方向综合设计报告

一、 设计目的和内容

作为全球顶尖的职业篮球联盟，NBA 的职业性早已不局限于纯篮球领域。小到主教练的一次调遣换人，大到整支球队的球员交易和市场经营，一个好的决策往往是建立在长年累月且方法得当的计算、分析和模拟之上的。虽然无时不在产生的成吨的比赛数据提供给了我们丰富的分析资源，但也常常隐含歧义数据点令人难以察觉。因此，本项目结合数据可视化技术，探求一种多方位的数据分析手段，使数据和人类感官充分互动融合，在一定程度上降低个别离群信息点所带来的偏见，同时，创造更多机会发掘以往难以直观得出的结论。本项目着眼于一项强大的数据真实正负值，一个算法并未公开的，声称可以用于综合衡量一名篮球运动员的场上表现的数据。它的一个令人信服之处在于，每当一个赛季的 RPM 出炉，总会出现一些基础数据并不突出的球员，高居 RPM 榜前列，这似乎表明他们在场上的贡献不能简单填充基础数据单，但却可以有效地对球队赢球大局做出较高贡献。这看似可以许多数据分析师的猜想。本着对种类繁多的数据之间关系的好奇，我利用网络上公开的 NBA 数据，重点挖掘控球后卫之一位置每赛季的各项数据均值与一系列真实正负值以及球队表现之间的关系模式。

基于相关性较强的球员基础和高级数据，重点分析 NBA 后卫克里斯保罗的过去 14 个赛季中个人表现的变化，从侧边观察其个人能力发展轨迹，比赛风格变迁与变换球队搭档不同主攻手对其比赛风格所带来的变化，并与相近时期的顶级后卫进行对比。

我建立了一个动态网页，可视化地协助和呈现数据挖掘和分析结果。我主要使用了 python 的 lxml 和 requests 库从 www.basketball-reference.com和 http://www.espn.com/nba/statistics/rpm爬取了相关球员的历年数据，使用 pandas 库清洗数据，并作相关性分析处理，输出为 csv 文件存储在本地。在前端部分，使用了 HTML+CSS+JavaScript+D3.JS 库进行网站搭建。同时，学习储备了相关数据关联分析、数据可视化和人机交互的有关知识，指引我更好地完成本项目。

二、 背景知识

1. 真实正负值[1]，一部分是 RAPM，完全基于回合相关的正负值；一部分基于球员基础数据(box stat)，RPM 是继承自 RAPM(Regularized Adjusted Plus-Minus)，不过后者有统计回归的算法，前者没有公开。基本理念就是基于正负值和每一回合上场的 10 个球员来衡量单个球员的价值，会尽可能消除轮转和对手不同带来的影响，反映球员对比赛产生的影响力，不依赖于基础数据[2]。数据由 Jeremias Engelmann 提供，ESPN 展示。
2. 四大因素[3]，是从净得分（net score）得出的指标，这些指标与获胜的篮球比赛最密切相关。这些因素也可以确定团队的战略优势和劣势。团队的进攻和防守都可以应用四个因素，因此它给了我们八个因素。（1）投篮得分：有效投篮命中率 eFG%=（投篮命中数+0.5*三分投篮命中数）/投篮个数，注意投篮命中数已经包括三分投篮命中数在内；（2）保护球：失误率 TOV%=失误数/（投篮个数+0.44*罚篮个数+失误数）；（3）进攻篮板：进攻篮板 ORtg=进攻篮板/（进攻篮板+对手防守篮板）；（4）搏得罚篮机会：罚篮命中数/投篮个数，或者是罚篮个数/投篮个数。尽管这些是决定 NBA 胜利与失败的四个基本因素，但这些因素的重要性并不相同。 Dean Oliver 发现以下权重如下：（1）投篮（40%）（2）失误（25%）（3）篮板（20%）（4）罚球（15%）。可以看到，投篮是最重要的因素，其次是失误，篮板和罚球。每个统计数据衡量的是一项单独的技能，因此没有理由团队会表现出色，如果出现投篮少，丢失过多篮板球和很少能制造罚球的现象。同时，例如，在一个有着糟糕命中率的夜晚，一个球队可以通过在另一个方面的出色表现来弥补进攻方面的不足，进而弥补一个方面的糟糕表现。

三、 系统方案和实现

本系统分为数据处理和相关性分析，和前端可视化两大组成部分。

1. 首先要获取数据。NBA 的基础数据和大多数高阶数据在网上都有公开。经过比较分析，选定 www.basketball-reference.com 为主要获取数据的网站，因为它维护良好，门类齐全，格式统一，提供多种动态增删改数据的方式，方便批量获取。鉴于数据量较大，手工下载颇费功夫，因此借助 Python 的 lxml 和 requests 来处理 html 格式网页和网络请求，可统一从的<tr>，<td>标签

下批量获取数据值。另外，ESPN 提供的 RPM 数据较为权威，并且结果也完全公开。因此，我的 RPM 数据即采用 ESPN 官网公布的一系列 RPM，包括 ORPM 和 DRPM，以及 WINS(WIN SHARE)。爬取方法与第一个网站类似。由于 ESPN 的 RPM 在 13-14 赛季才首次公开，并且准确的赛季 RPM 应当在赛季结束后统一计算，赛季中前半段的 RPM 波动很大，因此我的相关分析是始于 2013-14 赛季，终于上一赛季 2018-19 赛季，故我所获取的基础数据和一些高阶数据也都是来自于相应的六个赛季。

