实验报告

——数据分析与特征提取

PB19020499 桂栋南

目录

[1. 数据分析 3](#_Toc5798)

[1.1. 单个特征的分布 3](#_Toc22097)

[1.2. 异常值分析 3](#_Toc23100)

[1.3. 特征间的关系 4](#_Toc31786)

[1.3.1. 与胜率之间的相关性(以player1为例) 4](#_Toc27136)

[1.3.2. player属性内部的相关性（以player1为例） 5](#_Toc20362)

[1.3.3. 不同player之间的相关性 5](#_Toc21501)

[1.4. 特征与预测目标 5](#_Toc8420)

[1.4.1. ChampionID对胜率的影响 5](#_Toc32539)

[1.4.2. player\_lane(player\_role) 对胜率的影响 7](#_Toc23199)

[1.4.3. 考察总经济对胜率的影响 8](#_Toc22212)

[2. 特征提取 9](#_Toc8885)

[2.1. 特征变换 9](#_Toc30838)

[2.1.1. str -> int 9](#_Toc6759)

[2.1.2. 取log 10](#_Toc6946)

[2.2. 组合特征 10](#_Toc13498)

[2.2.1. 两队的经济差 10](#_Toc26186)

[2.2.2. 尝试组合firstblood,firsttower,firstinhibitor 11](#_Toc8583)

[2.3. 特征聚集 12](#_Toc24373)

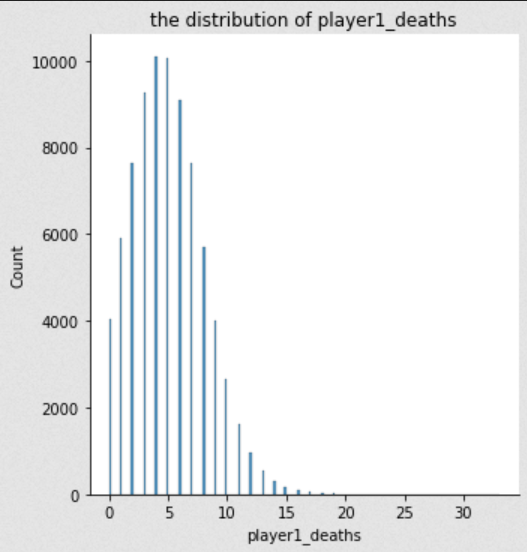
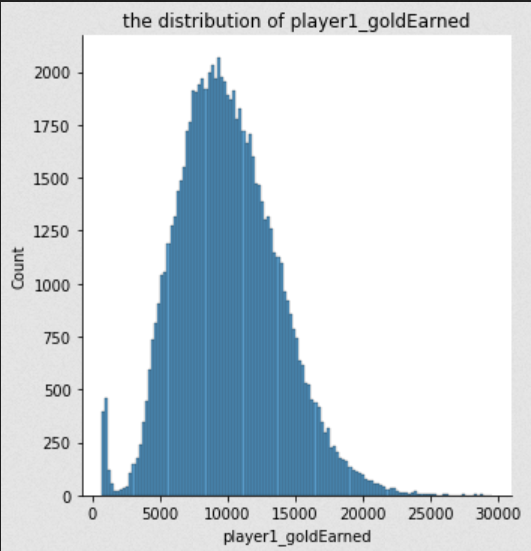
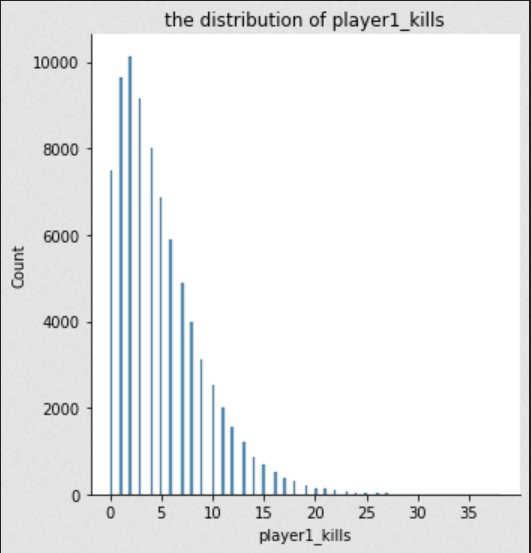
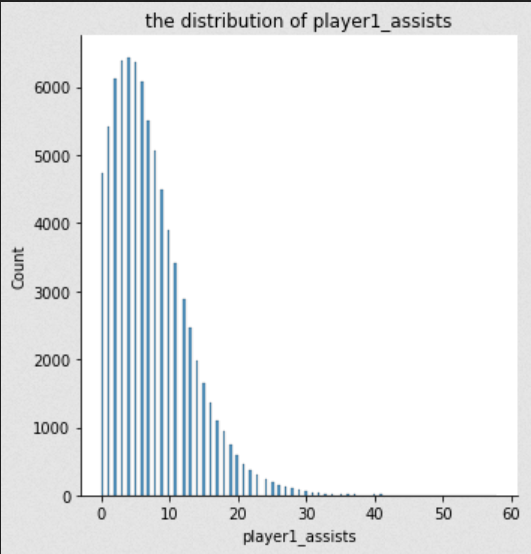
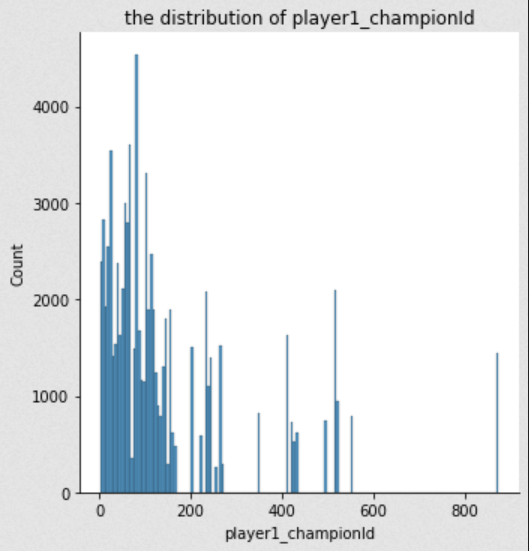
[2.3.1. 利用PCA实现数据降维 12](#_Toc31339)

