
 WorldCupPlayers.csv	2019-09-29 3:11	Microsoft Excel 逗...	2,10
 WorldCups.csv	2019-09-29 3:11	Microsoft Excel 逗...	

总共包含 3 张表

①WorldCups.csv

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTeams	MatchesPlayed	Attendance
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	13	18	590.549
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	16	17	363.000

该表包含了至 2014 年为止 20 届世界杯的年份、举办国家、四强队伍、总进球数、总球队数、总比赛数、以及参与人数

②WorldCupMatches.csv

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Attendance	Half-time Home Goals	Half-time Away Goals	Referee	Assistant 1	Assistant 2	RoundID	MatchID	Home Team Initials
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico		4444.0	3.0	0.0	LOMBARDI Domingo (URU)	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)	201.0	1096.0	FRA
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium		18346.0	2.0	0.0	MACIAS Jose (ARG)	MATEUCCI Francisco (URU)	WARNKEN Alberto (CHI)	201.0	1090.0	USA

该表统计了每一场比赛的场次、时间、阶段、比赛场地、城市、主客队名称和
碎屑、得分数、是否加时、观众数、半场主客队得分、裁判（主裁和助理裁
判）

总计 852 条数据

③WorldCupPlayers.csv

RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
0	201	1096	FRA CAUDRON Raoul (FRA)	S	0	Alex THEPOT	GK	NaN
1	201	1096	MEX LUQUE Juan (MEX)	S	0	Oscar BONFIGLIO	GK	NaN

该表包含了球员的队伍简称、教练名字、球员号码、球员名字、是否首发、球
员位置（GK 表示 goaltender 守门员，C 表示 captain 队长）、以及一些赛场事件
总计 255959 条数据

三 可视化工具

我选择的可视化工具为 Python 和 Tableau，选择该工具主要有以下原因：

1. 使用 Python 进行可视化的原因主要是本人最近在跟随实验室老师学习
机器学习相关知识，刚好可以利用 python 知识来进行数据处理和画
图，也方便我掌握和学习 python
2. Python 的 matplotlib 库的代码写起来和 Matlab 极为相似，在实验三四五
Matlab 实验的基础上可以方便我快速进行上手学习
3. 选择 tableau 的原因主要是这个工具可以快捷的拖拽实现可视化，并且
与数据源进行实时连接，且最重要的是在地图的画图方面可以自动处理
数据得到经纬度，简洁美观。

四 数据预处理

4.1 空值数据清洗

由于 python 读取 csv 文件会自动将空值表示为 NaN，因此需要进行数据清
洗，将所有 NaN 值去除，具体方法为使用 dropna()方法

4.2 国家名统一

由于历史原因，德国不同时期的名称不同，为方便后续的分析，固进
行国家名统一，具体方法为 replace('Germany FR', 'Germany')，将德国名称
统一为 Germany

4.3 特殊数值处理

分析发现在 WorldCups.csv 的 Attendance 属性的表示形式为国际常用的三位分节法：三位一个阶然后进行分割，利于阅读但不利于进行数据分析。

因此将小数点去除并转化为整数，具体方法为 world_cups['Attendance'] = world_cups['Attendance'].str.replace('.', '').astype('int64')

清洗后的数据如下所示：

WorldCups.csv

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTeams	MatchesPlayed	Attendance
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	13	18	590549
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	16	17	363000

WorldCupPlayers.csv

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
35	201	1090	USA	MILLAR Bob (USA)	S	0	Tom FLORIE	C	G45'
74	201	1093	BRA	DE CARVALHO Pindaro (BRA)	S	0	PREGUINHO	C	G62'

WorldCupMatches.csv

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Attendance	Half-time Home Goals	Half-time Away Goals	Referee	Assistant 1	Assistant 2	RoundID	MatchID	Home Team Initials
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico		4444.0	3.0	0.0	LOMBARDI Domingo (URU)	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)	201.0	1096.0	FRA
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium		18346.0	2.0	0.0	MACIAS Jose (ARG)	MATEUCCI Francisco (URU)	WARNKEN Alberto (CHI)	201.0	1090.0	USA

五 可视化方案以及部分可视化过程

达到什么目的、采用什么可视化手段、预期达到什么效果

5.1 可视化方案

1. 首先分别对三个表（概要表、比赛表、选手表）进行相关性分析，预先通过热力图矩阵分析变量之间的相关性，方便选取强相关的变量进行后续分析。
2. 展开对于 WorldCups.csv 表的分析，首先分析累计获得冠军数最多的几个国家、其次分析获得过前四强的国家，同时，利用处理过程的中间结果暂存为获奖国家.csv 文件，后续使用 Tableau 进行地图层面的可视化分析。
3. 其次展开时间维度数据统计分析，分别针对观众数、参赛队伍、比赛数目进行分析得到相应的树形热力图以及双折线图。同时为了验证进球数与参赛队伍数目的正相关关系，还绘制了相应的散点图进行分析。

- 随后，我聚焦在对于 WorldCupPlayers.csv 表的分析，主要分析了主场效应的影响。首先，统计每一届所有比赛的主场得分以及客场得分绘制堆积柱形图进行得分对比分析，其次，针对主场队伍的比赛结果进行分析，计算发现，主场球队获胜的概率为 57%，平局的概率为 22%。之后，为了深入分析主客场球队得分，分别的主客场球队的得分分布绘制相应的小提琴图进一部分展开分析。
- 接着，我希望在进行可视化分析的同时可以加深对著名战役经典赛事等“冷知识”的了解，因此我绘制了观看人数前 10 的比赛，了解了著名的马拉卡纳惨案，绘制了观看人数最多的前 10 个比赛场地，了解了巴西著名的 Jornalista Má rio Filho 体育场，统计每场比赛的双方得分，绘制了最大分差的前 10 场比赛。
- 最后，除了普通的图表式可视化外，我还希望更加直观的展示出参加世界杯最多次数的国家、球员以及教练，因此使用了不同的背景模型（大力神杯、足球）绘制了词云图。然后还使用了 Tableau 绘制了地图，展示了夺冠及四强队伍的地理分布位置。

5.2 部分可视化操作

数据导入与预处理

2 读取数据

```
In [3]: world_cups = pd.read_csv('./data/WorldCups.csv')
world_cup_player = pd.read_csv('./data/WorldCupPlayers.csv')
world_cups_matches = pd.read_csv('./data/WorldCupMatches.csv')
```

```
In [4]: world_cups.head(2)
```

```
Out[4]:
```

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTeams	MatchesPlayed	Attendance
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	13	18	590549
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	16	17	363000

```
In [5]: world_cup_player.head(2)
```

```
Out[5]:
```

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
0	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Alex THEPOT	GK	NaN
1	201	1096	MEX	LUQUE Juan (MEX)	S	0	Oscar BONFIGLIO	GK	NaN

```
In [6]: world_cups_matches.head(2)
```

```
Out[6]:
```

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Attendance	Half-time Home Goals	Half-time Away Goals	Referee	Assistant 1	Assistant 2	R
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico		4444.0	3.0	0.0	LOMBARDI Domingo (URU)	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)	
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium		18346.0	2.0	0.0	MACIAS Jose (ARG)	MATEUCCI Francisco (URU)	WARNKEN Alberto (CHI)	

3 数据预处理

- 空值处理
- 国家名统一
- 字段类型转换

```
In [7]: world_cup_player = world_cup_player.dropna()
world_cups = world_cups.dropna()
world_cups_matches = world_cups_matches.dropna()
```

```
In [8]: world_cups = world_cups.replace('Germany FR', 'Germany')
world_cup_player = world_cup_player.replace('Germany FR', 'Germany')
world_cups_matches = world_cups_matches.replace('Germany FR', 'Germany')
```

```
In [9]: world_cups['Attendance'] = world_cups['Attendance'].str.replace('.', '').astype('int64')
```