2. 分析数据：（1）2013-14 赛季至 2018-19 赛季每百回合位置包括 PG（Point Guard，控球后卫）的 NBA 球员的数据，以及 RPM 数据。选择每百回合统计的数据，是因为真实正负值统计的结果也是对每百回合而言的。具体列表及备注[4]如下：

表 1 每百回合数据及 ESPN 系列正负值说明

缩写	名称	备注
PER	球员效率值	归一化后的球员每分钟产量，联盟平均值为 15
TS%	真实命中率	综合考虑两分球、三分球和发球的投篮效率评估
3PAr	三分球投篮数	三分球出手占总投篮数的百分比
FTr	罚球率	每次出手能得到的罚球数
ORB%	进攻篮板率	对球员在场上时能抢到进攻篮板的百分比的估计
DRB%	防守篮板率	对球员在场上时能抢到防守篮板的百分比的估计
AST%	助攻率	对球员在场上时队友通过其助攻所投中投篮所占百分比的估计
STL%	抢断率	对球员在场上送出抢断终结的对手回合所占对手所占总回合的百分比的估计
BLK%	盖帽率	对球员在场上送出的盖帽占对手两分球出手的

		百分比的估计
TOV%	失误率	百回合的失误数估计
USG%	球权使用率	对球员在场上时处理球的回合的百分比的估计
OWS	进攻胜利贡献值	
DWS	防守胜利贡献值	
WS	胜利贡献值	
WS/48	每48分钟胜利贡献值	联盟平均值为100
OBPM	进攻端正负值	
DBPM	防守端正负值	
BPM	正负值	
VORP	球员替换价值	球员在球场上起到的作用，与联盟该位置平均水平的球员的差异
ORPM	ESPN进攻真实正负值	
DRPM	ESPN防守真实正负值	
RPM	ESPN真实正负值	
RPM_WIN	ESPN胜利	包含了球员的真实正负值和他参与的回合

S	贡献值	

（2）投篮分时段效率评估数据：选择四名控球后卫（Steve Nash，Jason Kidd，Chris Paul，Stephen Curry）生涯中在分差 5 分之内的得分相关数据，包括 FG%，3P%和 eFG%，以及生涯每季场均相应数据。

3. Python 中提供了便捷的相关性分析的函数接口。由于数据项之间具为对称关系，在大样本下大部分 NBA 统计数据服从正态分布，我使用 pandas 下 corr（）函数计算部分数据之间的皮尔森相关系数，构建了一个静态 heatmap。由于需要在下一步中进行可视化交互，因此重构 DataFrame 格式，更改为每项数据由两个 NBA 数据项名称以及二者之间的相关性系数，共三个元素构成的一个 list，方便一一对应于将要画出的 Heatmap 中的每一个小方块。
4. 我所搭建的网页主要由三部分构成，分别是 HTML 静态网页布局，CSS 样式修饰文件，用于实现可视化及与交互的 JavaScript 文件，以及链接了一个进行动态数据可视化的 d3.js 库。

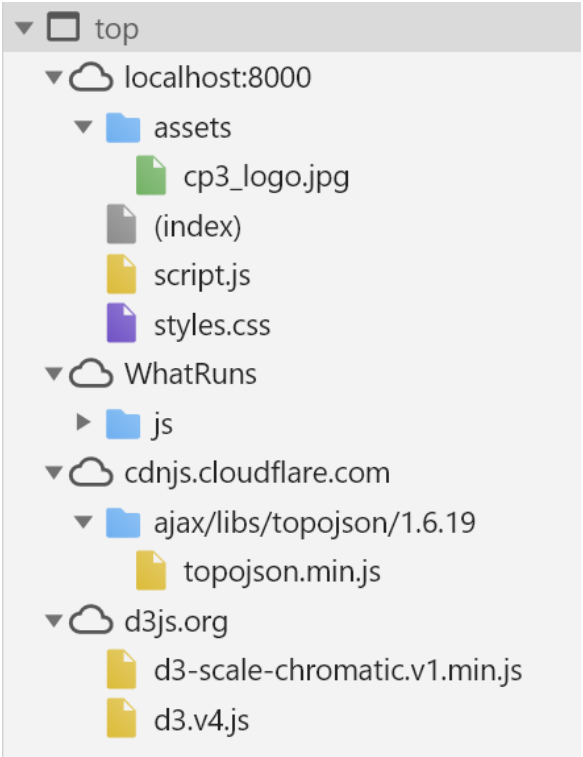

图 1 项目结构

5. 网页由以下几个部分组成：（1）一个信息顶栏，用于描述问题，进行总结；（2）一个热力图 Heatmap，包括浮动工具条 Tooltip，阈值拉动条 Slider，

以及菜单选择栏 Selector，用于相关分析可视化；（3）一个动态柱状图 Bar Chart，同样有多个球员可供显示选择，并且柱状图具有点击选中和显示数据功能，用于按时间序列展示各个球员的“助攻率-回合占有率”之比（AST%/USG%）的变化走势；（4）一个散点图 Scatter Plot，具有选中高亮功能和浮动工具条功能，用于比较各个球员的各个球员的“助攻率-回合占有率”之比，以及对应赛季的失误率（TOV%）；（5）一个折线图 Line Plot，用于按时间序列展示分析一名球员的比赛焦灼时刻的几项评估得分能力的数据和赛季相应的场均数据的比较。