[2.3.2. 利用featureagglomeration实现特征聚集 12](#_Toc1212)

# 数据分析

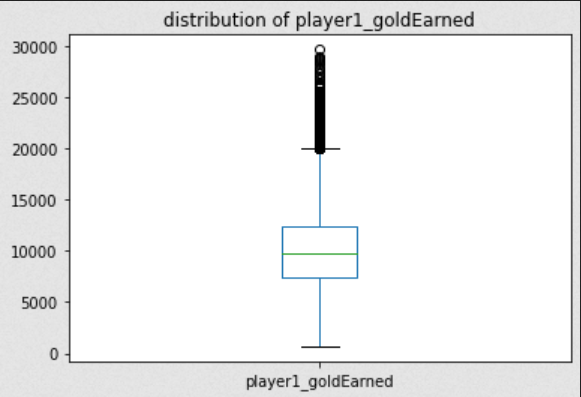
## 单个特征的分布

* player\_goldEarned的分布（以player1为例）
* player\_kills的分布（以player1为例）
* player\_deaths的分布（以player1为例）
* player\_assists的分布（以player1为例）
* player\_championId的分布（以player1为例）



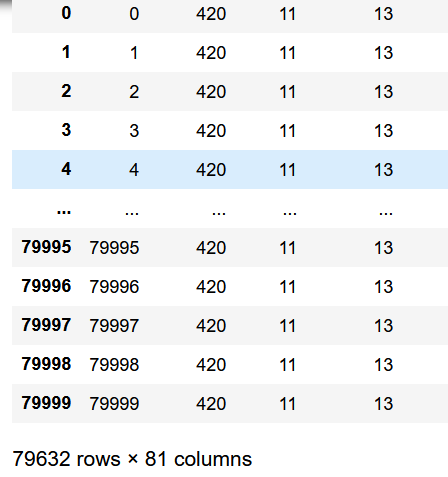
## 异常值分析

* 最后并没有采用异常值处理后的数据进行分析
* 以player1\_goldEarned 为例