绘制柱形图

5 统计获得过前四名的队伍

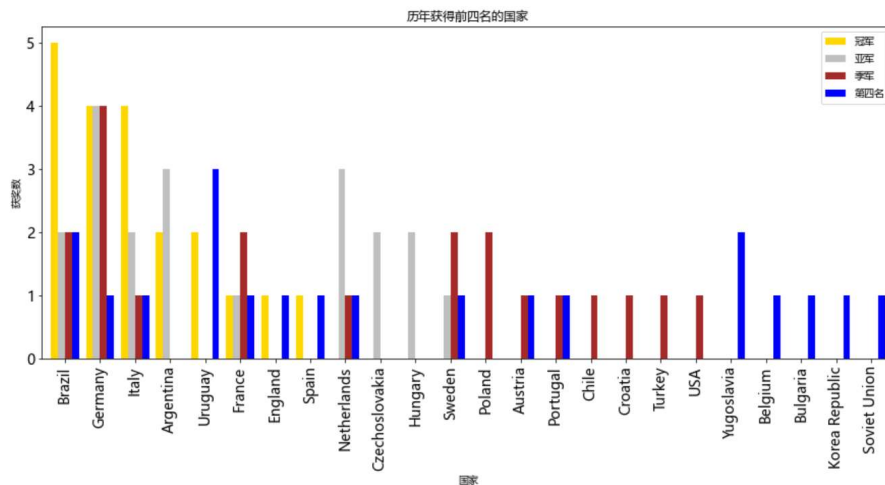
```
In [16]: # 只对world_cups表进行分析
plt.rc("font", family='Microsoft YaHei')
gold = world_cups["Winner"]
silver = world_cups["Runners-Up"]
bronze = world_cups["Third"]
fourth = world_cups["Fourth"]

gold_count = pd.DataFrame.from_dict(gold.value_counts())
silver_count = pd.DataFrame.from_dict(silver.value_counts())
bronze_count = pd.DataFrame.from_dict(bronze.value_counts())
fourth_count = pd.DataFrame.from_dict(fourth.value_counts())

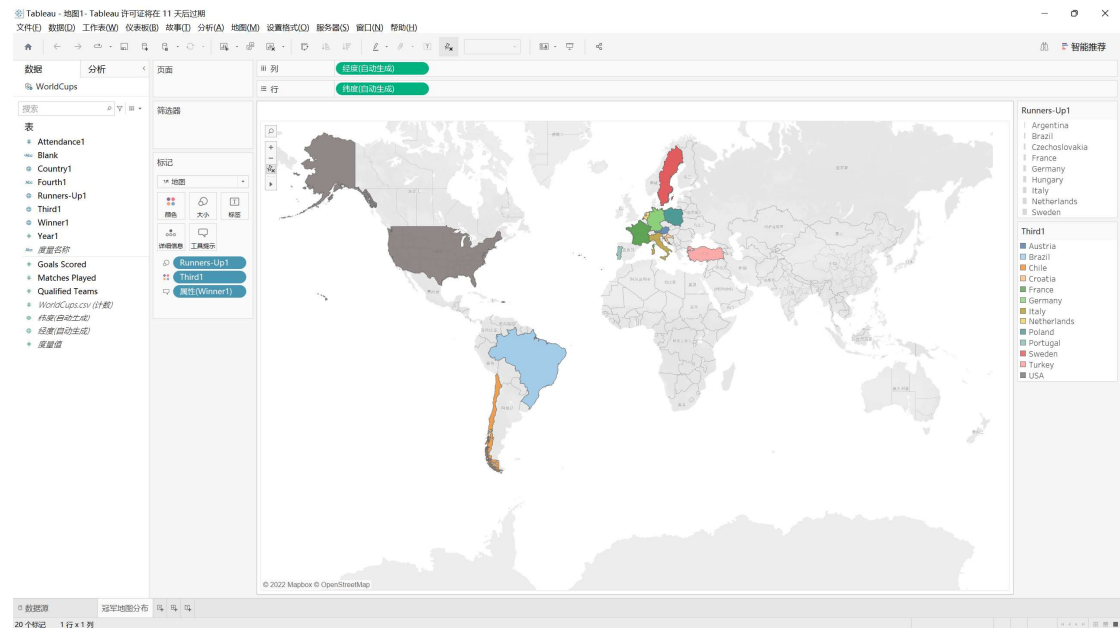
total_count = gold_count.join(silver_count, how='outer').join(bronze_count, how='outer').join(fourth_count, how='outer')
total_count = total_count.fillna(0)
total_count.columns = ['冠军', '亚军', '季军', '第四名']
total_count = total_count.astype('int64')
total_count = total_count.sort_values(by=['冠军', '亚军', '季军', '第四名'], ascending=False)

total_count.plot(y=['冠军', '亚军', '季军', '第四名'], kind="bar",
                 color=['gold', 'silver', 'brown', 'blue'], figsize=(15, 6), fontsize=14,
                 width=0.8, align='center')
plt.xlabel('国家')
plt.ylabel('获奖数')
plt.savefig('柱形图')
plt.title('历年获得前四名的国家')
```

Out[16]: Text(0.5, 1.0, '历年获得前四名的国家')



