**2020春季学期**

**北京大学《AI引论》课程小组项目书面报告**

**NBA球员投篮数据分析**

**毛一琛 (1900012906), 李鸿泽 (1900013037), 孙凌宇 (1900013036)**

**Email:1900012906@pku.edu.cn,1900013037@pku.edu.cn,1900013036@pku.edu.cn**

**院系：信息科学技术学院**

**指导老师: 童云海**

**2020年5月**



**摘 要**

在本项目中，我们从Kaggle平台上获取分析所需的NBA比赛投篮数据。首先，对数据集进行数据预处理，使其满足完整性、一致性约束。对数据集进行相关性分析，做出相关性热力图，启发下一步的关联规则挖掘。数据分析共分为五个部分：运用数据可视化方法对主场优势效应进行验证，并且探究投篮距离、防守距离等因素对于球员投篮选择的影响。之后，我们分析了得分前三名球员的投篮选择与命中率。进一步地，我们选取投篮距离等特征，并自定义了命中率新的特征，运用决策树的方法对投篮命中率进行分析，同时进行优化以解决拟合率低和算法不稳定的问题。在附加题中，我们运用基于投篮距离特征的K-Means算法进行聚类，并引入带权向量距离，对球员能力进行分类。

**关键词：**投篮选择、决策树、聚类分析

**目 录**

[**1** 项目背景（宋体，小三） 4](#_Toc41429100)

[**2** 相关工作 4](#_Toc41429101)

[**3** 方法 4](#_Toc41429102)

[**3.1** 数据预处理 4](#_Toc41429103)

[**3.2** 相关性分析与相关性热力图 4](#_Toc41429104)

[**3.3** 数据可视化 4](#_Toc41429105)

[**3.4** 决策树 4](#_Toc41429106)

[**3.5** 聚类分析 4](#_Toc41429107)

[**4** 实验 4](#_Toc41429108)

[**4.1** 数据预处理 4](#_Toc41429109)

[**4.2** Task1 6](#_Toc41429110)

[**4.3** Task2 10](#_Toc41429111)

[**4.4** Task3 15](#_Toc41429112)

[**4.5** Task4 17](#_Toc41429113)

[**4.6** Task5 附加题 18](#_Toc41429114)

[**5** 结论 20](#_Toc41429115)

[小组分工 21](#_Toc41429116)

[致谢 21](#_Toc41429127)

[参考文献 21](#_Toc41429128)

# 项目背景（宋体，小三）

美国职业篮球联赛（National Basketball Association），简称美职篮（NBA），是由北美30支职业球队组成的男子职业篮球联盟，是美国四大职业体育联盟之一，也是世界上最受欢迎的篮球比赛之一。随着篮球技术的不断发展，数据分析在战术制定、球员训练甚至营销推广中的地位越来越重要，评价一个球员的标准早已全部数据化，每个球队也都配备了专门的数据分析部门。因此，在本文中，我们小组运用可视化、决策树、聚类分析等多种方式对NBA投篮数据进行分析。

# 相关工作

在NBA比赛中早已有许多关于球员和比赛的数据分析，比如每场比赛的得分、篮板、抢断等数据，用于描述球员能力的雷达图，以及主场优势效应的总结。本文的部分分析是对相关工作的验证与延伸。

# 方法

# 数据预处理

# 相关性分析与相关性热力图

# 数据可视化

# 决策树

# 聚类分析

# 实验

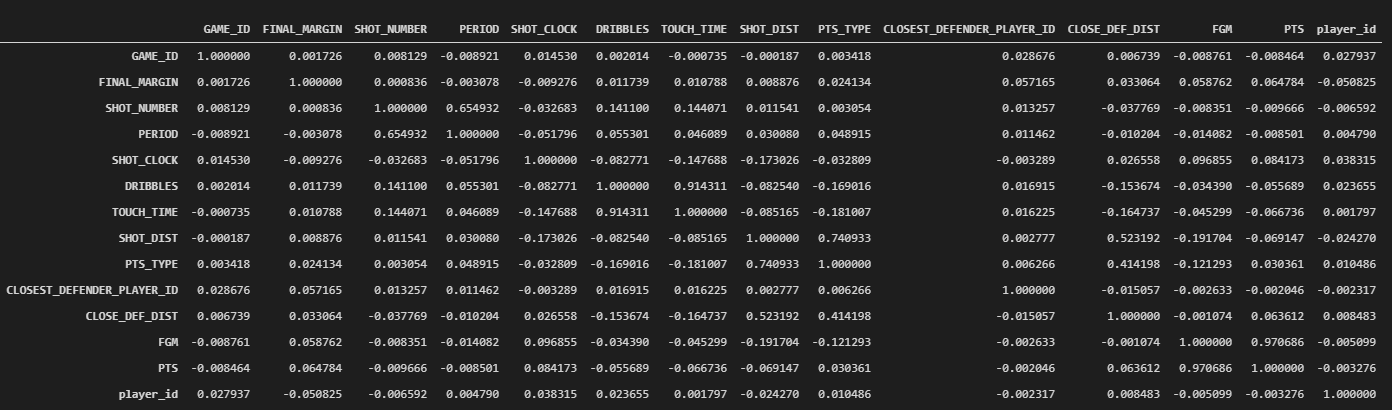
# 数据预处理

从Kaggle网站上获取的数据总体质量较高，满足完整性约束和唯一性约束。但是还有部分数据存在问题：

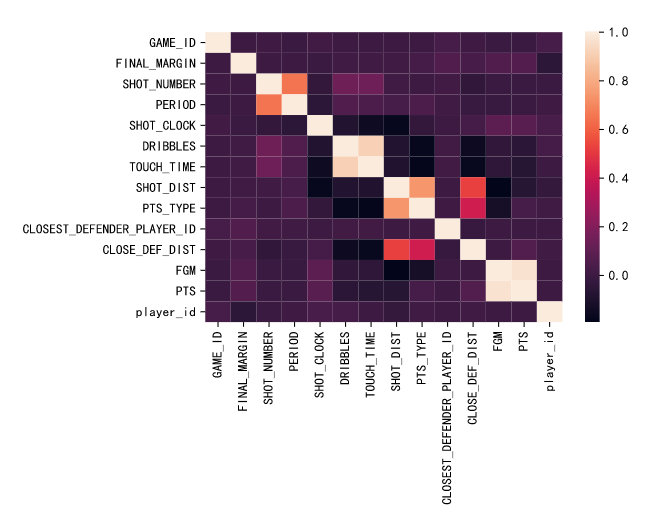
1. SHOT\_CLOCK项部分数据缺失。经观察，大部分缺失项对应的本节剩余时间GAME\_CLOCK小于等于24秒，根据比赛规则，此时不再记录进攻时间。其余情况可能是由犯规或其他原因造成的数据缺失。在不影响分析结果的情况下，我们将该项缺失数据用对应GAME\_CLOCK填充。其中在我们分析第一个问题时大量数据重复，我们在创建新的数据表的时候将其合并。