  + 箱型图分析
  + 正态拟合分析



离群值较多，不易分析，采用正态拟合尝试

* 3\*sigma 离群值舍弃

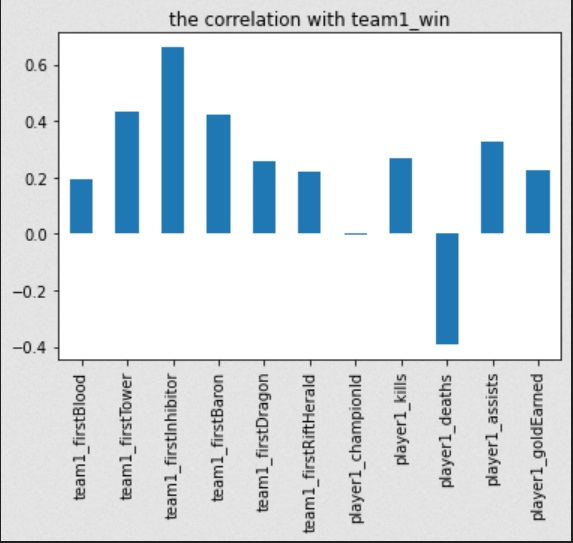


剩余79632，大致删除300+条数据。

## 特征间的关系

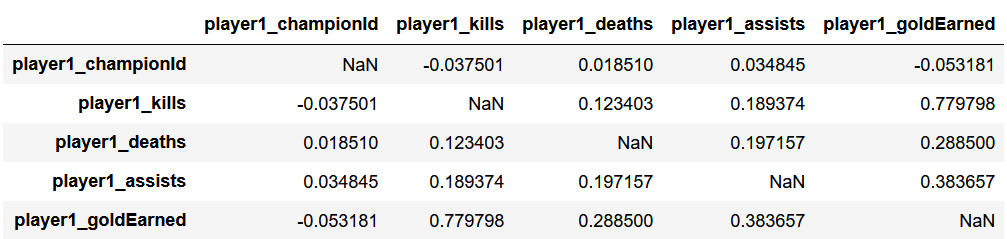
### 与胜率之间的相关性(以player1为例)

* Pearson 相关性
  + 可以看出team1\_firstInhibitor与胜率相关性最大



### player属性内部的相关性（以player1为例）

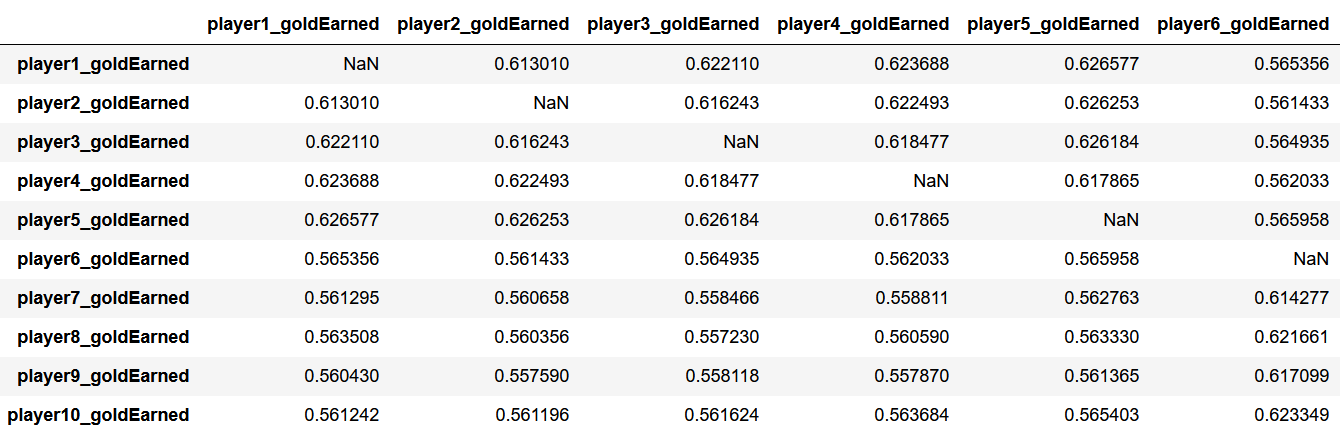
可以看出player内部kills与goldEarned的相关性最高（与游戏尝试相符）