6. Heatmap 可视化动态交互。Heatmap 是一个通过颜色分布使得用户可以直观地判断成对数据之间的相关关系的一类图。通过阅读文章，我发现人眼对形状，尤其是相似形状的大小关系判断尤为敏锐，因此在基础的 Heatmap 设计之上，改变了其每一个小方块的布局位置以及方块大小。同颜色分布同理，方块面积越大，说明这两项数据的线性相关性越强，，并且具有较深的颜色。因此，我们的人眼很容易从几百个小方块中识别出那些又大、又深的色块，再将鼠标移动当相应色块上方，便可通过直接阅读浮动条，知道是哪一对数据了。
7. 浮动工具条 Tooltip 的实现是为了方便用户寻找定位两个相应数据项，因为数据较多，排布密集。我定义了三个鼠标事件，分别是鼠标悬浮事件、鼠标移动事件以及鼠标移开事件，以实现随着鼠标动态滑过 Heatmap 方块区域，跟随展示相应数据项及其相关系数。
8. 添加了一个滑动条 Slider，实现了过滤低于某一相关性阈值的数据块。我的实现方法是同样添加输入响应事件，设置滑动条范围为 0 到 1，每当该输入响应事件被触发，比较各个方块相关性系数绝对值和当前滑动条所在处数值，若低于滑动条当前数值，则降低相应方块的遮盖度 Opacity。
9. 在热力图、柱状图和折线图中都实现了下拉选择框 Selector，可以选择不同年份、运动员或数据项用于显示。这几处下拉选择框的共性是都要在 JavaScript 中实现组件的更新，如大小，横、纵坐标位置，并且在更新时利用 d3.js 的 transition（）和 duration（）方法实现自动补充渐变效果。不同之处在于，在柱状图和折线图中，由于参与比较的几名运动员所处时期，或数据项的取值范围不是完全一致的（横坐标是根据具体赛季年份标定的，

而不是运动员的生涯第几个赛季；纵坐标为某项数据取值），故还需更新坐标轴，导致了有组件（长方块条、线条）增减的情况。此时，要先使用 remove（）删除无用组件，然后更新组件的相应属性 attr（）。如需要增加渐变效果，例如淡出，则可设置 transition（）前的遮盖度为 1，某个 duration（）之后为 0。

10. 在柱状图和散点图中，我使用了颜色和大小渐变。我添加了 https://d3js.org/d3-scale-chromatic.v1.min.js 用于线性的、分类别的颜色缩放标定，以突出球员的优势项，淡化弱势项。失误率 TOV%能在一定程度上评价球员的 AST%/USG%的含金量，因此，在散点图中对小圆点的大小（半径）根据对应球员、年份的 TOV%线性缩放，使得原本二维的图像包含了更为丰富的信息，得出的评价也可以更加客观。

四、 设计结果

1. 网页运行

   使用 Python 3 的 http.server 模块开启一个本地 Web 服务器，加载数据文件，使用 localhost：8000 访问 index.html。可以使用 ngrok，将本地的 Web 服务映射到外网一个临时分配的 IP 地址，供他人测试使用。

