In early parts of the projects, basic statistical analysis was done to the dataset of performances of all players recorded from professional matches in ATP tournaments. For example, the abilities of various players in the aspect of serving.

In the later phase of the project, an approach to predict who is going to win the final in tennis tournament was raised by applying machine learning models.

The basic idea and motivation in background is that the performance of the final players in the past 6 matches (in grand slam) may reveal something about their chances to win the championship in the final match.

The features are based on the comparison between two players. The classifier aims at separate the cases that first player wins and second player wins.

[1]

Let's first look at the abilities of serving of various tennis players. The features that can represent the ability of serving could be: the frequency of aces, breakpoints saving rate, 1st serve won rate. In some sense the 1st serve in rate can be also taken into consideration.

Data source: the match data from 2000 to 2016. In total 46799 matches.

The correlation matrix of related attributes:

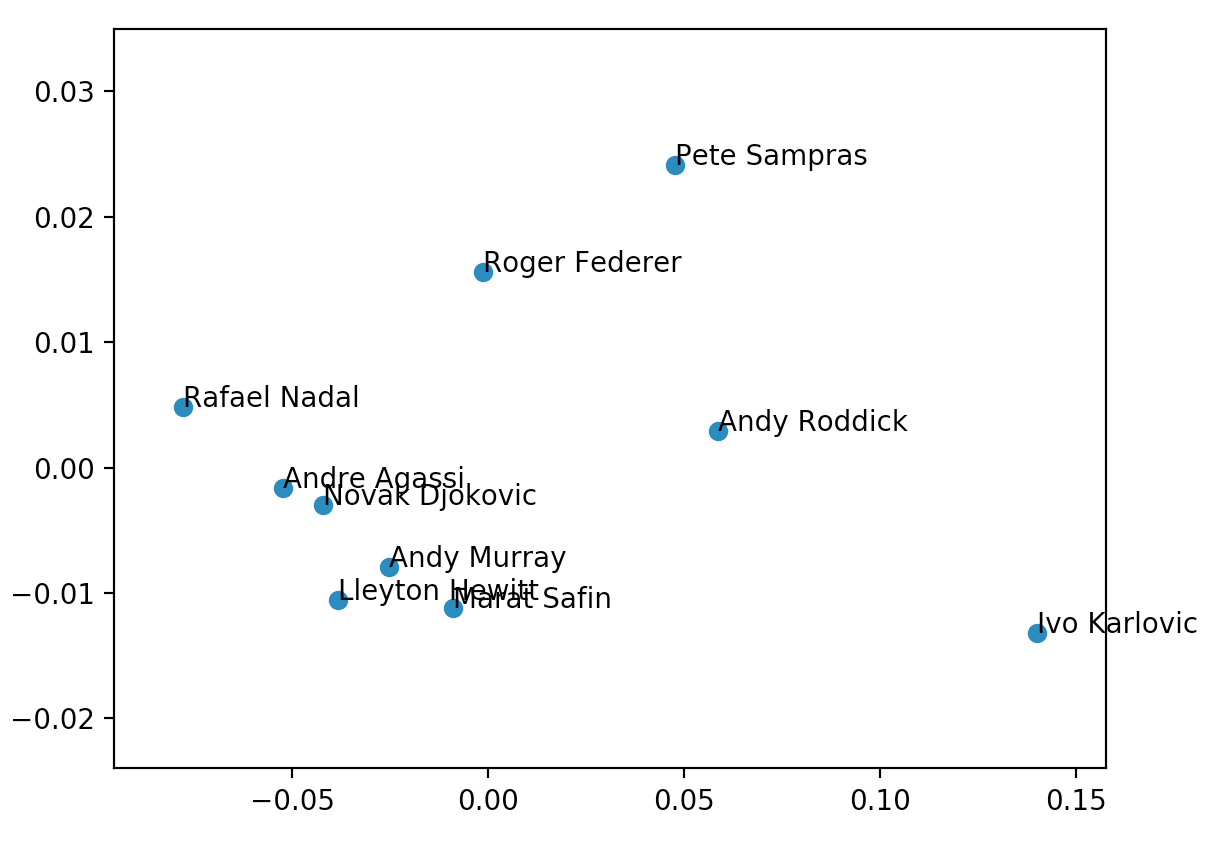
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ace | df | 1stIn | 1stWon | bpFaced | bpSaved |
| ace | 1 | 0.341444 | 0.138945 | 0.917376 | -0.721957 | -0.567597 |
| df | 0.341444 | 1 | -0.570393 | 0.348611 | 0.080848 | -0.199852 |
| 1stIn | 0.138945 | -0.570393 | 1 | 0.157907 | -0.617207 | -0.004676 |
| 1stWon | 0.917376 | 0.348611 | 0.157907 | 1 | -0.843422 | -0.662172 |
| bpFaced | -0.721957 | 0.080848 | -0.617207 | -0.843422 | 1 | 0.550028 |

The evident facts are that: aces and 1st serve won are highly correlated, and the more aces one has served, the less break points he needs to face.