2、TOUCH\_TIME项部分数据出现异常值。经统计共有312条数据的TOUCH\_TIME为负数，根据定义为持球时间，不应该为负数。观察发现所有异常值对应的投篮前运球数DRIBBLES均为0，结合实际情况猜测应为在接球就投时数据记录出错，因此将所有异常值置为0。

3、数据存在冗余。例如player\_id与player\_name均可单独满足唯一性约束；得分PTS由PTS\_TYPE与SHOT\_RESULT决定。在分析中，将冗余项删除，以降低变量维度。

进一步地，对数据进行相关性分析，以指导进一步的探索方向。结果如下表：

为了更直观地观察各变量之间的相关性关系，做出相关性热力图如下：



在图中可以观察得到：PERIOD与SHOT\_NUMBER，TOUCH\_TIME与DRIBBLES，PTS\_TYPE与SHOT\_DIST，PTS与FGM之间具有较强的相关性，而通过这些变量之间的定义可以发现，这都是显然的（比如持球时间与投篮前运球个数呈正相关，是否是三分球由投篮距离直接决定），并没有太大的价值。而CLOSE\_DEF\_DIST与SHOT\_DIST，CLOSE\_DEF\_DIST与PTS\_TYPE之间具有不太显著的相关性，这启发我们再进一步的研究中重点关注。

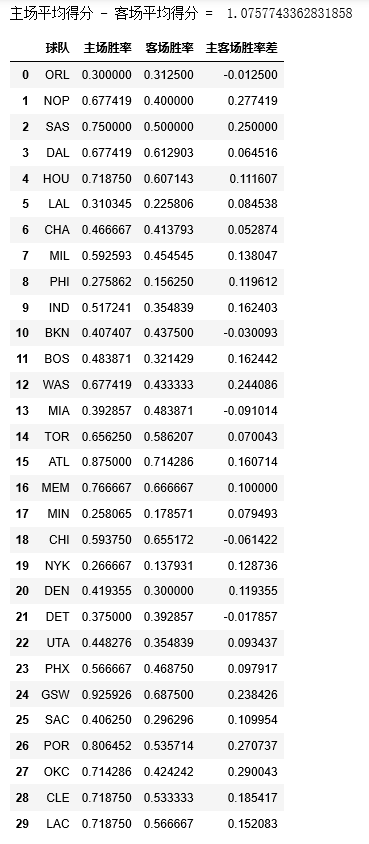
此外，其他变量之间相关性不显著并不说明它们没有有价值的关系，而是这些关系可能比较复杂，不能只用相关性简单描述，这也是进一步研究的重要原因。

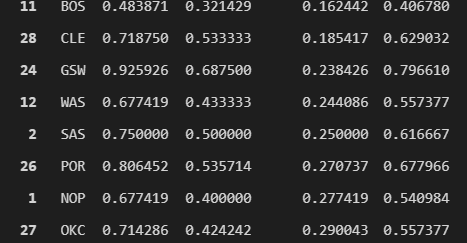
# Task1

第一个任务是探究主客场因素对每一支球队比赛结果的影响。

主场优势效应认为，运动队在主场比赛时取胜的几率要高于在客场比赛的取胜率。在本问题中，我们研究主客场胜率差别与分差分布来验证主场优势效应。

实现方法是分别计算出每一支球队的的主场胜率（主场胜/（主场胜+客场胜）），客场胜率，主客场胜率差，以及每支球队的总胜率。通过总胜率我们可以看出在当前数据下每支球队的的相对实力。将其按主客场胜率差排序后得到下表。