使用 Tableau 绘制地图

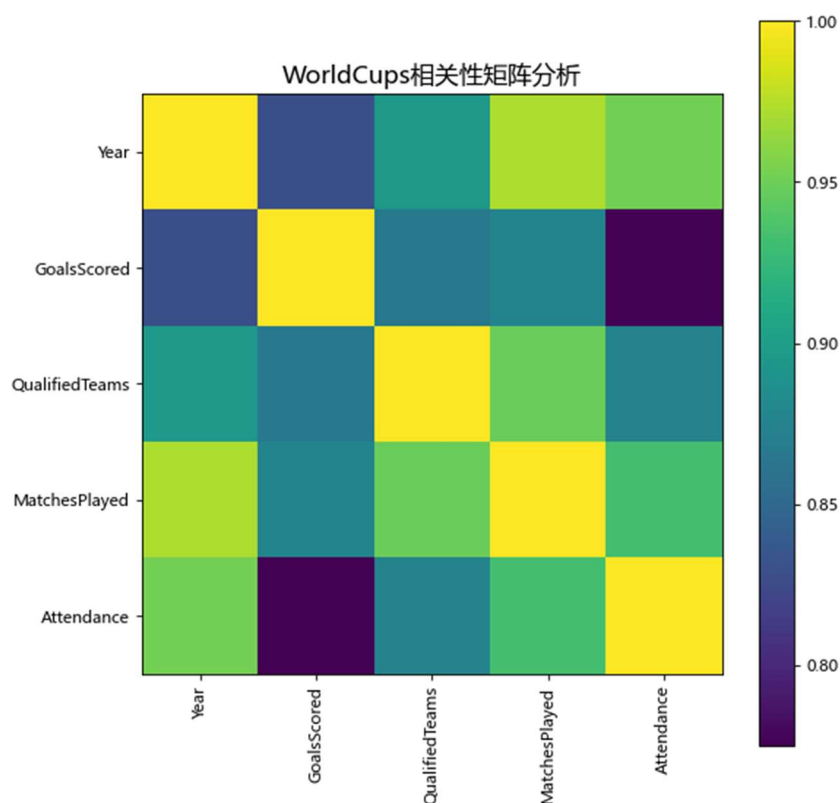


先连接到 WorldCups.csv 文件后，将想要显示颜色的数据拖入颜色标记即可进行显示，并且可以添加“说明及摘要”等辅助信息

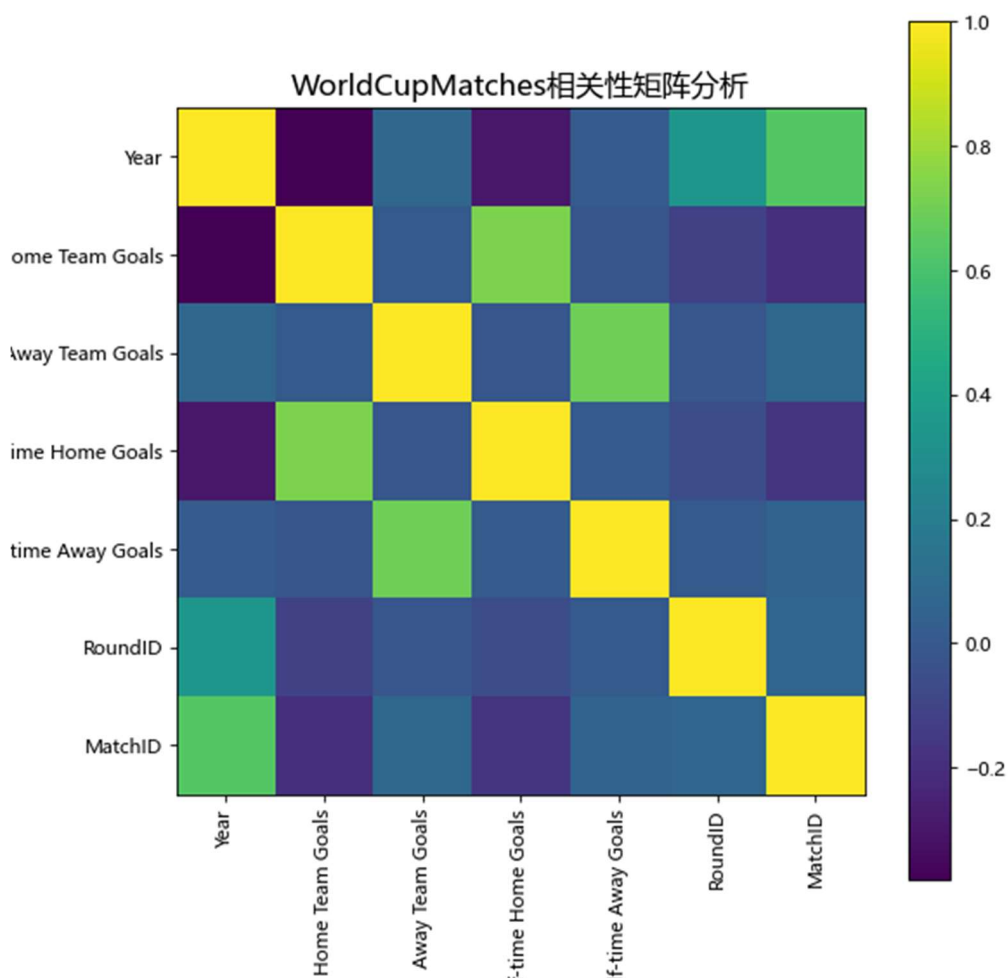
六 可视化结果

6.1 WorldCups 和 WorldCupMatches 热力图相关性分析

相关分析是对变量两两之间的相关程度进行分析。使用相关性分析可以帮助我们进一步可视化的变量选取有较大帮助



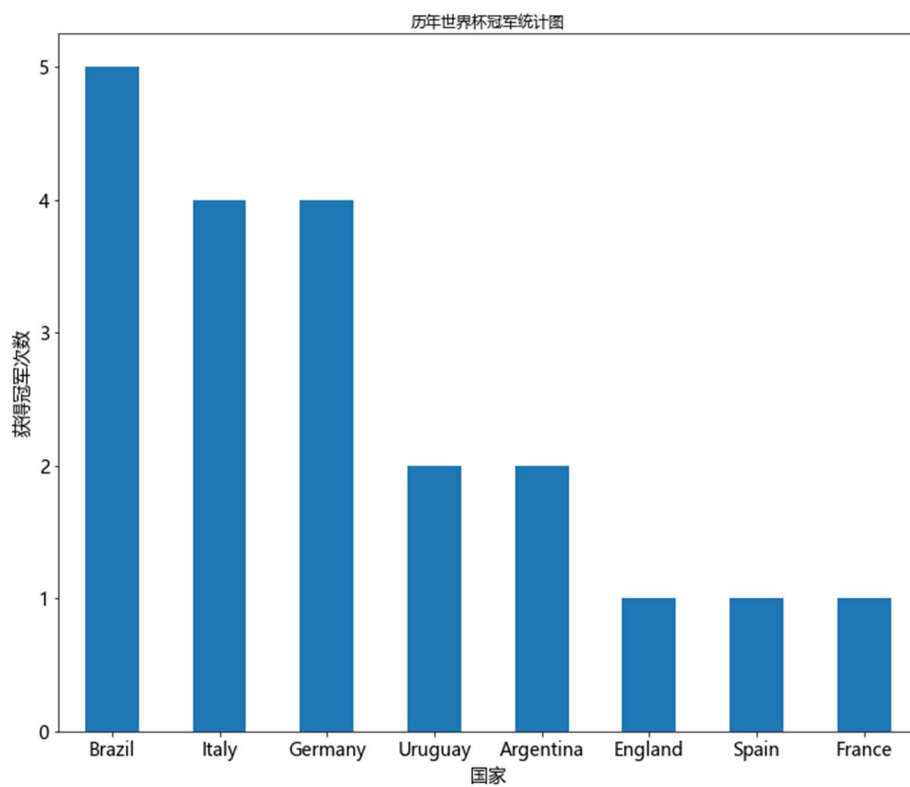
分析：可以看到年份与所有其他变量的皮尔森相关系数 P 值都是在 0.8 以上的，因此本文都对变量两两之间进行了可视化分析，此外可以看到 MatchesPlayed 与 QualifiedTeams 也是强相关的。



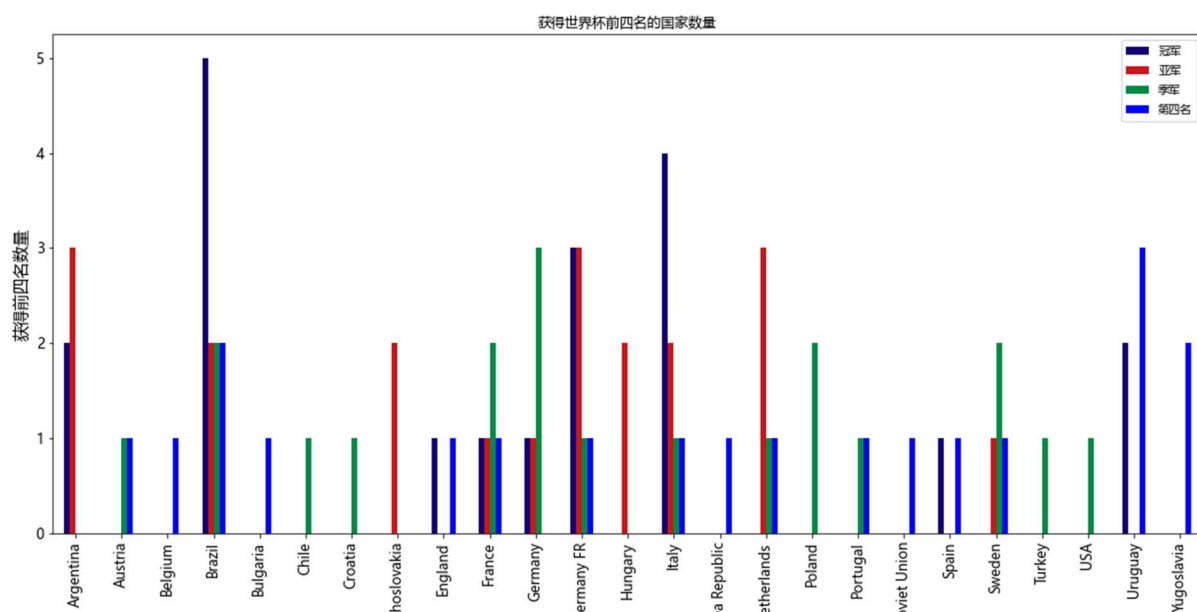
分析：可以看到主客队的得分情况以及半场得分情况都是绿颜色，强相关的，因此值得进一部分分析，而 RoundID 于 MatchID 虽然也呈现强相关，这主要是由于认为设定每一年的 ID 值格式一致导致，因此舍弃不进行进一步分析。