2. 总体界面展示

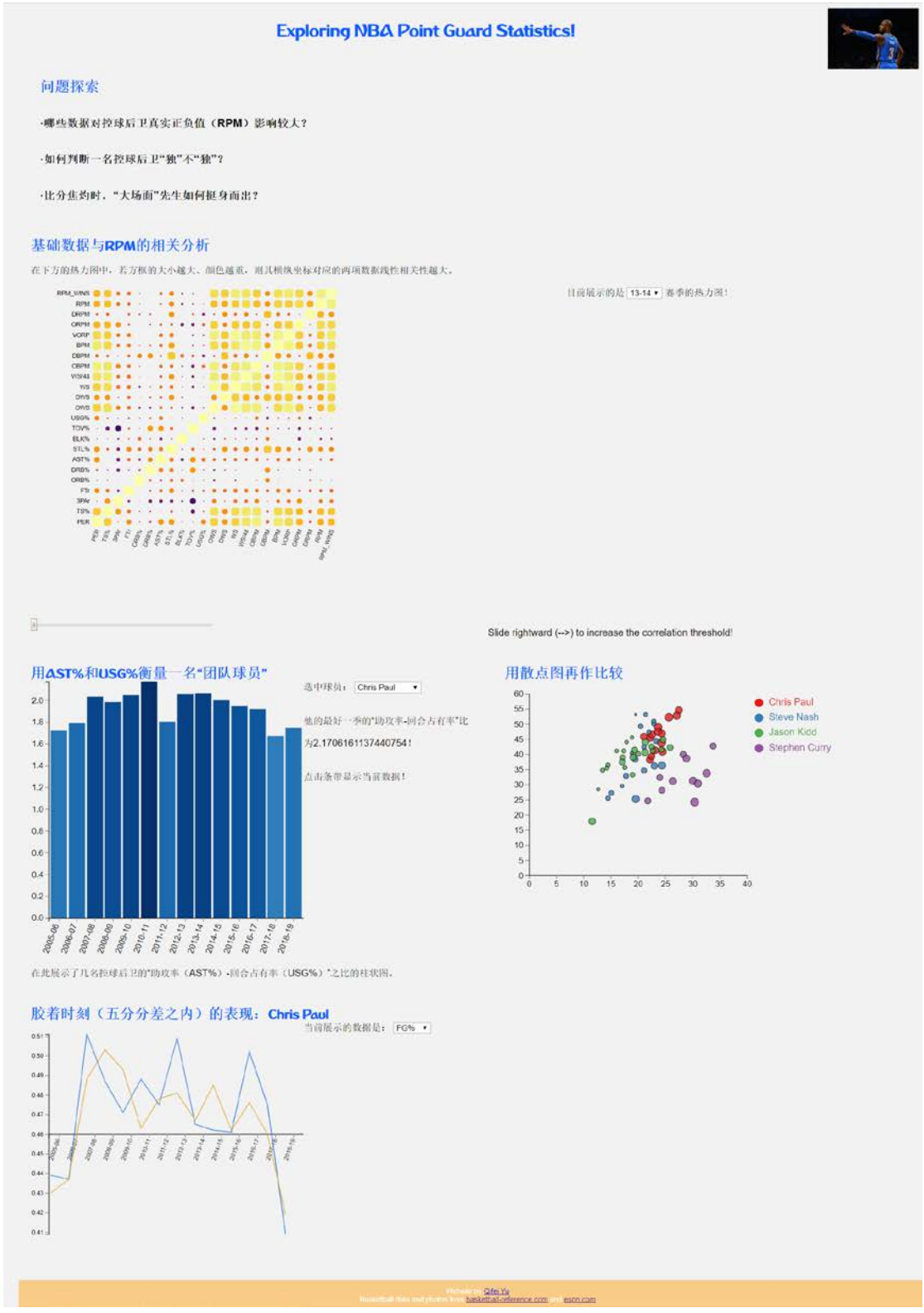

图 1 index.html 网页全景

3. 第一部分：基础数据与一系列真实正负值的相关分析

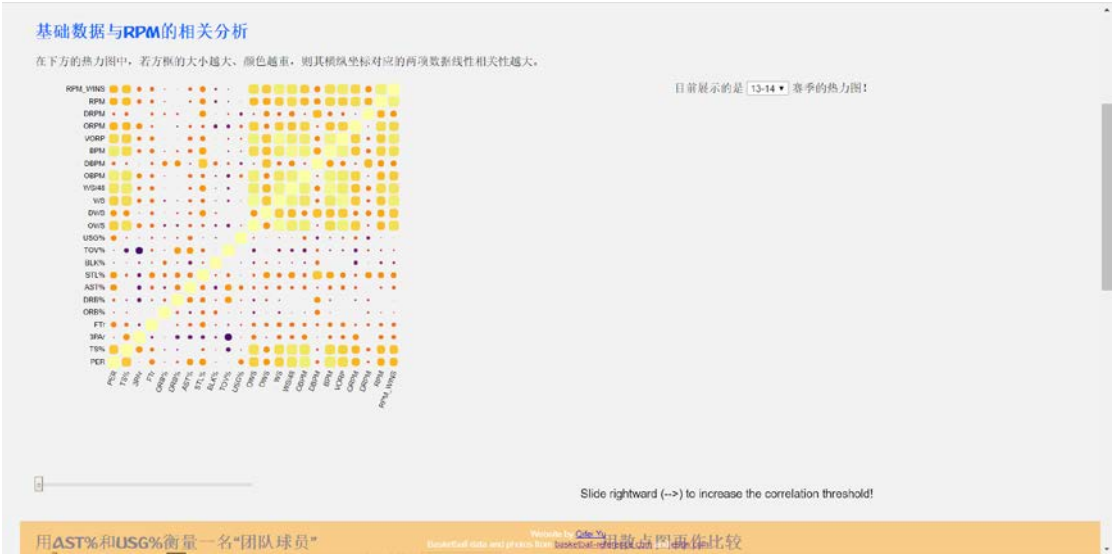

图 3 相关分析：Heatmap

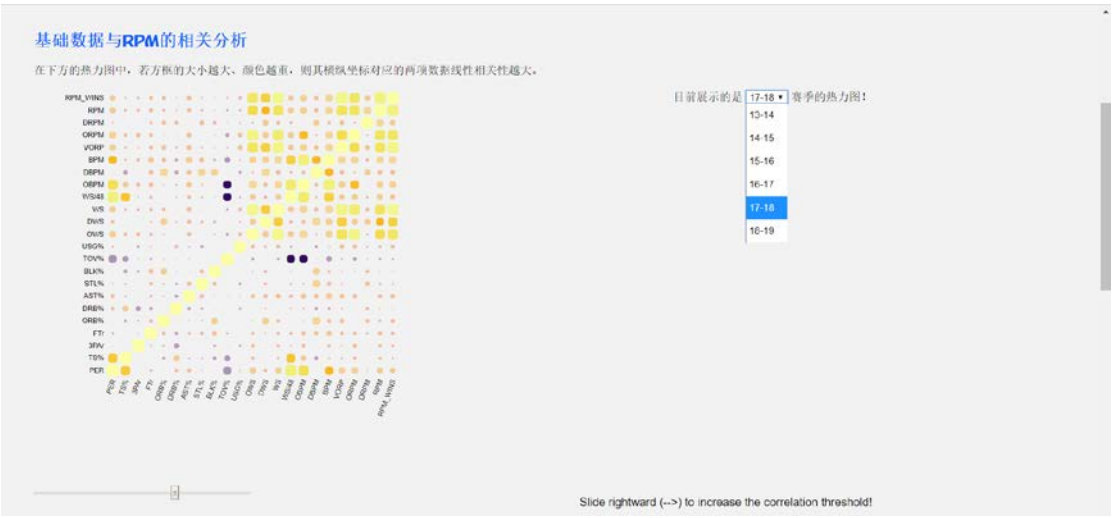

图 4 更换赛季显示，并滑动调整相关度阈值