### 不同player之间的相关性

以goldEarned为例

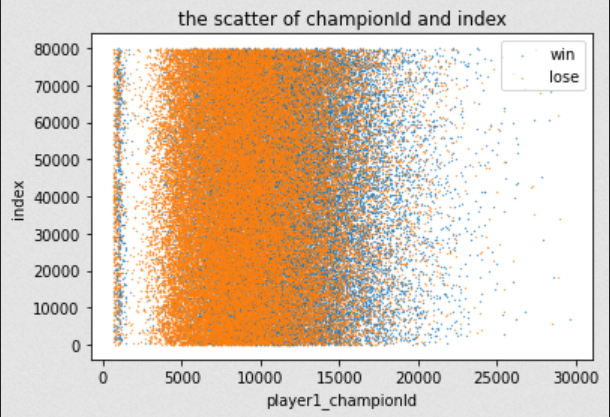
* 竟然发现不同player的经济是正相关的，不分队伍，震惊！
* 这或许可以从游戏时长的角度来解释



## 特征与预测目标

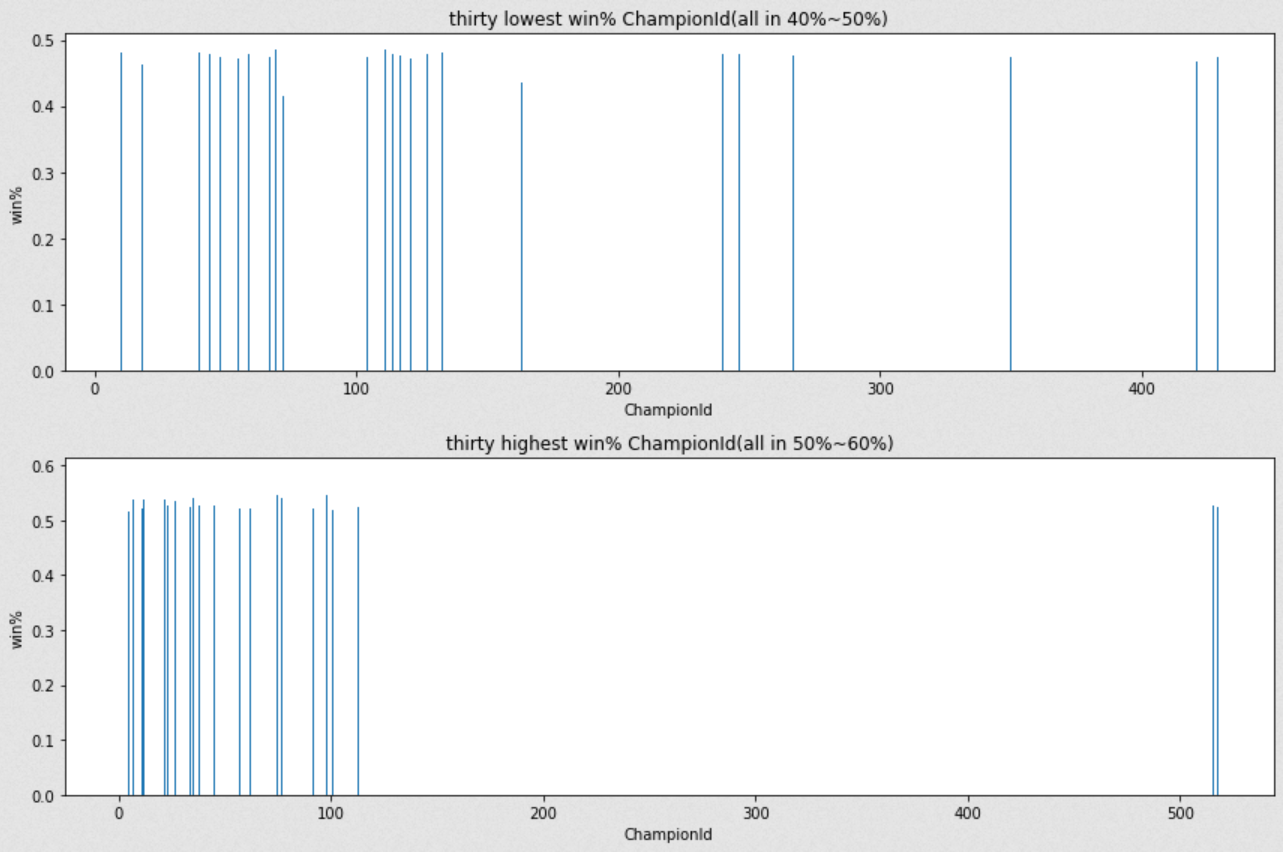
### ChampionID对胜率的影响

大致分布的散点图（仅player1）

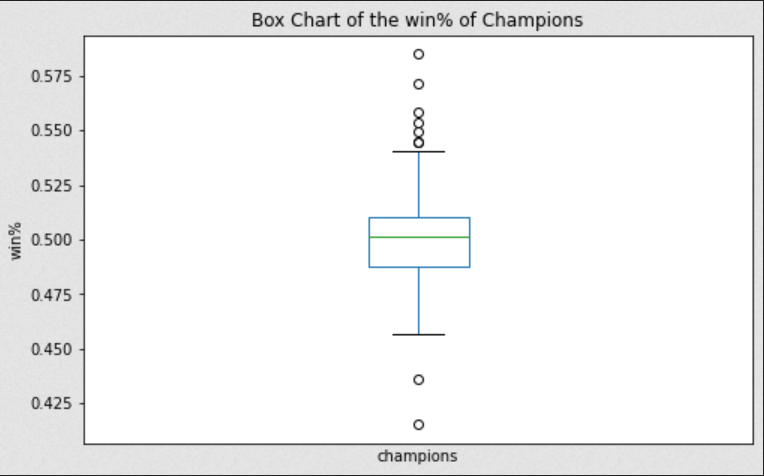


各个英雄胜率的统计图

* 直方图
  + 胜率前30和后30位的ID分布



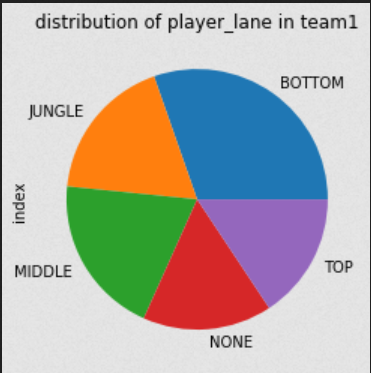