6.2 历届冠军及四强队伍分析

横轴为国家名称，纵轴为获得冠军的次数，从左往右按照获得次数递减排序。



分析：可以看到直到 2014 年仅有 8 个国家获得过世界杯的冠军，其中巴西、意大利以及德国都是常胜将军，分别获得过 5 次、4 次、4 次的好成绩。



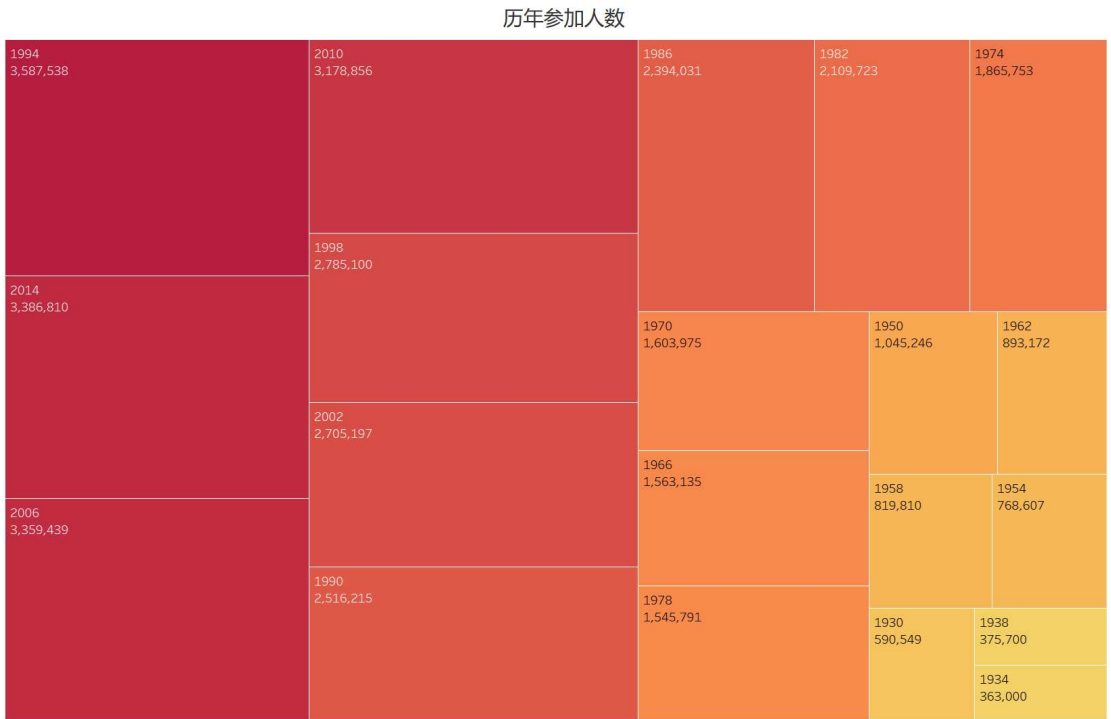
横轴表示国家，纵轴表示是否获得过前四名以及分别获得的数目

分析：历史上曾有 25 支球队闯进过四强，其中可以直观的看到巴西队在前四名奖项上都有不俗的成绩，此外捷克斯洛伐克以及匈牙利都仅仅只获得过一次亚军的荣誉，经查询资料发现，捷克斯洛伐克球队没有延续辉煌的原因是国家的战争导致其 1992 年就解散，球队不复存在。

6.3 各届世界杯参与人数、参赛队伍、比赛数目分析

参赛人数树形热力图分析

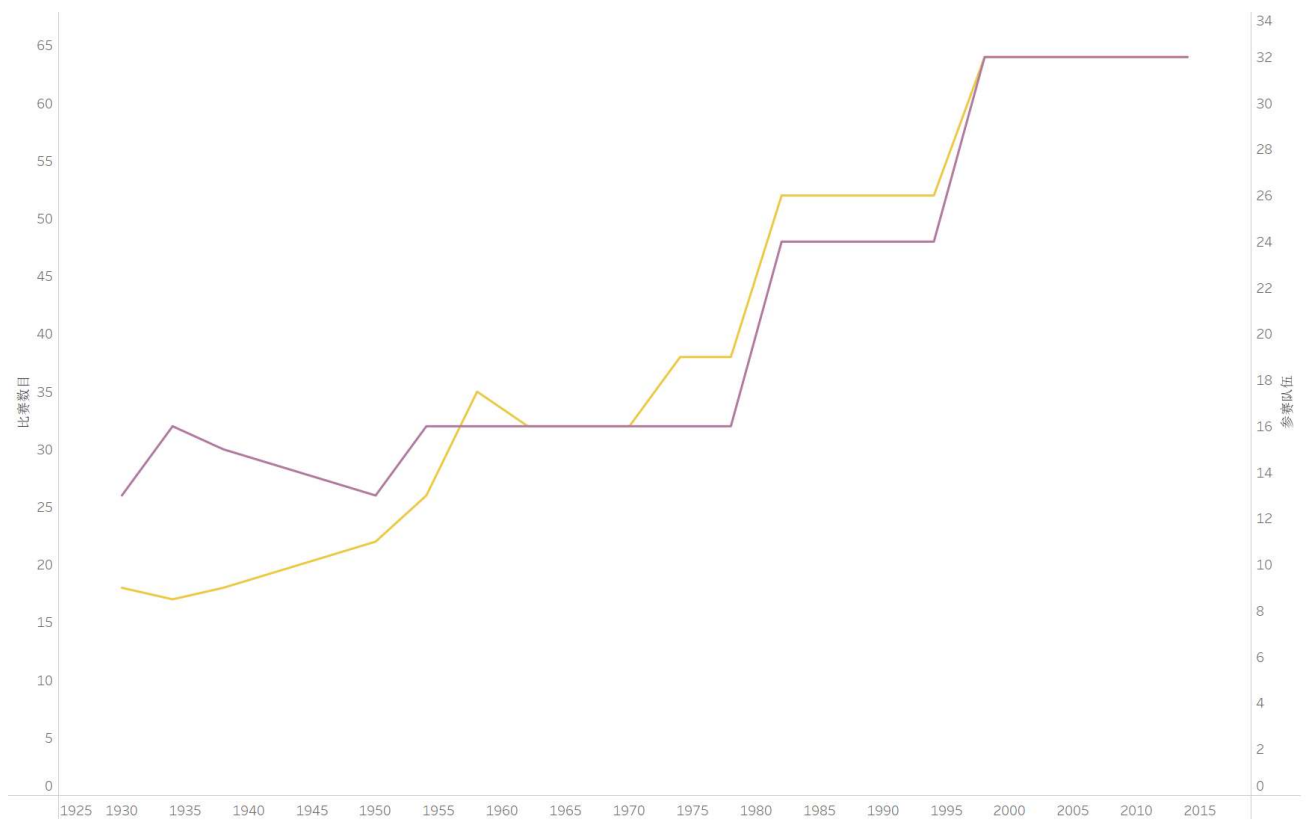
使用树状热力图能直观用方块的颜色，大小排序展示每一届世界杯的总参与人数的相对大小。



分析：可以直观的看到排名前四的参与人数中 2006 2010 2014 都有很高的参加人数，侧面反映出随着世界经济的发展，人们有了更多的经济实力来支持参与到世界杯活动中去。