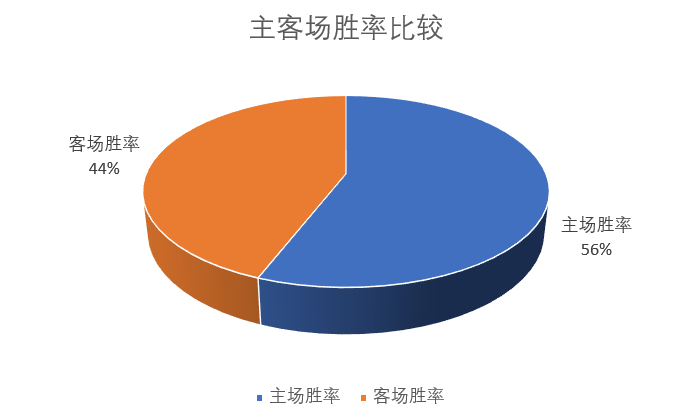


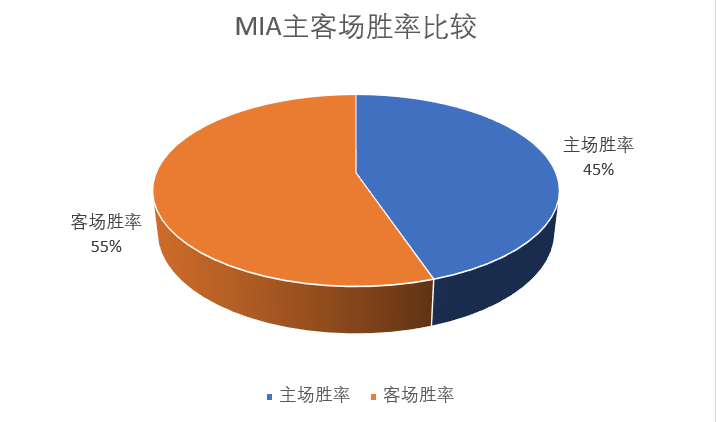
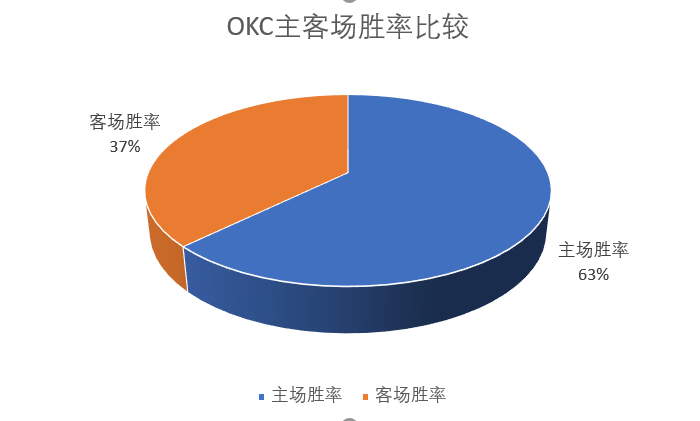
从上图看出大部分球队的主场胜率要高于客场胜率，其中OKC,NOP等球队主场胜率要高出25%以上,说明主客场因素对这些球队的影响比较大。



从上图看出部分球队主场胜率略低于客场胜率，但相差都在0.1以下，可能是由于偶然因素、数据量的问题引起的。

我们将胜率用饼状图的形式可视化，可以直观地看到，总体的主场胜率为56%，客场胜率为44%；主客场胜率差最大的OKC队主场胜率占总体的63%，客场胜率占总体的37%；主客场胜率差最小的是MIA队，分别占45%与55%。可以得到，主场优势效应对球队的影响确实存在且较为显著。