[2]

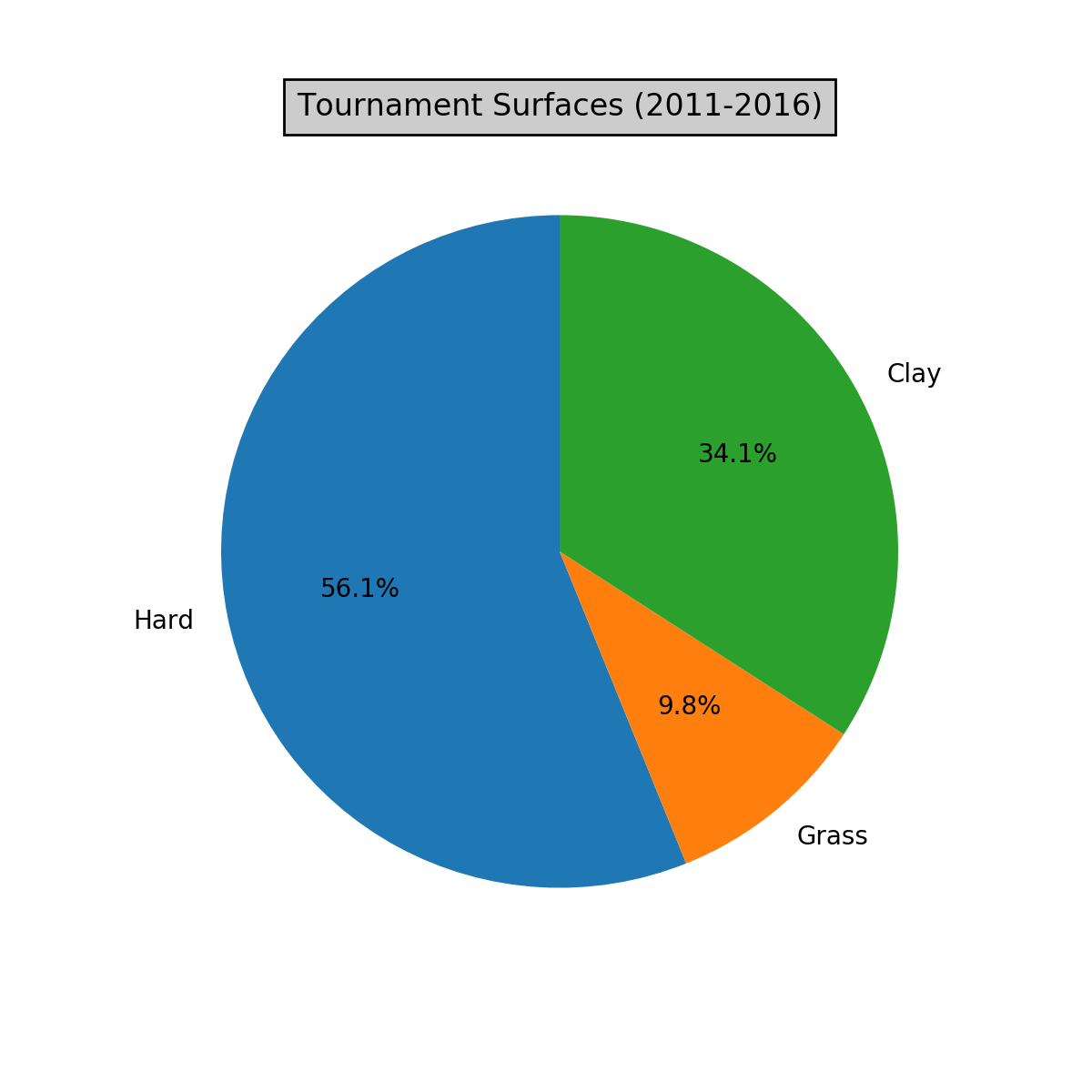
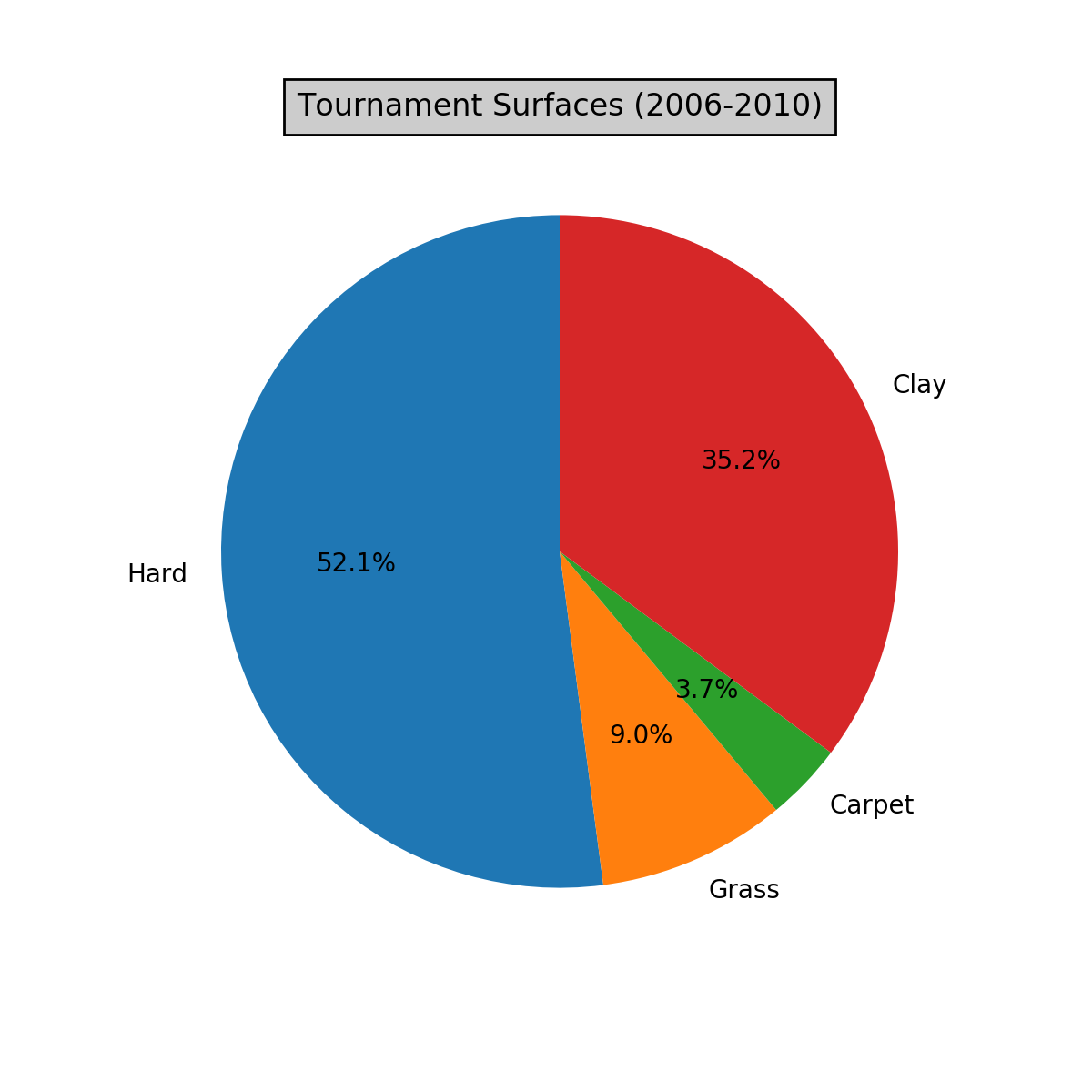