比赛队伍、比赛数目折线图分析

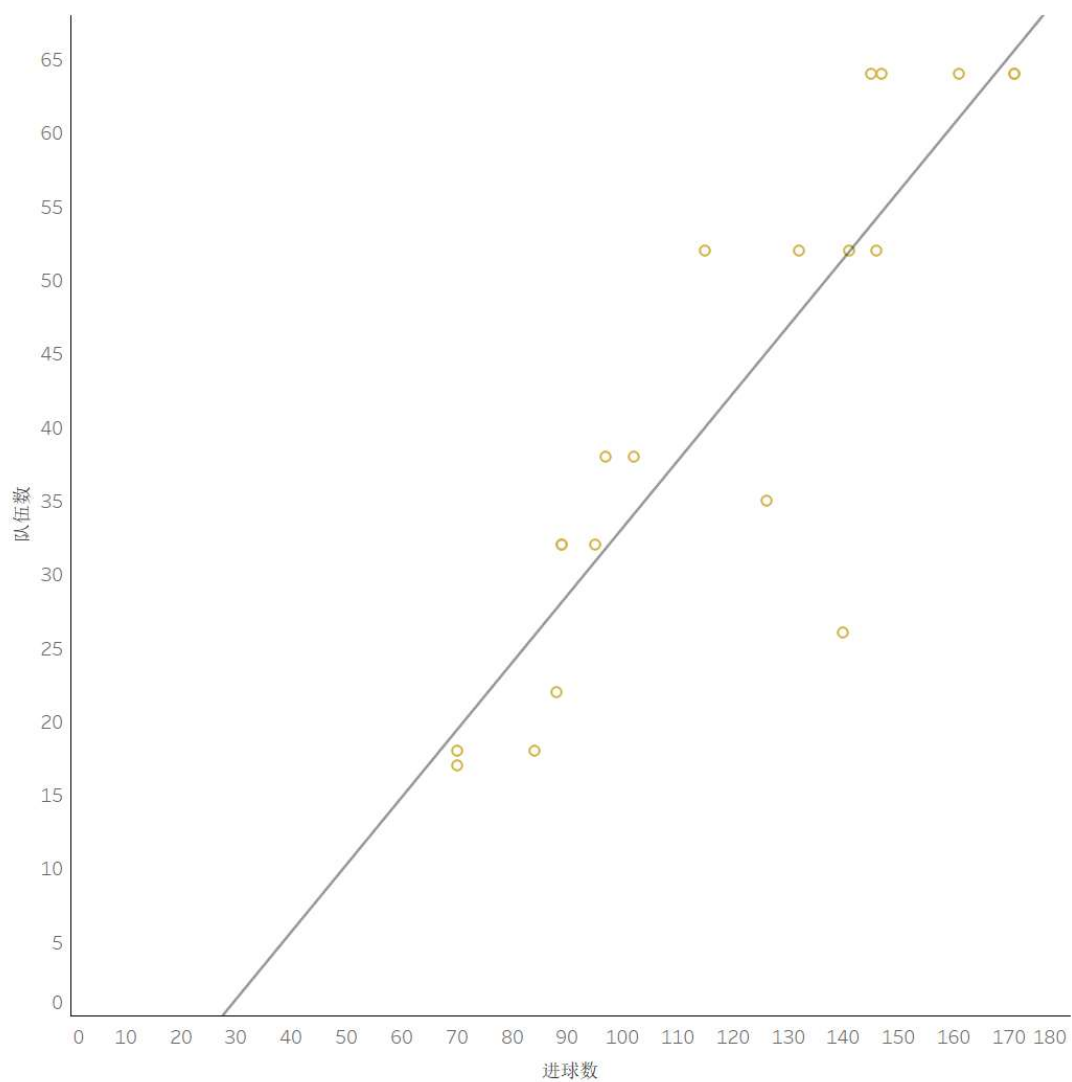
横轴为年份，纵轴表示每一年的参赛队伍数以及进行的比赛数目



分析：通过折线图走势可以看到，随着年份的增加，参与世界杯的队伍数目从原来的 13 到近期的 32 只队伍，比赛的数目也因此不断的增加，全世界七大洲的不同国家都力争参与到这场盛大的赛事中。

6.4 比赛进球数队伍数散点图

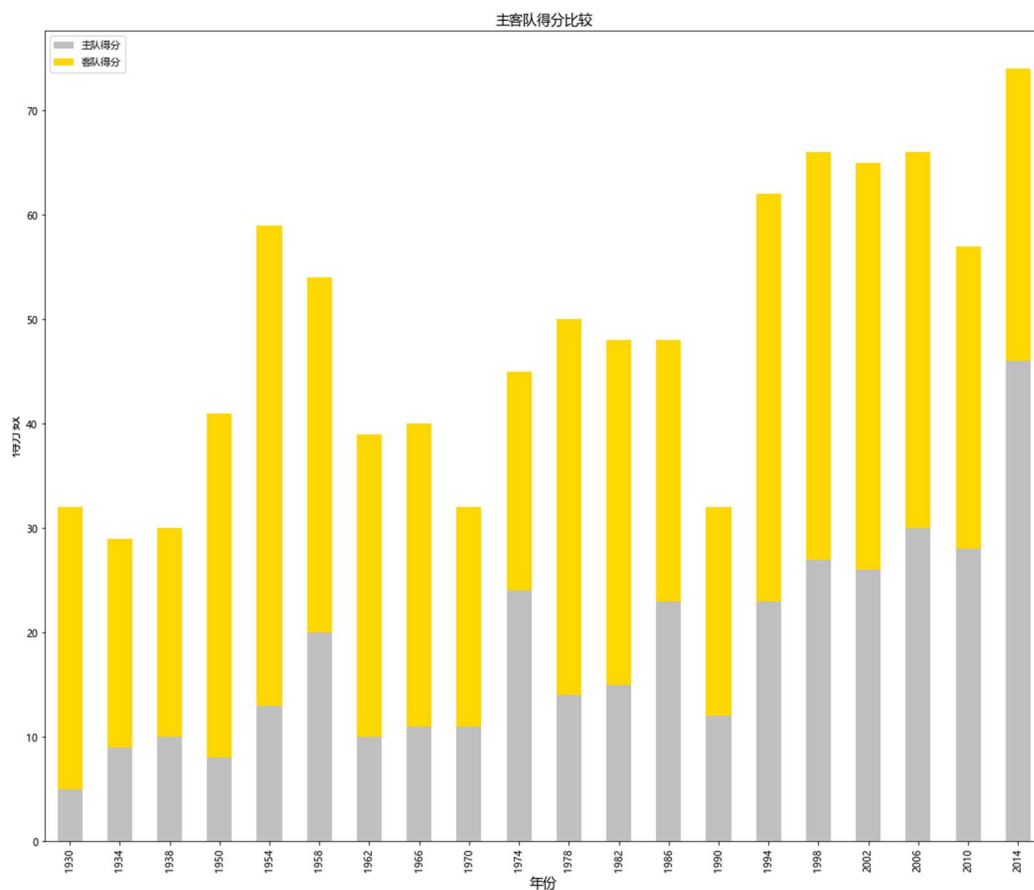
横轴表示进球数、纵轴表示队伍数



分析：由趋势线可以明显看出随着进球数的递增，队伍数表现出明显的正相关。

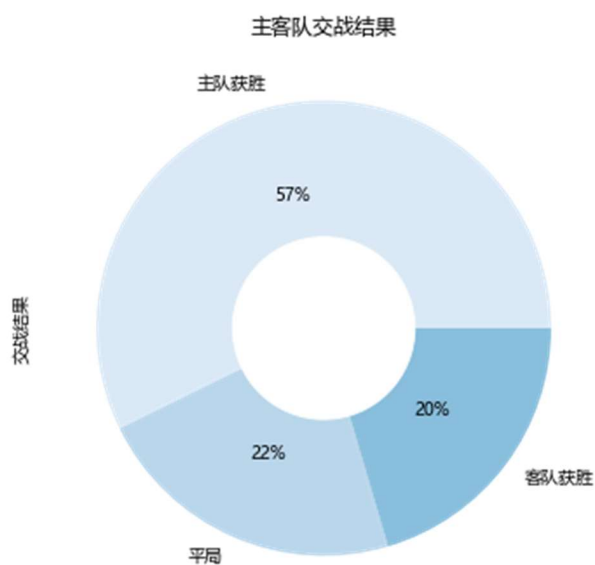
6.5 主客队得分对比及结果比较

堆积柱形图可以使人们一眼看出各个数据的大小，易于比较数据之间的差别。统计每一届比赛主客场得分加总后进行比较如下：



分析：可以直观的看到主场得分基本是客场得分的一倍以上，这就是体育领域经常提及的主场优势，因为球队在主场驻地的体育馆与外来队伍比赛，主场球队赛程会较松散，比客队更熟悉当地气候等，同时场馆大多为自家球迷，因此可以获得更多的加油助威

饼图可以准确反映不同数据之间的比例关系和相对大小关系



进一步的，分析 WorldCupMatches 表中的进行主客队交战结果，可以验证前文的分析，在所有的交战中，主队获得胜利高达 57%，获得平均为 22%，而只有 20%的可能输球，可见主场优势对于一个球队的重要性。