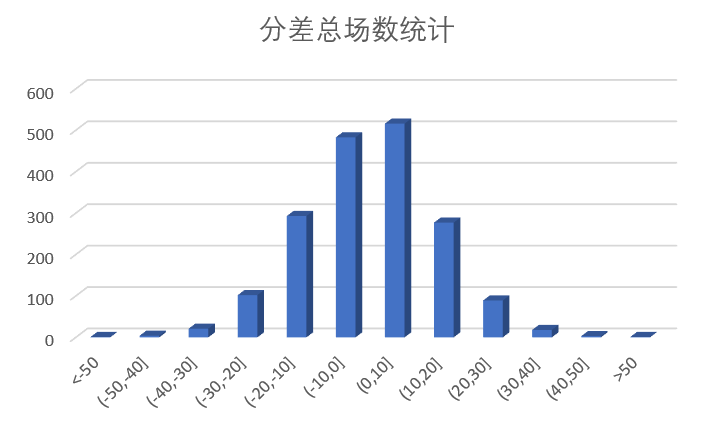




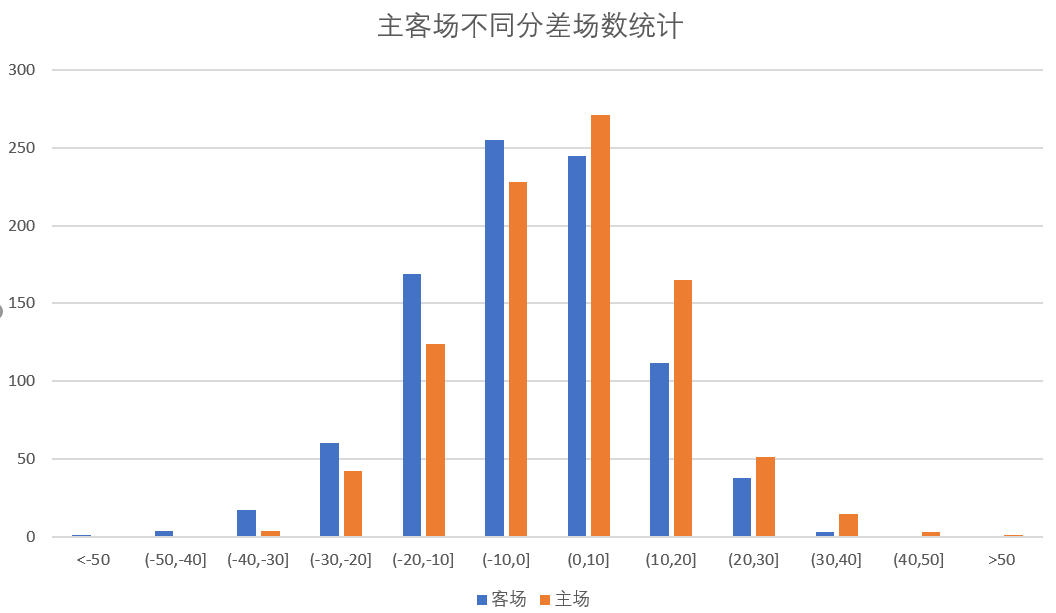
进一步，我们分析了主客场对分差的影响。首先统计所有场数的

分差分布，在考虑区间端点产生误差的情况下，分差总体呈对称分布，

可以由此验证数据集的完整性。



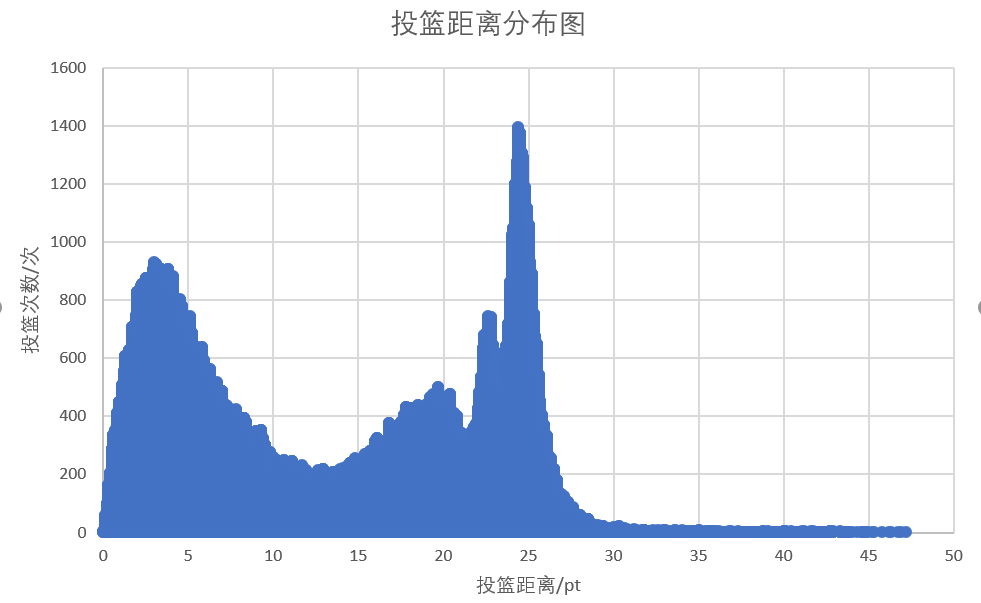
对主客场分别统计分差并作出分布图，可以看出在每一个大于零的分差段，主场均多于客场场次，同样的，在每一个小于零的分差段，客场场次多于主场。这说明主场优势效应在不同水平的球队之间的对局里都有体现。值得注意的是，所有分差大于40分的胜方球队均为主场（也就是败方全为客场），这也表明主场优势效应不仅可以增大胜利的几率，而且可以促进主场球队把胜利变成一场大比分胜利。

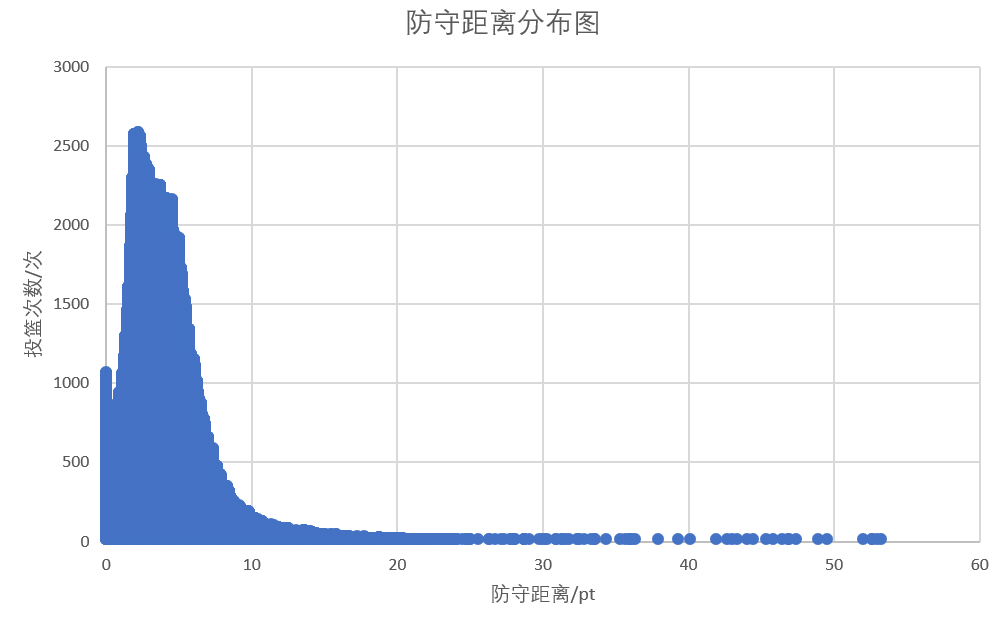


# Task2

第二个任务是探究球员们的投篮选择。我们这里主要考虑了投篮距离、防守距离、投篮时间、投篮前运球次数等因素。

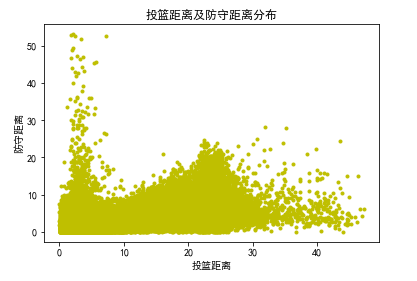
我们首先研究投篮距离和防守人距离这两个最明显的因素。分别做出投篮次数与投篮距离和防守距离的分布图如下：



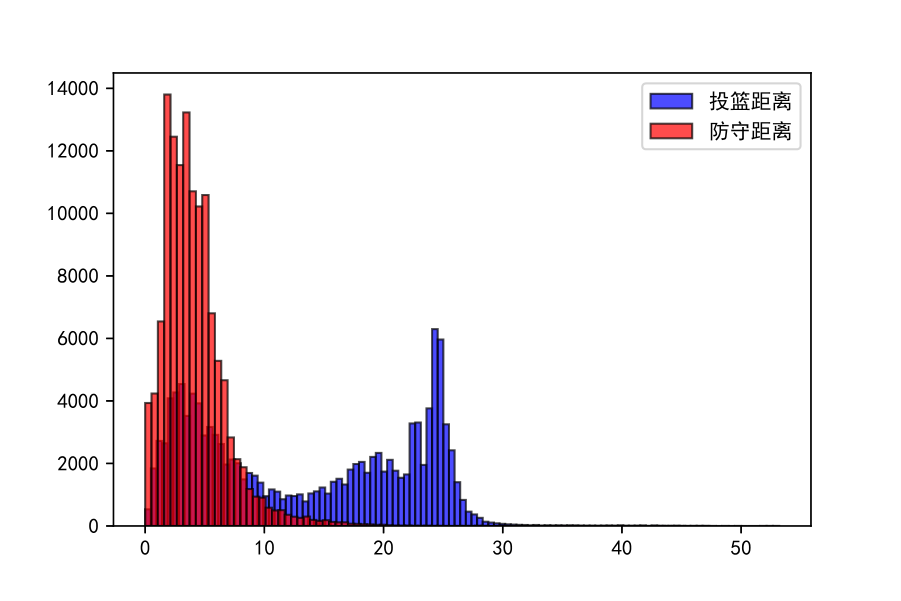


从投篮距离分布图可以看出，在0-5英尺、20-25英尺处的投篮次数最多，而在10-15英尺、30英尺以上的投篮次数最少。这说明NBA球员在比赛时倾向于短距离投篮和三分球，而不倾向于中距离投篮和超远距离三分球。这也是自然的，短距离投篮的命中率较高，三分球的得分较高，而中距离投篮和超远距离投篮的命中率相对较低，得分期望较小。根据图像可以推断三分线距篮筐距离应在20英尺-25英尺之间。查阅资料可知，NBA规定三分线距篮筐20.6英尺，推断符合事实。在之后对三位顶级得分手的分析我们可也以看出这一特点。