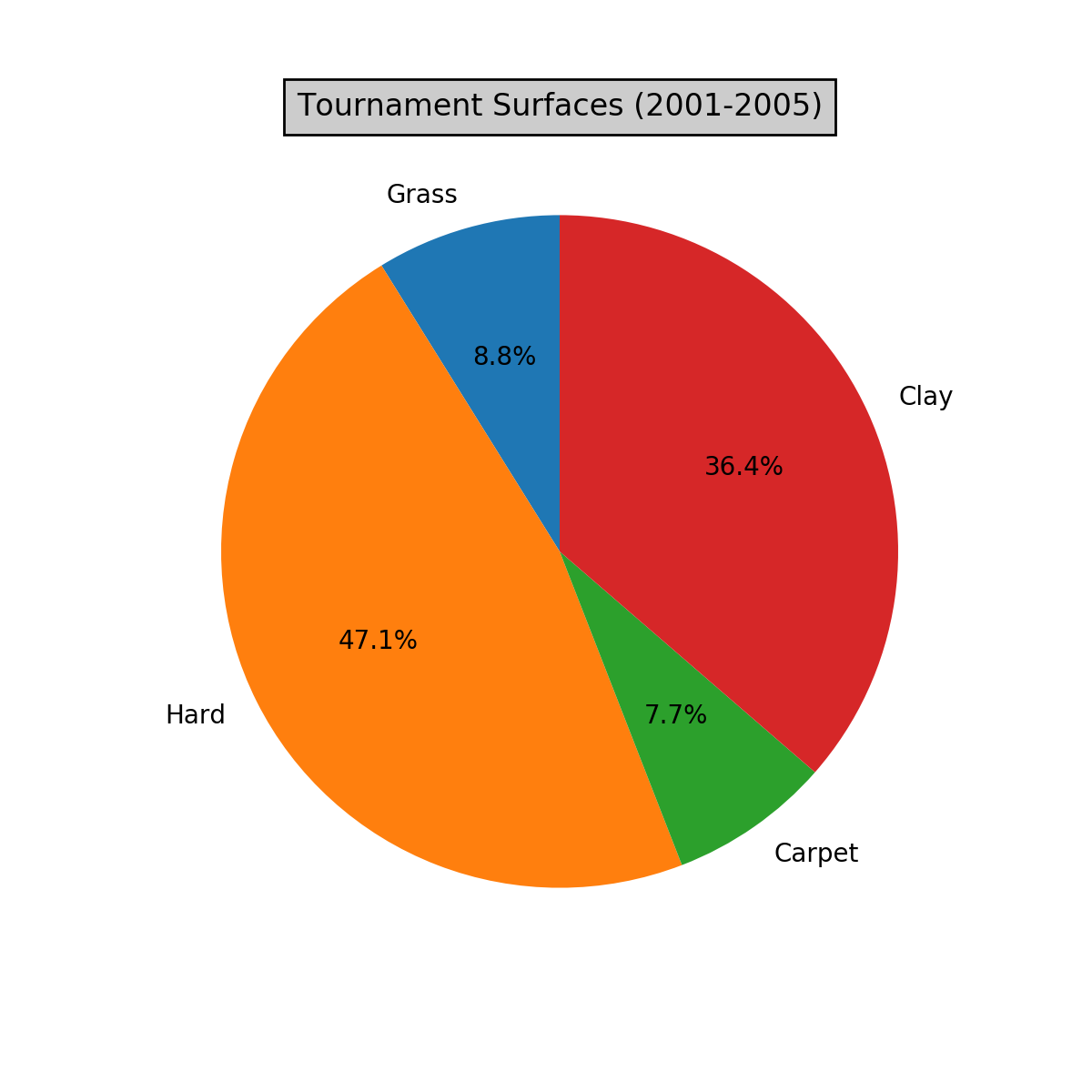
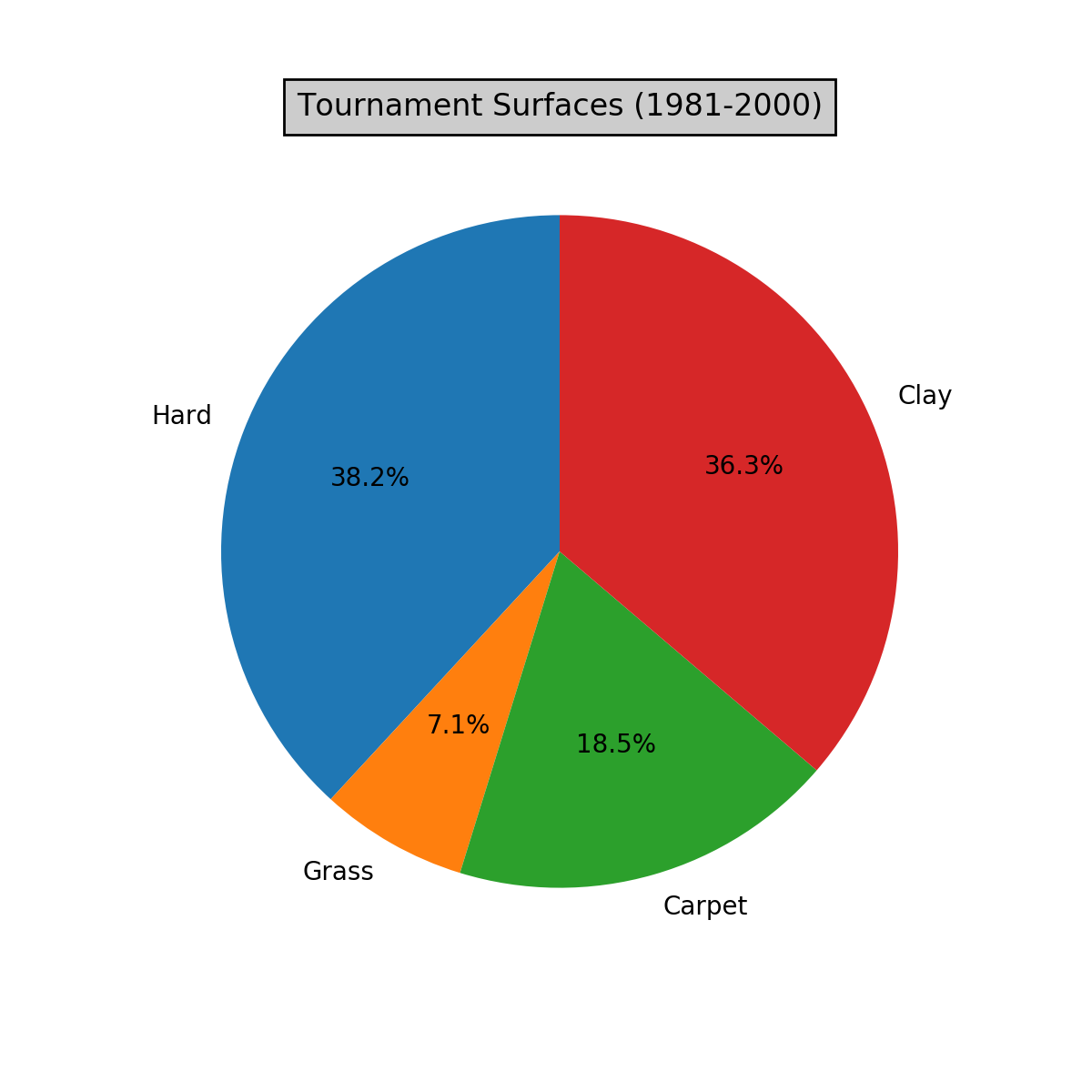
Got some descriptions of well-serve players? The first impressions the have come to mind could be: many aces, less double faults, high 1st serve in and high 1st serve won, less breakpoints faced, or high breakpoint saved ratio.