6.6 主客场得分分布分析

小提琴图结合了箱线图以及核密度图的优势，可以直观的看到数据的基本属性以及整体分布情况。

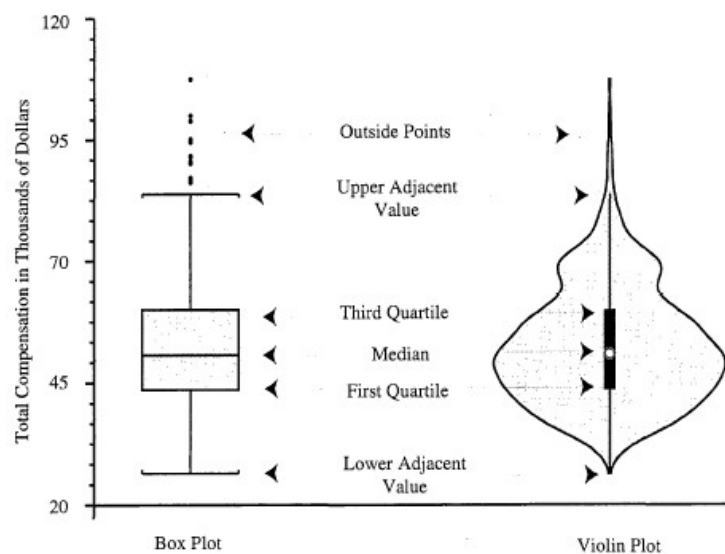
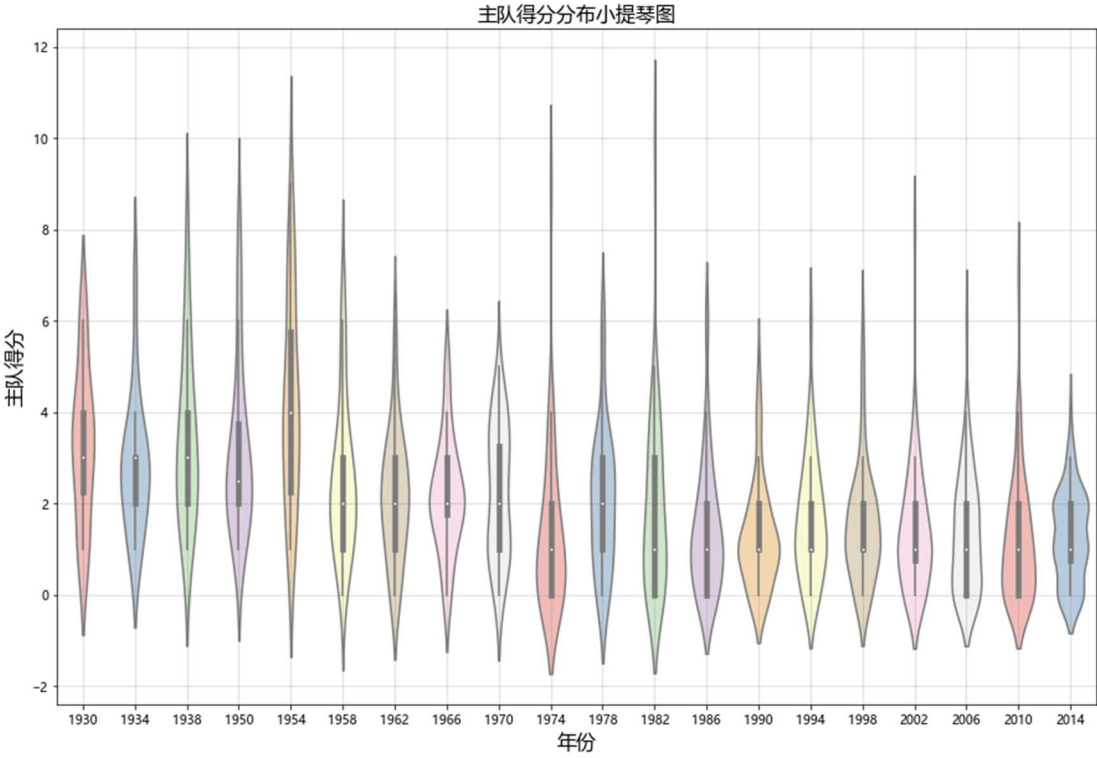


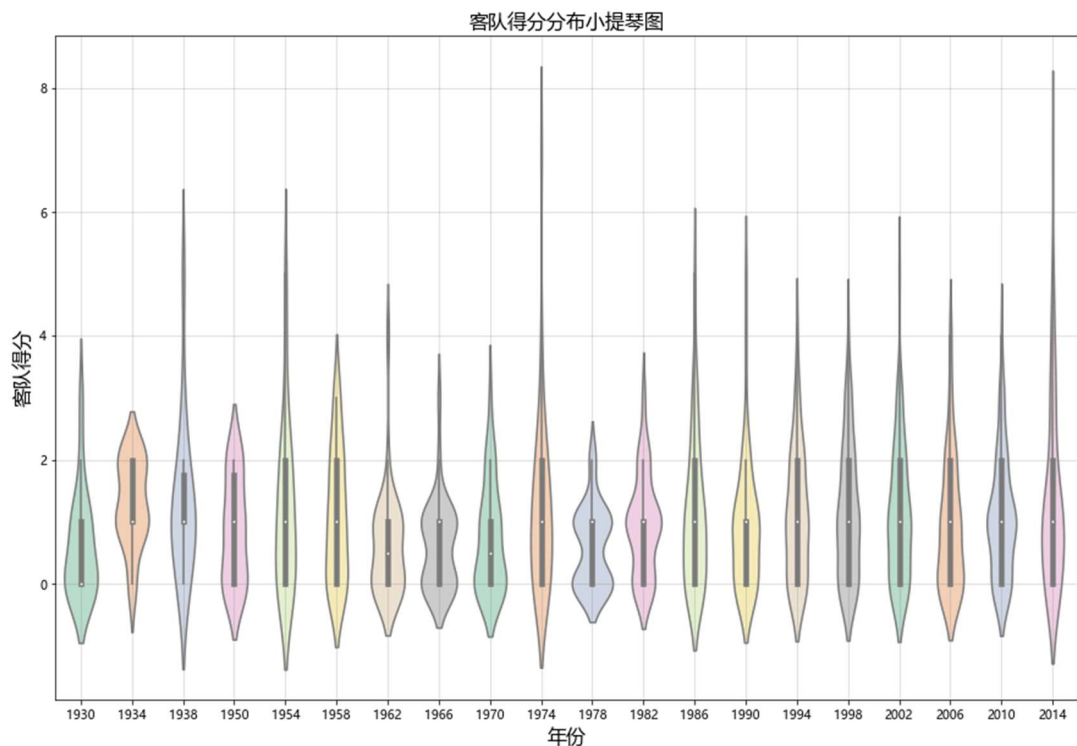
Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