* 箱图
  + 不同英雄胜率的总体分布



### player\_lane(player\_role) 对胜率的影响

先大致了解player\_lane的分布情况

* 仅对team1进行了初步分析

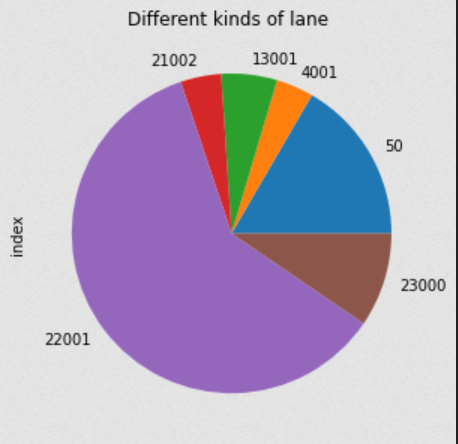


player\_lane对胜率的影响

* 采用map方法获取不同lane组合

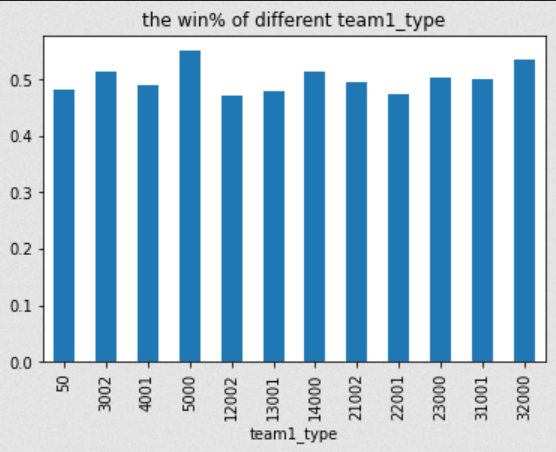
不同组合的饼图

* 其中第一位表示BOTTOM的数目，而后表示JUNGLE,MIDDLE,NONE,TOP的数目
* 0略去不表



不同lane的组合方式的胜率直方图