If we consider the above mentioned features and do the principle component analysis in order to visualize the players’ abilities onto a 2-D plane, we can obtain the following:

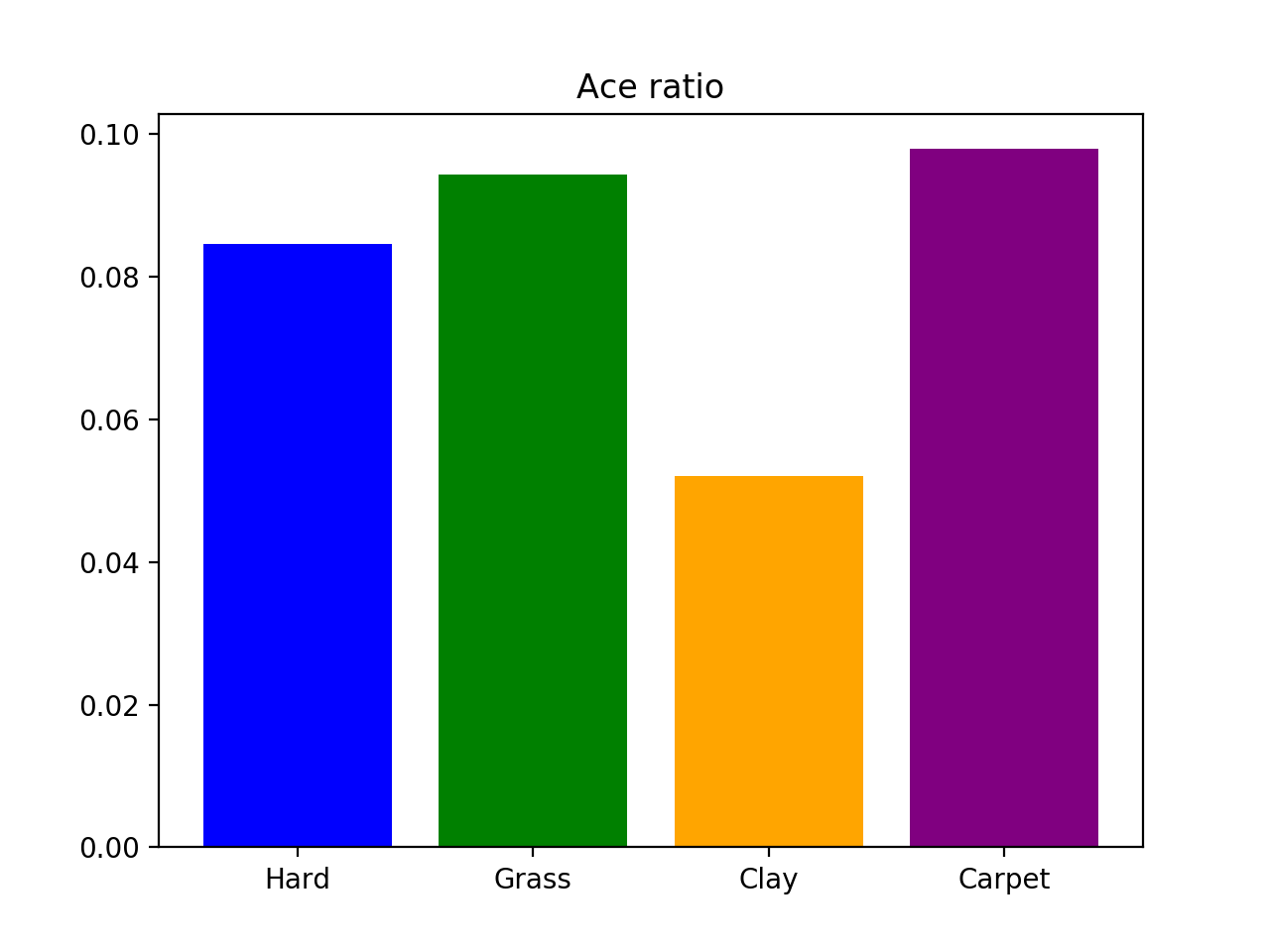


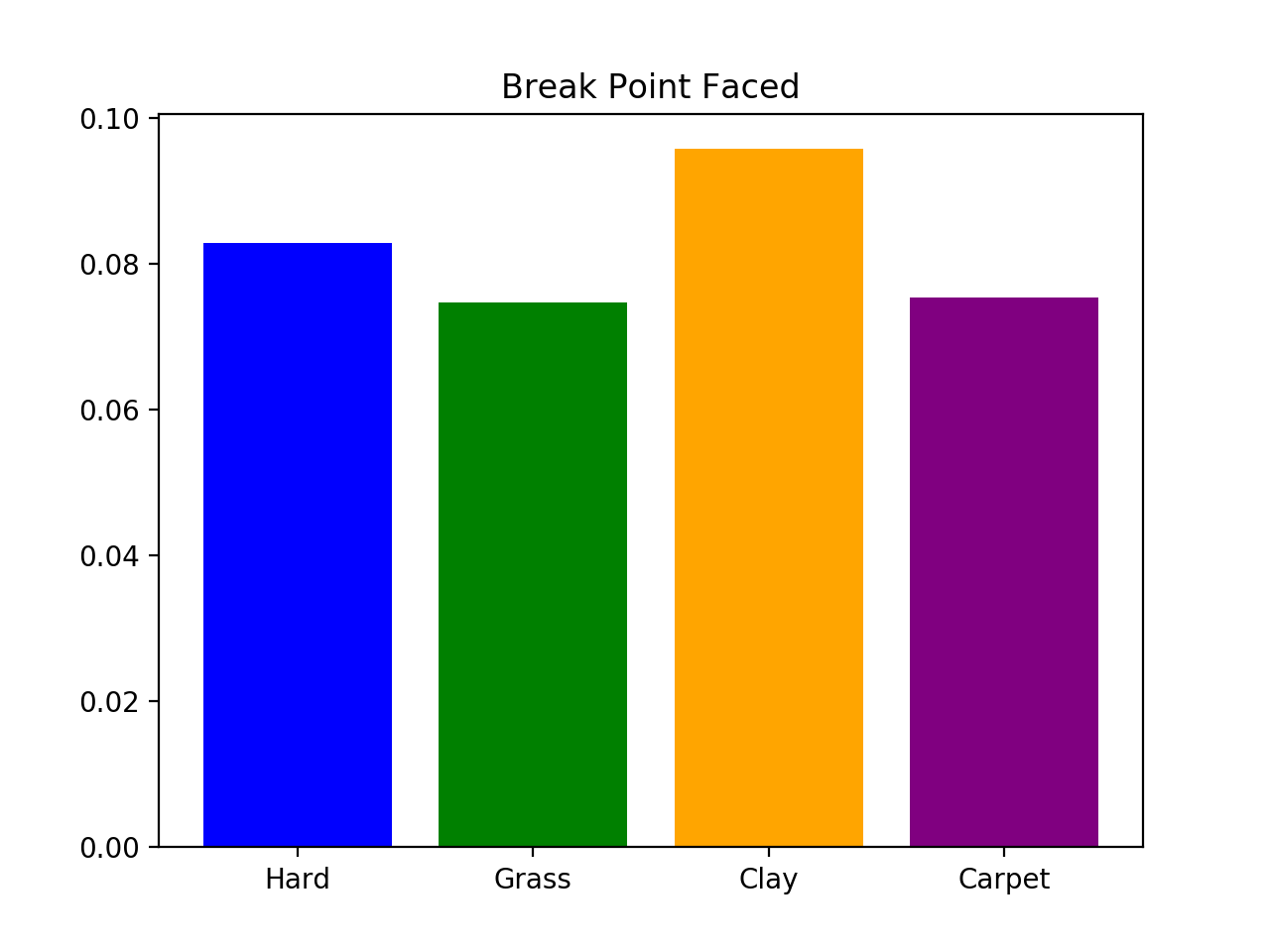
[3] Basic Statistics

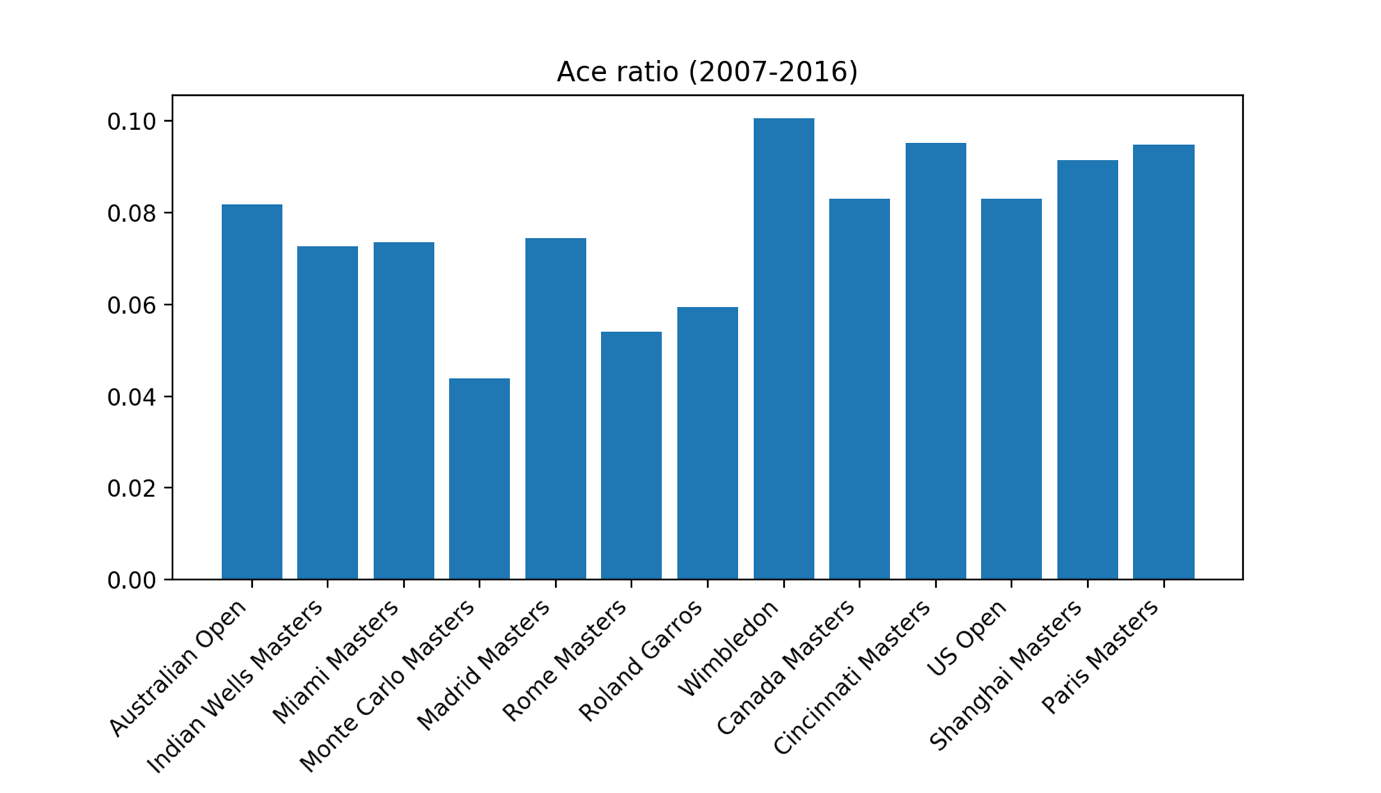
* The changes of surface types



* 平均而言，哪些场地最容易发出ACE？
* Averagely, how is the performance of ACEs on different types of surfaces?