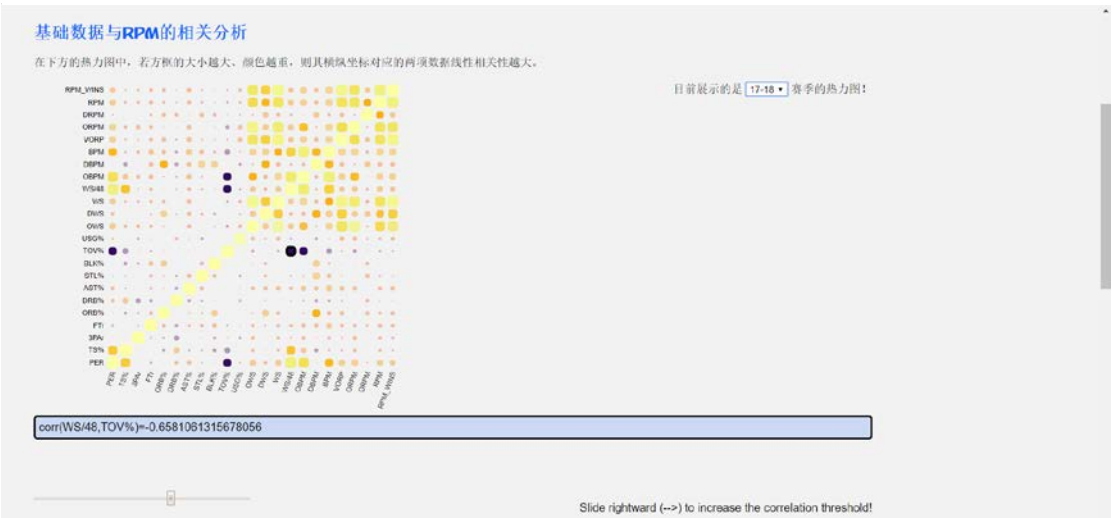

图 5 鼠标选择感兴趣的强相关性数据对

4. 第二部分：条状图和散点图，评估控球后卫的分享球表现

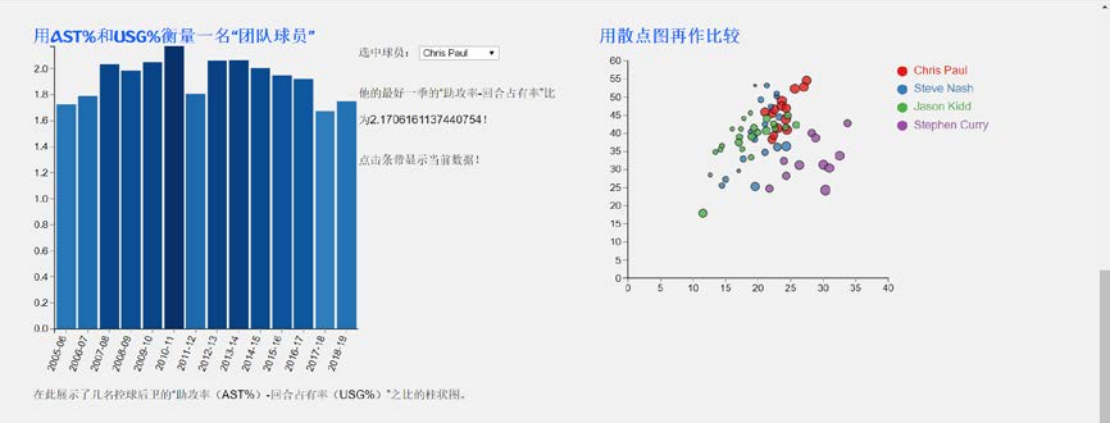

图 6 分享求表现与评价：Bar Chart 和 Scatter Plot

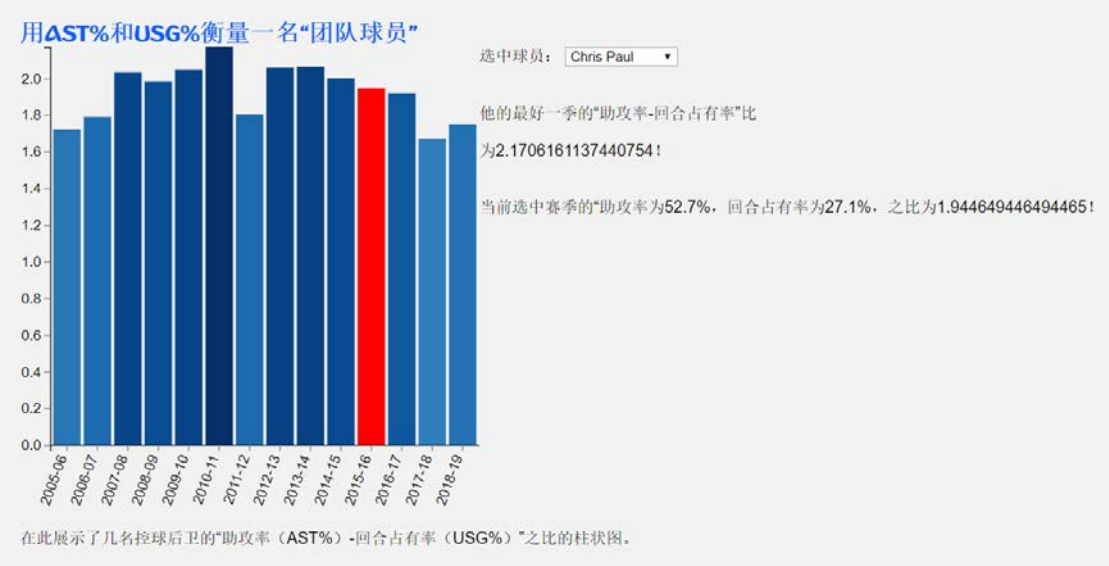

图 7 选中一个条块，展示数据