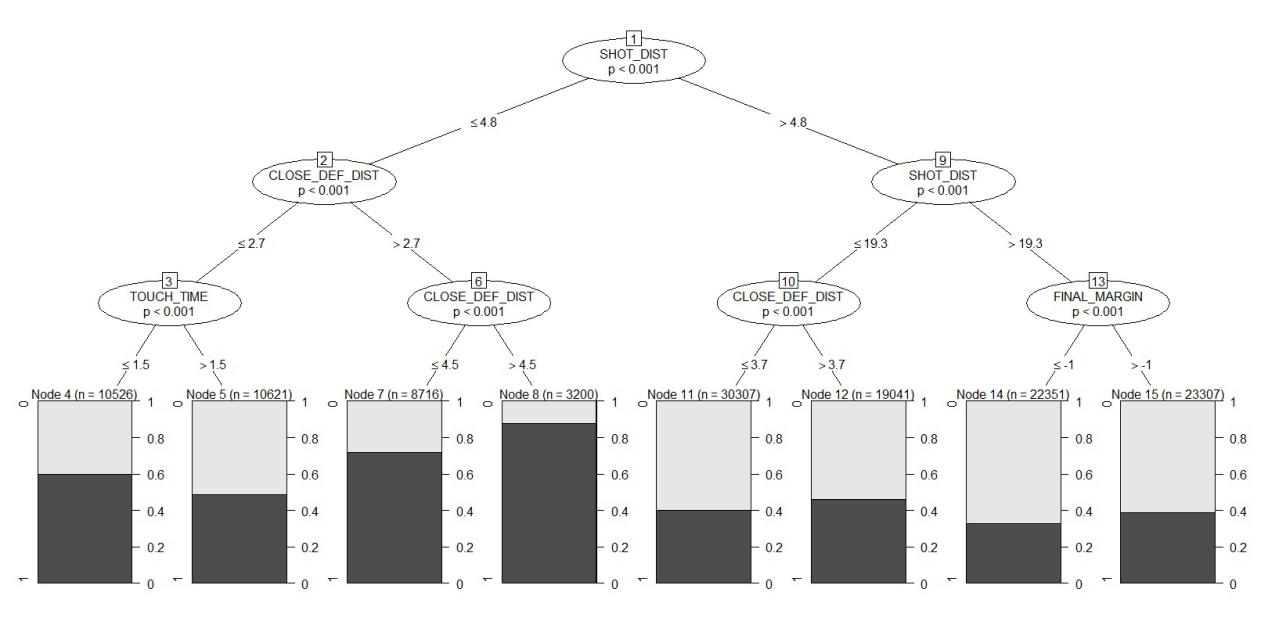
绝大多数投篮的防守距离在0-10英尺之间，高峰出现在2.2英尺处。我们推测，这应该是由进攻者与防守者博弈之后选择的结果。虽然防守者距离越远命中率越高，但是实际上防守者会尽可能缩小防守距离，因此进攻者只能在平衡之后选择防守者相对较远的时候投篮，也就是2.2英尺左右。



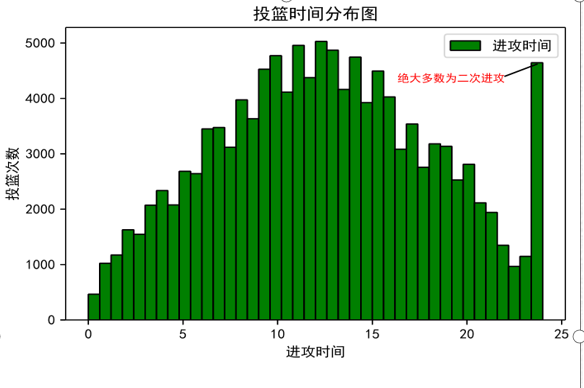
做出投篮距离与防守距离的散点图可以看出，对于大部分数据，在三分线内的时候防守距离上限随投篮距离增大而增大。对于投篮距离较小而防守距离很大的情况，应当为快攻导致的防守球员失位，是进攻球员的绝佳机会。而对于大于30英尺的投篮，数据总量较小，数据分布无明显规律。



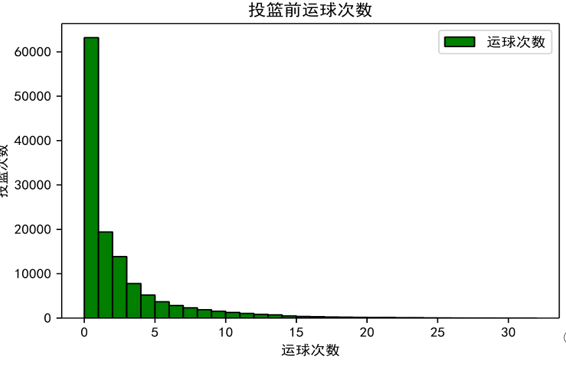
在进一步分析中，我们探究球员在什么条件下的投篮命中率较高，可以依此制定得分效率更高的进攻策略和防守策略。使用决策树分析SHOT\_DIST、CLOSE\_EDF\_DIST、TOUCH\_TIME等数据的结果如下图（具体实现见r代码部分）：