上图是将箱线图与小提琴图进行对比，可以看到小提琴图不仅可以显示中位数、四分位数范围、较高、较低的相邻值以外，还可以直观的显示数据的整体数据分布范围

参考链接：<https://zhuanlan.zhihu.com/p/376055263>

以下是我基于 Python 的 seaborn 库画的主客场得分小提琴图。





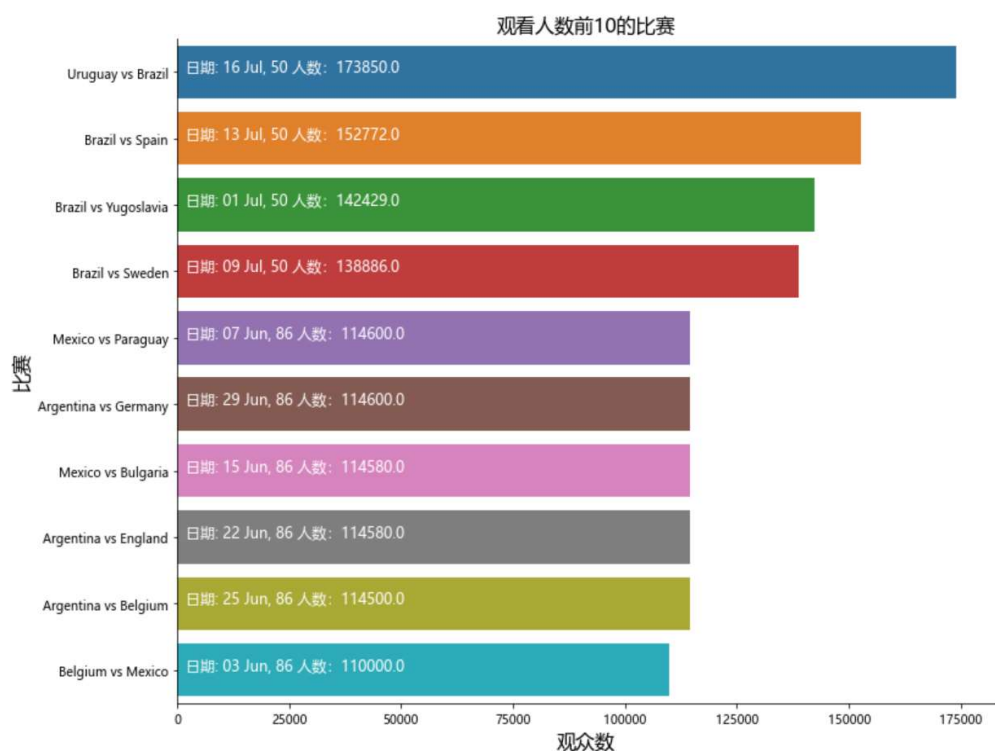
分析：小提琴图中较宽的部分代表观测值取值的概率较高，较窄的部分则对应于较低的概率。可以看到主队客队作战得分的中位数基本都为 1 分到 2 分之间，客队的小提琴分布相较于主队而言比较扁，即客队得分比较不稳定，且主队有更多可能球队得分高达 8 颗以上，客队基本无法达到这么高的数量。

6.7 著名比赛及体育场分析

条形图能够使人们一眼看出各个数据的大小，并且如果按照数据大小进行排序的话可以使人们直接清楚看到排名前几的数据。

观看人数前 10 的比赛

横轴为观众数，纵轴为比赛的双方，在条形图上添加日期以及人数的标签。

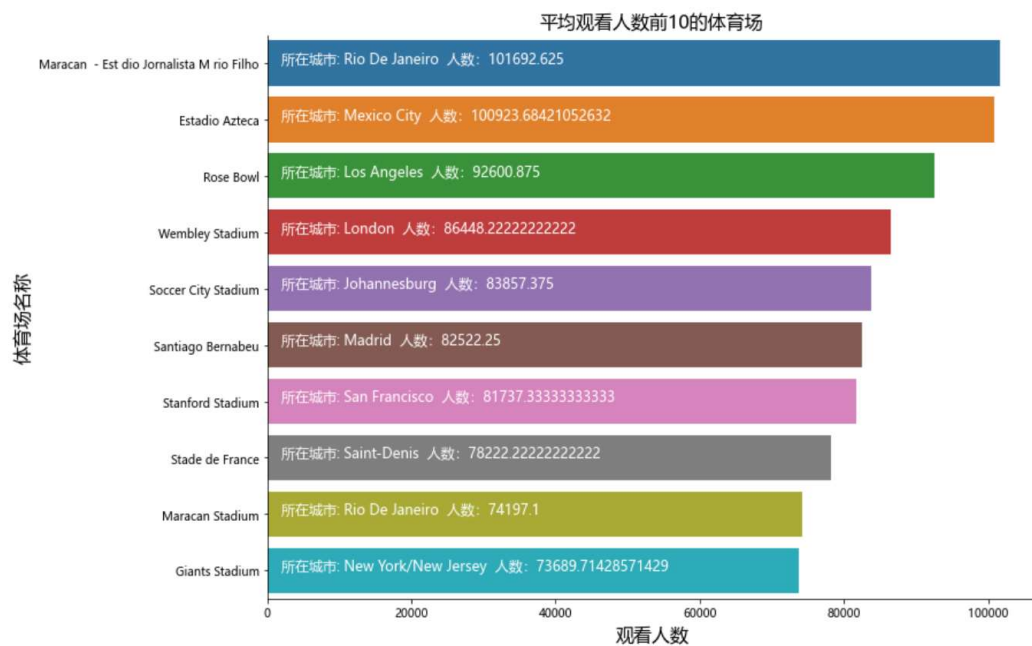


分析：可以直观的看到在观看人数前 10 的比赛中，前 4 场比赛都为巴西队与不同国家的比赛，可以直接反应巴西队的支持热度以及观众的期待指数。查阅资料发现，1950 年为巴西主办的世界杯并且是二战后的第一个世界杯，因此不难解释人们满心欢喜战争结束投身体育娱乐事业的热情。

此外值得一提的是，观看人数第一的比赛：1950 年乌拉圭对阵巴西，这场比赛被称为马拉卡纳惨案，这场比赛，巴西队在打平即可夺冠的情况下却以 1 比 2 负于对手，错失冠军。比赛结束后，有数名巴西球迷因承受不了球队失冠的压力而自杀、猝死。

比赛场地条形图

横轴为观众数，纵轴为比赛的场地，在条形图上添加所在城市以及平均人数的标签。

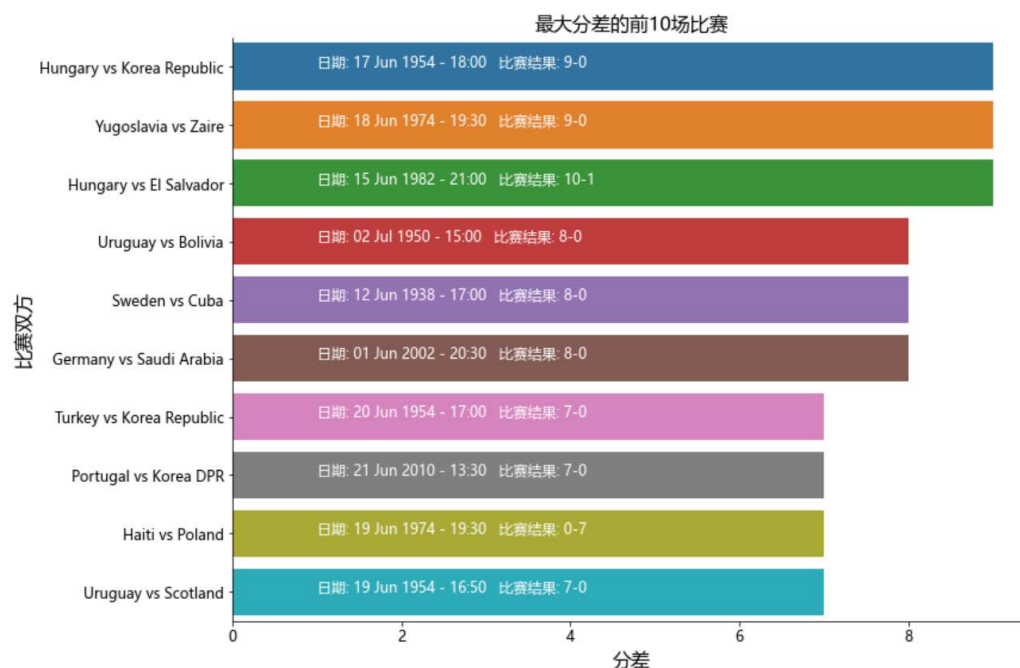


分析：排名第一的体育场为来自巴西里约热内卢的体育场，经过查阅资料可以发现 Jornalista Má rio Filho 又名马拉卡纳（Maracanã）创造了参加足球比赛的世界纪录。根据国际足联的说法，2014 年 7 月 16 日，近 20 万人去了马拉卡纳,观众总数为 199.854，这是有史以来参加足球比赛的最大人群！

参考链接：<https://www.itinari.com/zh/the-country-of-football-maracana-stadium-in-rio-de-janeiro-gzug>