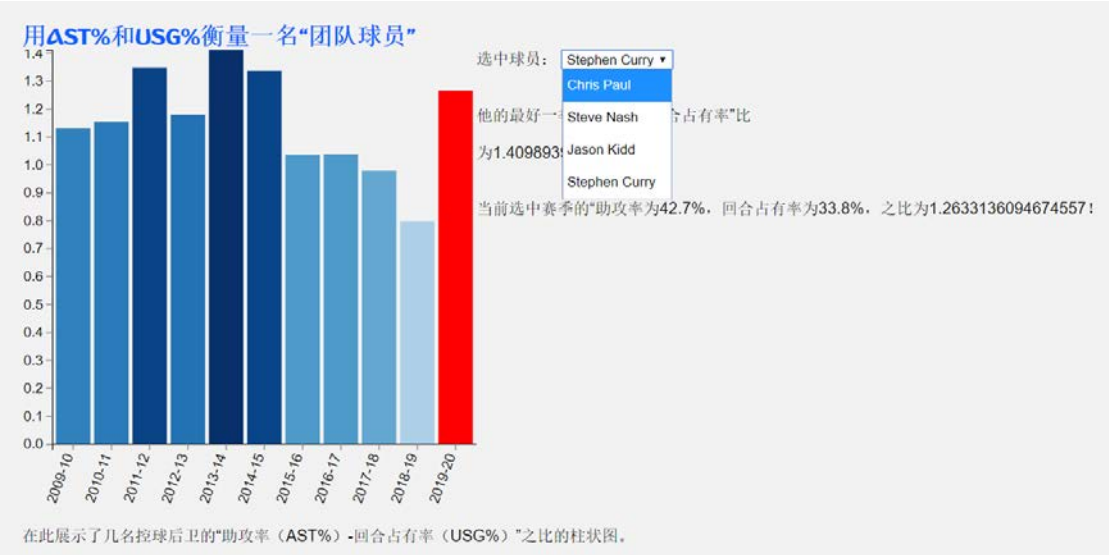

图 8 更换球员显示

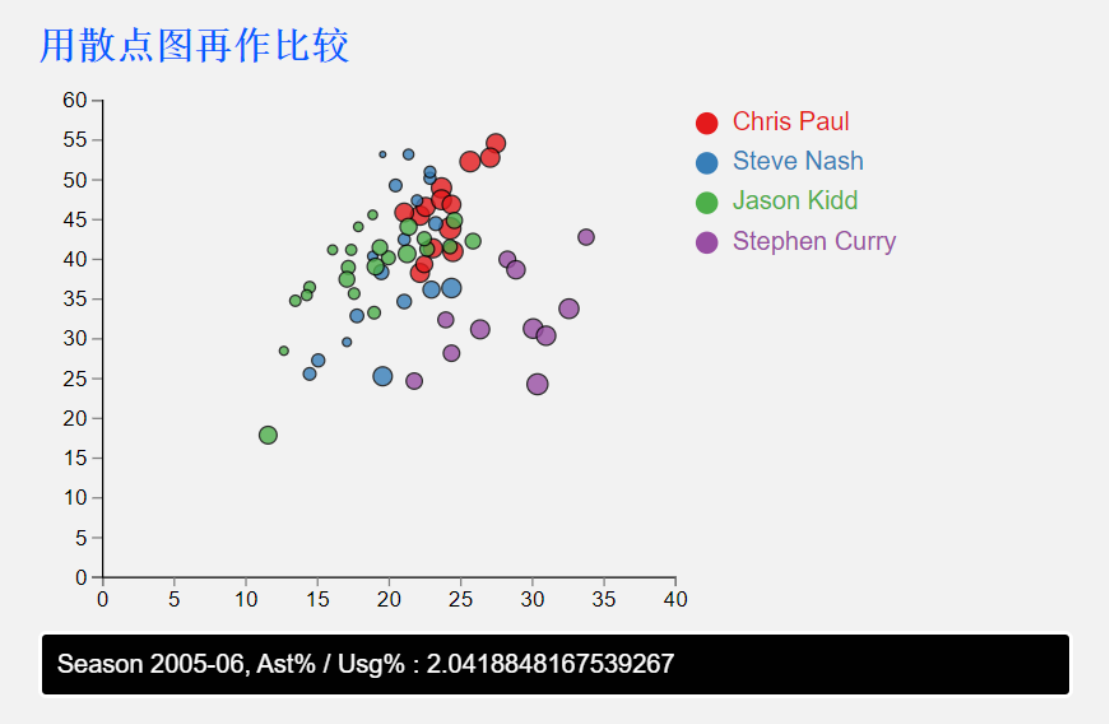

图 9 鼠标位于某一圆点上方，显示详细数据。圆点大小反映 TOV%表现，TOV%越大，失误出现的比例越高，圆点越小。较为理想的表现为，在不过分占有 USG%的情况下，AST%越高，TOV%越小，该球员传球意愿较强，且成功率较高。可视化呈现为偏右上角，且圆点面积较大。