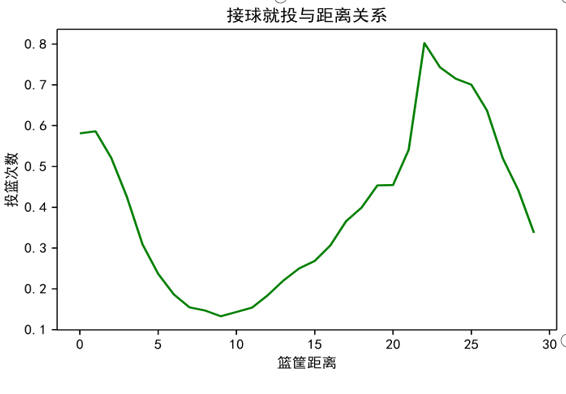
从结果中可以得到：当投篮距离小于4.8英尺，防守距离超过4.5英尺时，投篮命中率可以达到90%左右。同时，持球时间大于1.5s的命中率显著低于持球时间小于1.5s的命中率。这样，就可以通过团队协作与个人突破等方式尽可能创造投篮距离较小、防守距离较大并且尽可能接球就投的条件，使得进攻更具有侵略性。同时，防守球员也应当尽量避免这种条件的发生。



利用SHOT\_CLOCK这组数据做出进攻时间分布的图（进攻方在24秒内必须投篮，并且球必须触及篮框，否则即违例）。可以看出攻击时间大多集中在10到15秒之间，而最后出现峰值是由二次进攻引起的（进攻方球员出手未进时，进攻方摘下前场篮板补篮，或者摘下前场篮板重新组织进攻的行为。进攻方得球进行二次进攻时，进攻时间由24秒改为14秒）



从运球次数的分布图可以看出，投篮数与运球次数呈显著的负相关，其中绝大多数投篮前运球次数为0，这说明在NBA比赛中，球员更倾向于通过团队配合与传球完成得分，而不是单纯依靠球员的个人能力单打独斗。



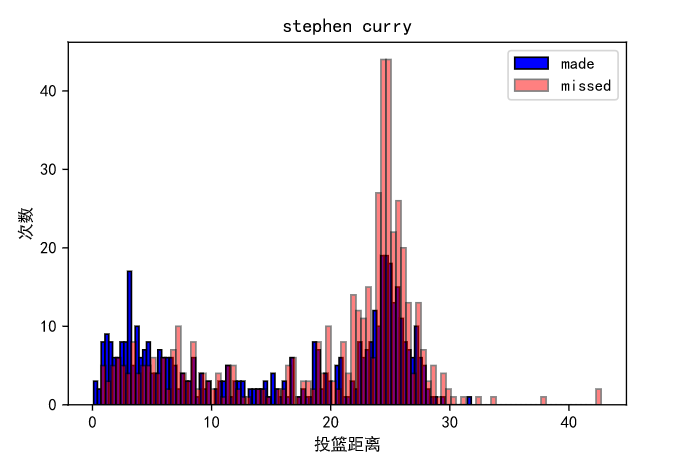
# 我们对投篮次数为0的数据进行进一步分析。将投篮前运球次数为0

# 为接球就投的投篮，做出接球就投的投篮次数与投篮距离的折线图。我们观察到与之前分析相似的趋势：在内线与三分线处最多，中投与超远距离投篮较少。但不同的是，与总投篮次数相比，在接球就投中，中投的比例更少，这也在一定程度上验证了之前的猜测，中投是球员不得已而为之，内线与三分才是战术和球员更加喜欢的投篮方式。

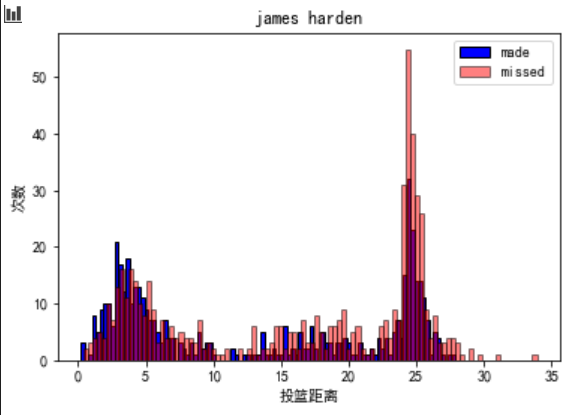
# Task3

第三个任务是找到得分前三的球员，分析他们在不同距离下投篮命中情况。

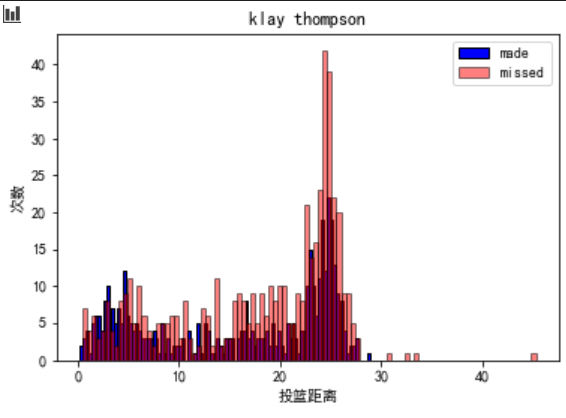
我们制作出子表格，统计出每一位球员的得分情况再进行排序。排在前三位的分别是Stephen Curry,James Harden,Klay Thompson.



Stephen Curry在近距离投篮的时候命中率高于未命中率，而距离在三分线左右时命中率低于未命中率。投篮集中在内线和三分.