比赛分差条形图

横轴为分差数，纵轴为比赛的双方，在条形图上添加比赛日期以及比赛的结果。



分析：分析最大分差可以帮助我们了解历史上的“耻辱战”，看出交战双方的悬殊实力，查阅得知，1982年匈牙利对战萨尔瓦多，这是世界杯史上少有的决赛阶段还出现悬殊分差的比赛，国内正在打仗的萨尔瓦多竟然以10比1输掉了比赛。

参考链接：<https://zhidao.baidu.com/question/174560478.html>

6.8 著名足球强国、球员、教练分析

使用开源的 wordCloud 库来进行表示，可以方便的编辑词云的形状背景以及颜色

胜利词云图

颜色编码国家类型，大小编码获得胜利的次数

利用大力神杯的图为模型图，将获得过胜利的国家进行表示，可以看到巴西、德国和意大利是最多的



球员词云图

颜色编码球员类型，大小编码参与次数

利用足球的图为模型图，将参与过世界杯的球员名称进行表示



教练词云图

颜色编码教练类型，大小编码参与次数

将参与过世界杯的球员的教练用词云进行表示

可以看到 雷斯·菲腊比·史高拉利（Luiz Felipe Scolari）、佩雷拉（Carlos A. Parreira）这些传奇教练的存在，查阅维基百科后可以验证结果

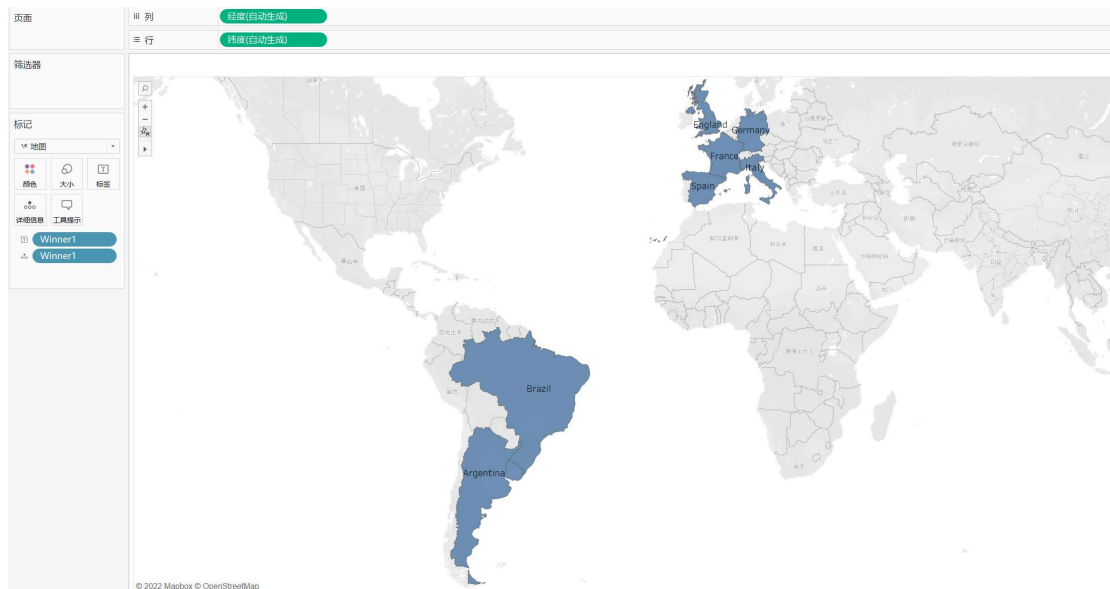


- 粗体字的教练名称代表仍然活跃的教练。
- 目前共有十四位教练曾经参与过三届世界杯或以上，包括十一位来自欧洲国家及三位来自巴西。
- 博拉·米卢蒂诺维奇是历史上参与过最多届世界杯次数而没有夺得冠军的教练。
- 马里奥·扎加洛除了在1970年、1974年及1998年以教练身份带领巴西，还在1994年以助教身份协助巴西夺冠。

教练名称	届数	国籍	参与届数
佩雷拉 (Carlos A. Parreira)	6届	巴西	1982, 1990, 1994 , 1998, 2006, 2010
博拉·米卢蒂诺维奇 (Bora Milutinovic)	5届	南斯拉夫	1986, 1990, 1994, {1998, 2002
奥斯卡·塔巴雷斯 (Óscar Tabárez)	4届	乌拉圭	1990, 2010, 2014, 2018
赫尔穆特·舍恩 (Helmut Schön)	4届	西德	1966, 1970, 1974 , 1978
温特伯顿 (Walter Winterbottom)	4届	英格兰	1950, 1954, 1958, 1962
塞普·赫爾貝格 (Sepp Herberger)	4届	德国 / 西德	1938, 1954 , 1958, 1962
拉约斯·巴洛堤 (Lajos Baroti)	4届	匈牙利	1958, 1962, 1966, 1978
亨利·米歇尔 (Henri Michel)	4届	法国	1986, 1994, 1998, 2006
安素·比亚索 (Enzo Bearzot)	3届	义大利	1978, 1982 , 1986
马里奥·扎加洛 (Mario Zagallo)	3届	巴西	1970 , 1974, 1998
居伊·泰斯 (Guy Thys)	3届	比利时	1982, 1986, 1990
加夫里尔·卡查林 (Gavril Kachalin)	3届	苏联	1958, 1962, 1970
雷斯·菲腊比·史高拉利 (Luiz Felipe Scolari)	3届	巴西	2002 , 2006, 2014
胡斯·轩迪克 (Guus Hiddink)	3届	荷兰	1998, 2002, 2006
卡尔·列宾 (Karl Rappan)	3届	奥地利	1938, 1954, 1962

6.9 历史获奖队伍地理可视化分析

使用 Tableau 工具进行地图数据便捷显示，可以看到可以看到历史上的冠军基本上全部在南美以及西欧



接着，使用 python 处理得到获奖过的国家后再使用 Tableau 工具进行可视化



分析：从地理位置来看，明显可以发现足球强国聚集在足球氛围浓烈的南非以及俱乐部举办的热火朝天的西欧，这些国家的球员从一系列甲级乙级联赛重重磨练脱颖而出会师大力神杯。遗憾的是亚洲仅仅前苏联和“采用特殊手段”的韩国进过四强，没有看到中国国家队的踪影...

七 总结与体会

对于世界杯往届历史数据的分析是帮助我们回望历史，了解曾经父辈甚至祖辈年代的足球历史，在进行分析的时候对于我来说最大的困难就是对于足球知识的缺乏，这也是驱动我进行本次数据可视化分析的主要原因。

经过本次可视化分析我成功使用了**相关性分析热力图**以及**散点图**进行变量之间潜在关系的分析。还使用了**柱形图**、**词云图**、以及**世界地图**对历史上获得冠军的国家的次数和频数以及进入前四强的国家的次数频数以及地理位置进行了深度分析。此外，还学会了使用**词云图**来进行大数据分析，分析历史上参加次数最多的国家、球员、教练，帮助我了解史上的辉煌球员和传奇主教；此外，还利用**树状热力图**直观分析了历史上每一届的参加人数，直观反映世界杯热度的与日俱增。随后，为了对“主场优势”效应加以分析，文章使用了**堆积条形图**分析主客队得分对比，**饼图**分析主队作战结果，**小提琴图**分析主客队得分分布情况。最后，为了对历史上比较著名的悬殊战役、著名的体育场、著名的比赛进行分析，分别使用**条形图**分析各数据前 10 名。

本次数据可视化作业对于我的**最大收获与体会**在于：

1. 对于世界杯的历史文化以及著名人物球队国家有了更深的理解
2. 学习了包括热力图、柱形图、树状热力图、词云图、堆积柱形图、条形图、饼图、小提琴图、地图的制作
3. 掌握了 python 画图的基本方法
4. 掌握了 Tableau 制作地图地理数据的基本方法
5. 掌握了进行数据分析数据可视化的基本流程

最后，感谢一直热心向我介绍数据可视化相关知识的任课老师，如果没有他的耐心讲解与教导，我可能也不会有独立进行数据可视化的想法，也学不到这么多的可视化技巧与方法。