The ace ratio (number of aces/total number of serves) in the recent 10 years show the generally average difficulties of serving an ace during each tournament. Wimbledon ranks the top, naming the tournament in which the aces happen most frequently. Although same as clay tournament as others, in Madrid Masters the players served more aces than in the other clay master tournaments.

* Who can do their 2nd serve in the way of 1st serve?

E1 = 1stIn \* 1stWon + (1-1stIn) \* 2ndIn\*2ndWon + 0

Where, 2ndIn = 1 – df/(1-1stIn)

E1 = 1stIn \* 1stWon + (1-1stIn-df) \* 2ndWon

E2 = 1stIn \* 1stWon + (1-1stIn) \* 1stIn \* 1stWon

If E2>E1, we can say that it is more worthy for the player to do their 2nd serve just the way as their first serve.

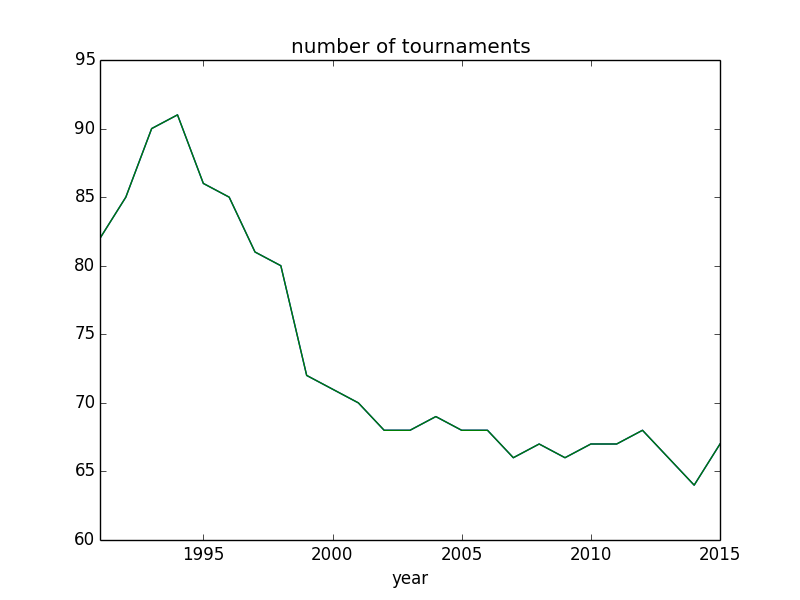
In some matches E2>E1 while in some other ones the opposite.

本文选取了1991年以来的比赛数据进行统计分析以及可视化。戴维斯杯的比赛不计入在内。共提取了83636（72640有效计分）场ATP单打比赛的数据。

1. 场地的变化

网球的场地总共有硬地（hard）、红土（clay）、草地（grass）、地毯（carpet）四种。

从1991年至今，一共举办了1832场赛事（不包含戴维斯杯，下同）。



一年的赛事数量总体呈现出下降的趋势，最高的在1994年达到一年有91项赛事。而进入到21世纪后，赛事数量降低并稳定在65-70项左右。