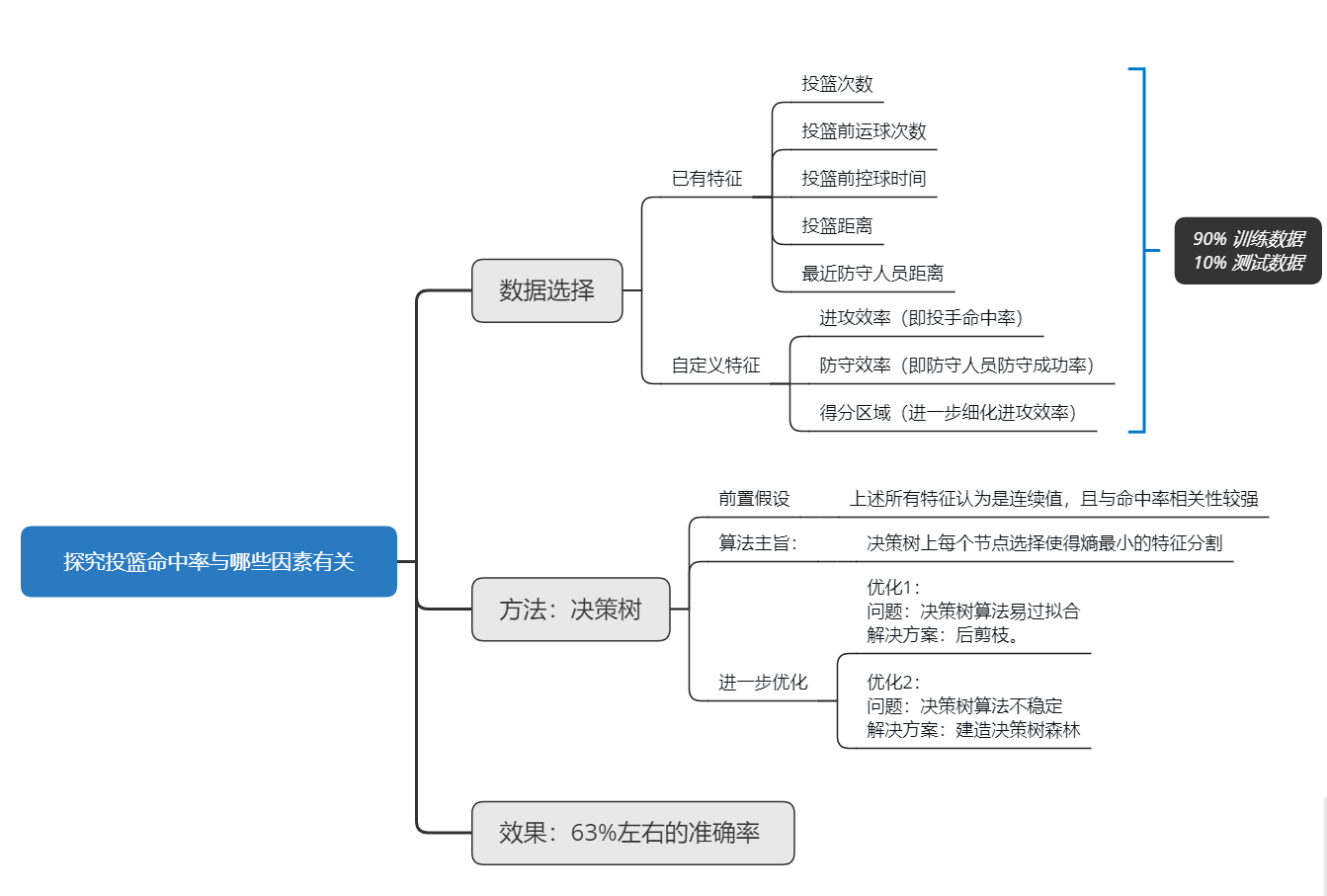
James Harden在较近距离时命中率高于未命中率，而三分线附近命中率低于未命中率。相比前后两位球员，他中投更少。



总的来说，三位球员得分能力很强，他们在内线附近成功率比较高，这也符合距离相对较近时命中率较高的认知。三分球的出手明显高于中投，而且三分球命中率也相对比较高。

# Task4

第四个任务是探究投篮命中率较高的方式。原有数据中我们选择了投篮前运球次数、控球时间、投篮距离、最近防守人距离为特征，同时自定义了球员命中率、防守成功率。得分区域三个特征。运用决策树的方法进行分析。



算法：决策树上每个节点选择使得熵最小的特征分割方案，自上而下构建决策树，枚举当前可分割的所有特征及其分割权值（此处假设所有的分割都为二分割），并计算该分割下的熵的期望值，选择熵期望值最小的分割方案，对子树递归构建。具体实现见代码及注释部分。

优化1：决策树算法存在易过拟合的问题：当设置最大深度为10时，在训练数据上正确率为90%以上，而在测试数据上只有57%左右的正确率。为解决该问题，运用剪枝的方法进行优化，将训练数据中分解出一部分数据作为测试数据，在完成建树后，利用测试数据进行剪枝，具体方案为对决策树自底向上，对于每个节点判断是通过决策树得到的正确率高，还是通过测试数据中众数结果得到的正确率高，取两者中的最优解。

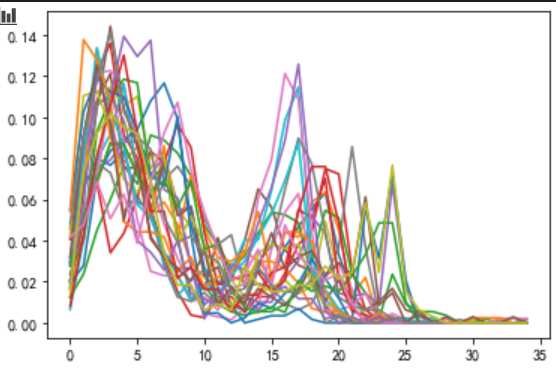
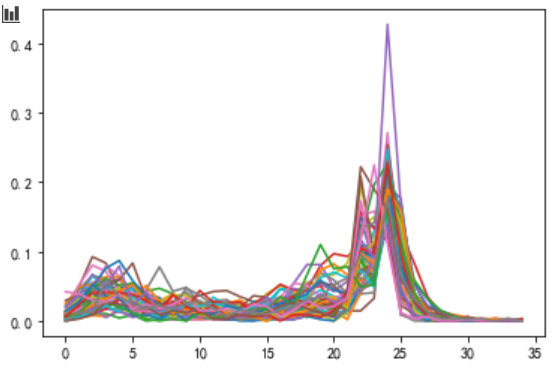
优化2：决策树算法存在不稳定的问题：数据轻微波动对决策树的构建影响极大。运用建造决策树森林的方法进行优化：取其中众数作为最终判断，结合优化1，具体实现为将测试数据分为9份，共构建9棵决策树，其中第i个决策树由除第i份测试数据外的8份数据，并利用第i份数据进行后剪枝所得。