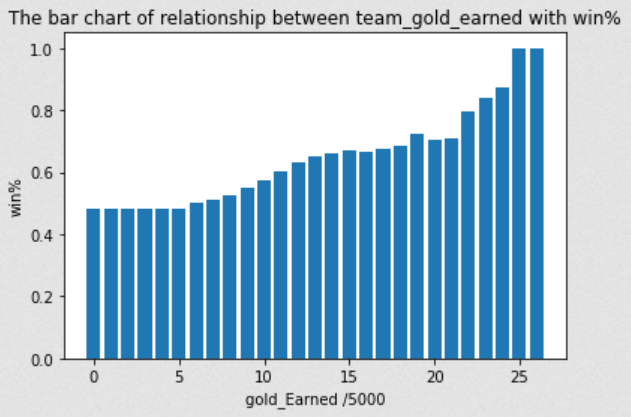
* 神奇的发现，可能五个打野的胜率高一点！



### 考察总经济对胜率的影响

画出不同经济与胜率之间的关系图

* 看出明显的正相关关系
* 当经济大于100000时，几乎必赢



# 特征提取

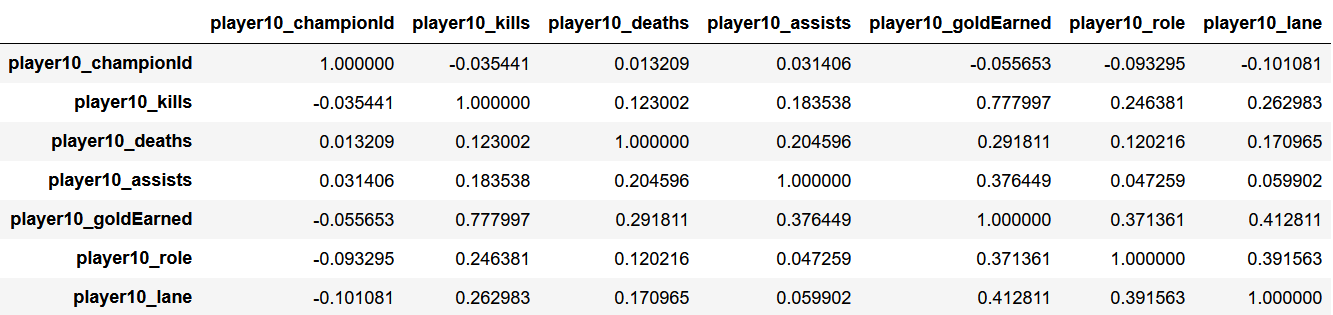
## 特征变换

### str -> int

将player10的str数据转换为int格式

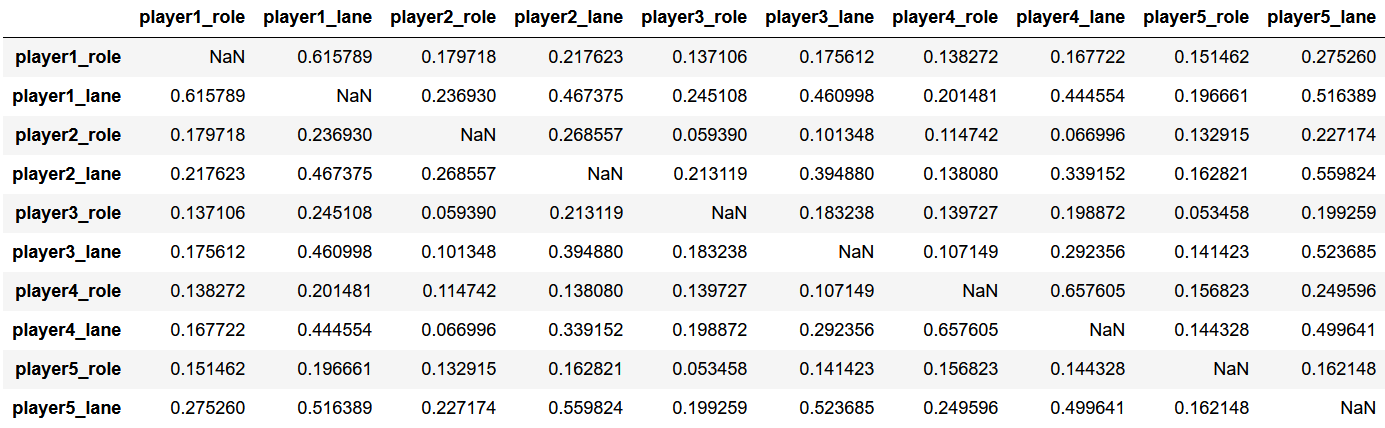
获得player10的各项数据之间的相关系数