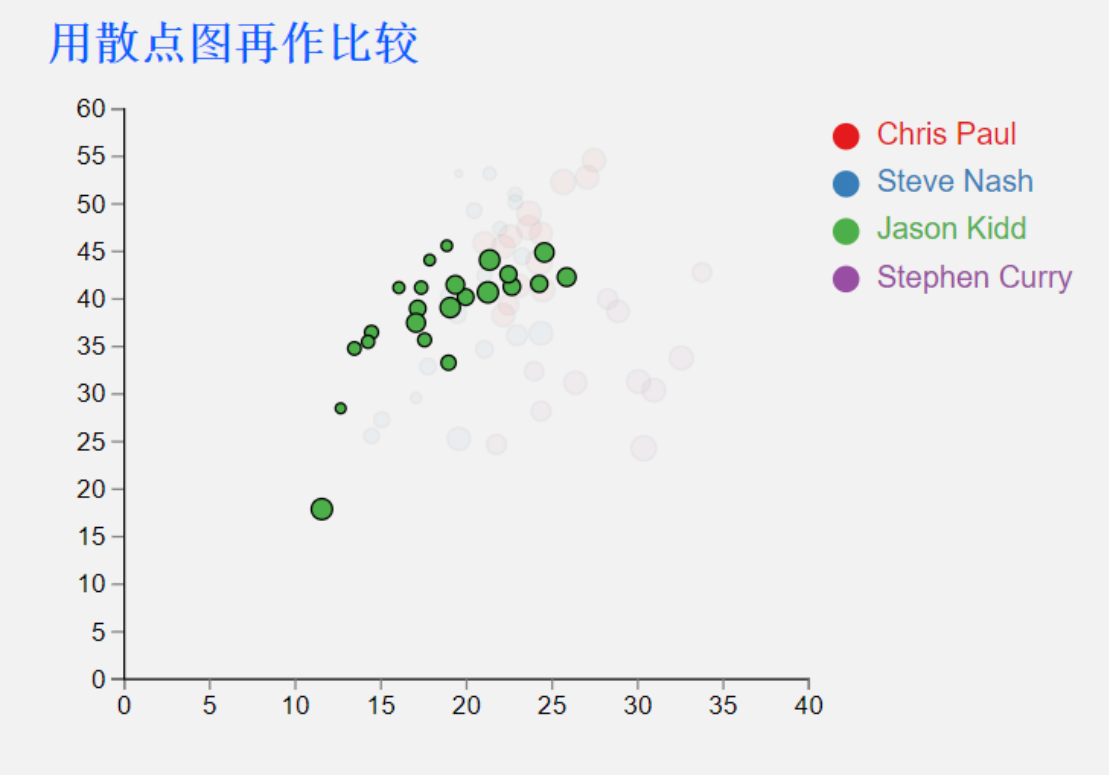

图 10 鼠标位于右侧圆点或标签上方，可以高亮显示当前球员数据而隐去其余球员

5. 第三部分：比分胶着时的得分表现

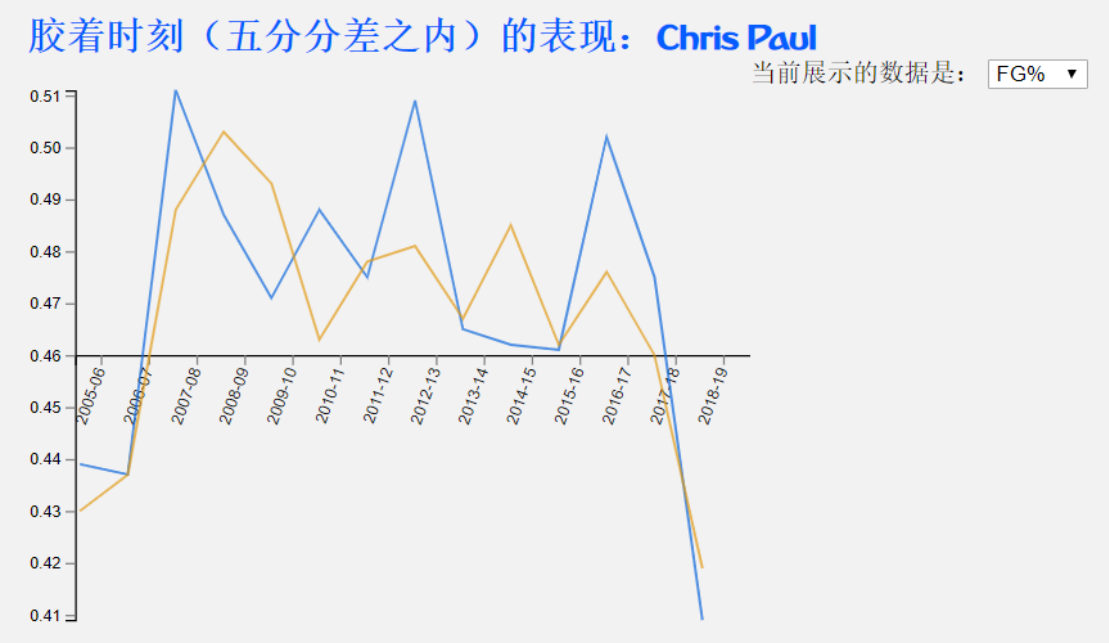

图 11 胶着时刻得分能力：Line Plot

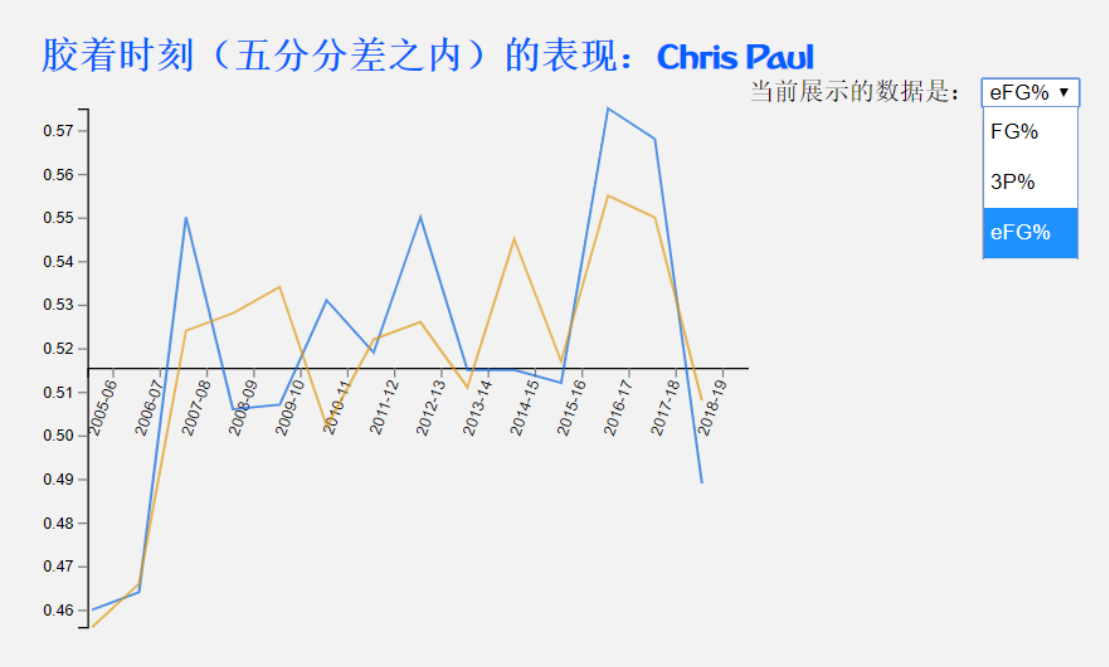

图 12 更改得分能力评估数据。通过与赛季场均数据比较，可以评价球员的关键时刻得分能力，并且可以观察其生涯的表现走势。

五、 总结与心得

通过完成本次设计，我借助可视化的方法，首先分析了控球后卫基础和高阶数据和 ESPN 真实正负值 RPM 的相关关系，发现除了赢球贡献值 WIN/48 和 RPM 有较强相关性之外，进攻端的数据，如 OBPM 相对防守端而言更能代表 RPM。同时发现，失误率 TOV%对控卫球员的赢球贡献值影响很大，可能的原因是后场的失误更容易被对手通过快攻转化为得分。因此，一些比较冒险的传控虽然有利于盘活球队进攻，但是若能控制失误率，会对球队赢球贡献极大。其次，我通过柱状图和散点图，分别从时间序和 TOV%角度，考察几名顶尖后卫的传球意愿和助攻能力。最后，我着重分析了一名控球后卫在比分胶着时刻的表现，因为一名好的控卫，不仅能在一般的时间段内，为状态出色的队友送出助攻，还可以在比分胶着，如队友频频打铁的危急时刻，挺身而出，化身得分手，打入关键球。从可视化结果中，可以观察到，该球员的关键时刻表现在几个赛季中比场均数据都更出色，杀手本色毕现。

我选择使用 HTML，CSS，JavaScript 和 d3.js 搭建一个动态网页的方式完成了可视化。这个实践过程让我拥有了前端实战经历，同时对可视化和人机交互技术有了更深刻的了解，在明确设计目标和考虑展示途径时，可以灵活的选择合适的方法，用最简洁、直观地方式向用户传达最关键的信息。

我意识到，篮球比赛是一个充满智慧的团队协作过程，有些贡献无法被数据记录，一些看似显然的数据可能也被各种因素影响，如出场阵容，比赛对手，和防守针对性等等。这次设计，我重点选择部分数据进行客观的挖掘分析，试图提供一个有趣的角度，玩味 NBA 数据。

六、 参考文献

[1]https://www.zhihu.com/question/27182727/answer/271206258

[2] https://zhuanlan.zhihu.com/p/50452888

[3]https://www.nbastuffer.com/analytics101/four-factors/

[4]https://www.basketball-reference.com/leagues/NBA_2014_per_poss.html