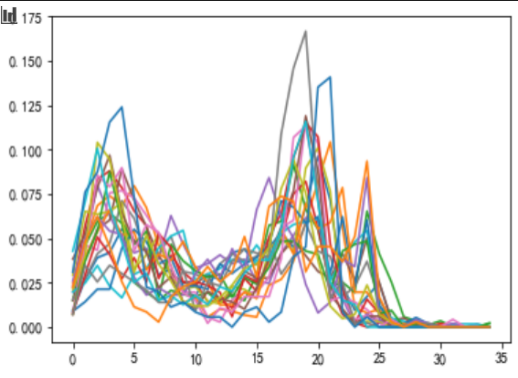
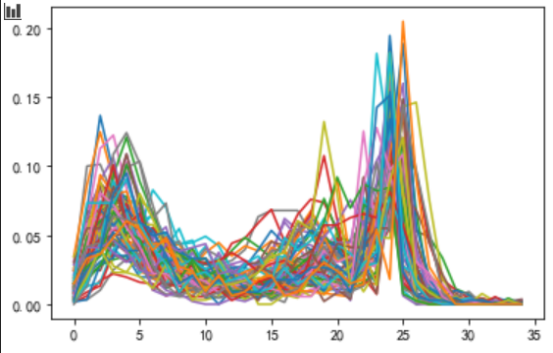
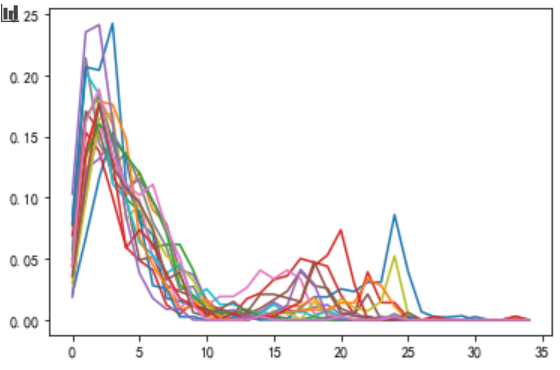
# Task5 附加题

我们小组完成的附加任务为基于投篮距离特征的球员分类任务。

数据处理，对于总投篮数大于320的球员构建投篮距离向量。（这里将相距一英尺内的投篮数与总投篮数的比值作为特征）。对所得的投篮距离向量，利用K-Means算法进行聚类，得到相似的向量，我们则认为对应球员具有一定的相似度。

取K等于5可以得到下列结果。





1.klay thompson（汤普森），jamal crawford（克劳福德），kevin love（乐福），avery bradley（布拉德利）为完全的外线进攻球员，三分是主要的进攻手段

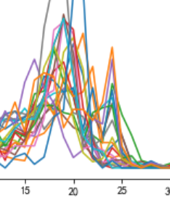
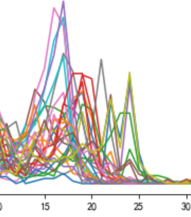
2. Nikola Vucevic（武切维奇），Pau Gasol（加索尔），Al Jefferson（杰弗森），DeMarcus Cousins（考辛斯） 为内线球员，同时具有一定的中远距离投射能力

3. Tyreke Evans(埃文斯), greg Monroe(门罗), Andre Drummond(德拉蒙德), Enes Kanter(坎特)为完全的内线球员，几乎全部进攻都是在三秒区内完成的。

4. james harden(哈登)，mnta ellis（埃利斯），damian lillard（利拉德），lebron james（詹姆斯），russell westbrook（威斯布鲁克），stephen curry（库里），kyrie irving（欧文），Chris Paul（保罗）归为一类。绝大多数球队的当家球星都是该类型的球员，在内线能上篮，在外线也有良好的三分投射能力。

5. LaMarcus Aldridge（阿尔德里奇），Blake Griffin（格里芬），Anthony Davis（戴维斯） 与第二类球员类似，但相对第二类球员在中远距离投射球更多。

问题：可以发现下面两类中分类仍存在不合理现象。问题的根源在于投篮距离的差异在不同区域是不同的，尤其是在三秒区以及三分线周围的区域，影响就较大。因此可以采用带权向量距离去描述两个球员的差异。



从现实角度看，他们本不应该分为同一类球员。

一般评价一名球员的进攻能力往往是通过进攻命中率，但是一名内线球员和一名外线球员的命中率往往是难以比较的。因此分类后，对于同类型的球员之间的数据也就有了一定的可比性。同样，对于防守效率的评价可以依靠进攻球员的进攻能力以及在该位置的得分效率来求解。

# 结论

在本项目中，我们首先通过对数据集进行相关性分析，以及观察相关性热力图，发现投篮距离、防守距离等变量对投篮选择与命中率有着较为显著的关联。在数据分析部分，我们运用数据可视化方法验证了主场优势效应，并且探究得到投篮距离、防守距离等因素对于球员投篮选择的显著影响。之后，我们分析了得分前三名球员的投篮选择与命中率，发现了他们之间的显著区别。进一步地，我们选取投篮距离等特征，并自定义了命中率新的特征，运用决策树的方法对投篮命中率进行分析，同时进行优化以解决拟合率低和算法不稳定的问题，得到了63%左右的正确率。在附加题中，我们运用基于投篮距离特征的K-Means算法进行聚类，并引入带权向量距离，对球员能力进行分类，并结合实际情况验证结果的合理性。

# 小组分工

# 报告部分：

# 数据预处理：李鸿泽

# Task1：孙凌宇、李鸿泽做图并分析

# Task2：毛一琛、李鸿泽做图，李鸿泽、孙凌宇、毛一琛分析

# Task3：毛一琛做图，孙凌宇分析

# Task4：毛一琛

# Task5：毛一琛

# 摘要、结论等其他部分：李鸿泽

# PPT部分：三人合作完成

# 海报部分：孙凌宇

# 致谢

感谢童云海老师一学期以来的教导，感谢两位助教在项目答疑中的提供的帮助。

# 参考文献

[1]决策树算法（一）——一些重要的数学概念 <https://www.cnblogs.com/wyuzl/p/7699738.html>

[2]决策树算法及Python实现 <https://blog.csdn.net/qq_34807908/article/details/81539536>

[3] Python统计分析可视化库seaborn <https://www.cnblogs.com/caiyishuai/p/11184166.html>