* int化的player10\_role和player10\_lane与其余值的相关性都不大



下面获取player1~5之间str值的相关性

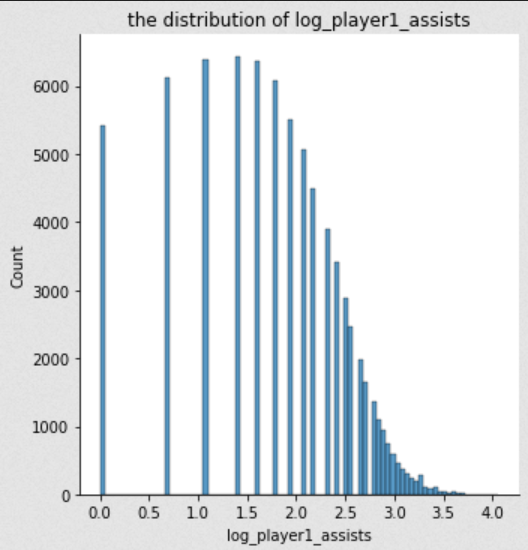
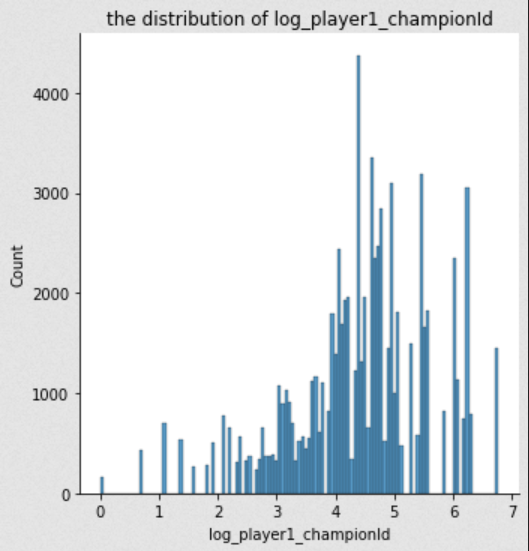
* 可以看出不同player间的lane和role的相关性也不高
* player1\_role和player1\_lane间的相关性最高

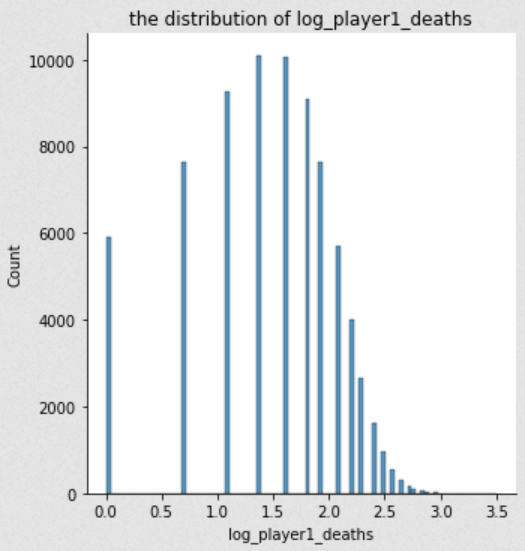
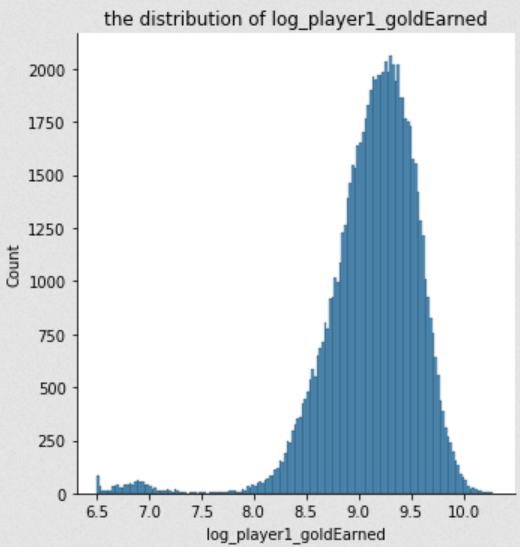
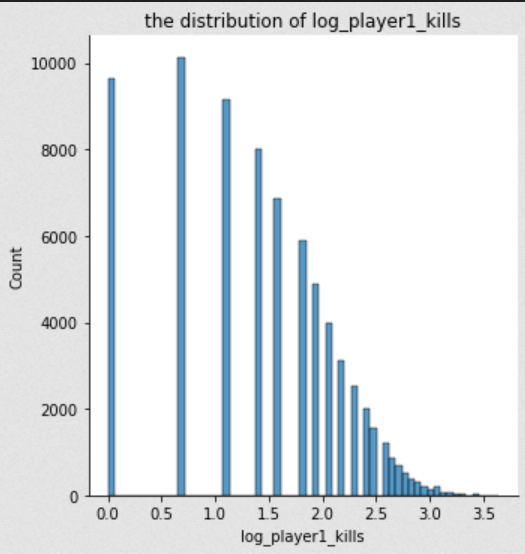


### 取log

以player1的各个数字特征为例

没有看见明显的分布趋势





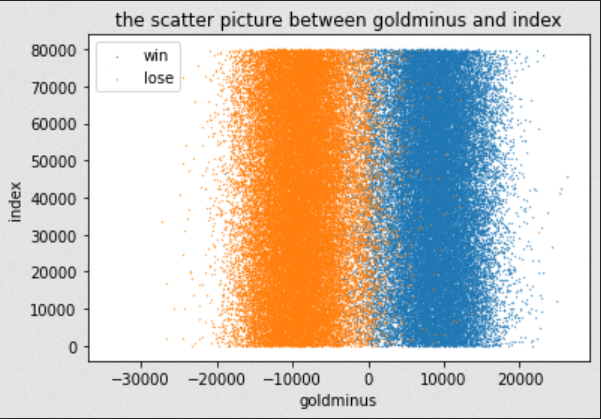
## 组合特征

### 两队的经济差

建立新的一列——经济差

画图查看大致趋势

* 可以明显看到分类聚集的现象
* 分类面大致为goldminus=0

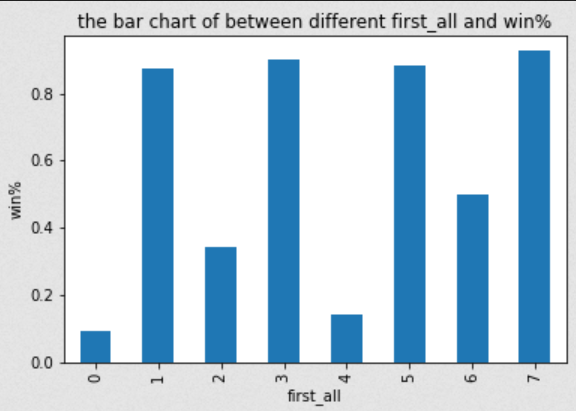


调用knn算法简单模拟

* 看出仅通过两队经济差来判断胜负，就有97%的准确率！

### 尝试组合firstblood,firsttower,firstinhibitor

尝试组合的原因是因为这三者与team1\_win的相关性相对最高



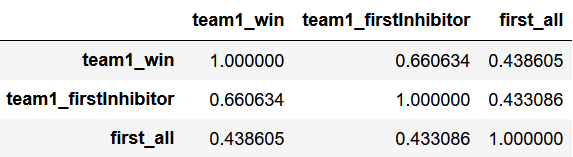
8种first\_all代表firstblood,firsttower,firstinhibitor的所有组合方式

是三种组合的二进制转十进制表示，如5代表101表明firstblood取1，firsttower取0，firstinhibiter取1

最大值取值处分别为001,011,101,111代表firstinhibiter对team1\_win的影响最大

三者的组合特征与胜率的相关性小于单纯的firstinhibiter相关性

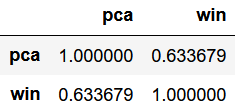
* 这个组合特征取的无效！
* sad! sad! sad!



## 特征聚集

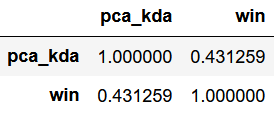
### 利用PCA实现数据降维

* 降维对象是team1除了team1\_win的所有对象



降维可能比较成功。

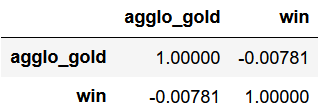
* 尝试对team1的kills,deaths,assists进行降维



失败了，合并后的聚集特征明显造成了相关性降低。

### 利用featureagglomeration实现特征聚集

* 失败了！不用看了！



两队的经济特征聚集到一维与胜负不相关。

可能的原因是游戏时长的随机性导致了经济特征的随机性。