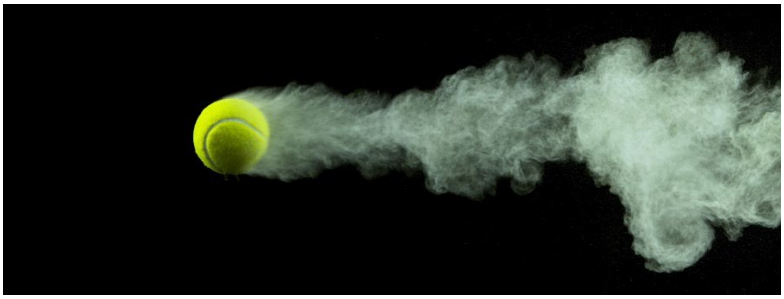
2024 MCM
问题 C：网球运动中的动量



在 2023 年温布尔登网球公开赛男子组决赛中，20 岁的西班牙新星卡洛斯-阿尔卡拉斯击败了 36 岁的诺瓦克-德约科维奇。这是德约科维奇自 2013 年以来首次在温布尔登输掉比赛，也结束了这位**大满贯**历史上最伟大球员之一的辉煌战绩。

[1]德约科维奇似乎注定会轻松获胜，他在第一盘以 6-1 的比分占据优势（7 局比赛中赢了 6 局）。然而，第二盘比赛却十分紧张，最终阿尔卡雷斯在决胜盘中以 7 - 6 获胜。第三盘与第一盘相反，阿尔卡拉兹以 6-1 的比分轻松获胜。第四盘开始后，年轻的西班牙人似乎完全控制了局面，但不知何故，比赛的走势再次发生了变化，德约科维奇完全控制了局面，以 6 - 3 的比分赢得了这一盘。第五盘也是最后一盘比赛开始后，德约科维奇延续了第四盘的优势，但比赛的走向再次发生了变化，阿尔卡拉斯以 6 - 4 取得了胜利。本场比赛的数据在提供的数据集中，"match_id "为 "2023-wimbledon-1701"。您可以使用 "set_no "列（等于 1）查看第一盘德约科维奇占优时的所有得分。 似乎占优的一方有时会出现多分甚至多局的惊人波动，这通常归因于 "势头"。

在字典中，"动量 "的定义是 "通过运动或一系列事件获得的力量或作用力。"[2] 在体育运动中，一支球队或一名球员可能会觉得他们在比赛中拥有动量或 "力量/作用力"，但很难衡量这种现象。此外，如果存在 "势 "的话，比赛中的各种事件是如何产生或改变 "势 "的，也不是一目了然的。

提供 2023 年温布尔登网球公开赛前两轮之后所有男子比赛中每一分的数据。您可以自行决定加入其他球员信息或其他数据，但必须完整记录数据来源。使用这些数据

- 建立一个模型，捕捉赛点发生时的比赛流程，并将其应用到一场或多场比赛中。您的模型应能确定哪位球员在比赛中的某个特定时间段表现更好，以及他们的表现好到什么程度。根据您的模型提供可视化的比赛流程描述。*注意：在网球比赛中，发球的一方赢得赛点/比赛的概率要高得多。您可能希望以某种方式将这一因素考虑到您的模型中。*
- 一位网球教练对 "势头 "在比赛中的作用持怀疑态度。相反，他假设比赛中的波动和一名球员的成功是随机的。请使用您的模型/度量来评估这一说法。

- 教练们很想知道，是否有一些指标可以帮助判断比赛的流程何时会从偏向一名球员变为偏向另一名球员。
    - 利用提供的至少一场比赛的数据，建立一个模型来预测比赛中的这些波动。哪些因素似乎最相关（如果有的话）？
    - 考虑到过去比赛 "势头 "波动的差异，您如何建议球员在新的比赛中对阵不同的球员？
- 在一场或多场其他比赛中测试您开发的模型。您对比赛中的波动预测得如何？如果模型有时表现不佳，您是否能找出未来模型中可能需要包含的任何因素？您的模型对其他比赛（如女子比赛）、锦标赛、球场表面和其他运动（如乒乓球）的通用性如何？
- 撰写一份不超过 25 页的报告，介绍您的研究结果，并附上一至两页的备忘录，总结您的研究结果，并就 "动力 "的作用以及如何让球员做好准备，应对网球比赛中影响比赛进程的事件，向教练提出建议。

您的 PDF 解决方案总页数不超过 25 页，其中应包括

- 一页摘要表。
- 目录
- 您的全套解决方案
- 一至两页的备忘录。
- 参考文献列表。
- 人工智能使用报告（如已使用，则不计入 25 页限制。）

注意：对于提交的完整材料，没有具体的最低页数要求。你可以用最多 25 页的篇幅来完成所有的解答工作，以及你想包含的任何附加信息（例如：图纸、图表、计算、表格）。我们接受部分解决方案。我们允许谨慎使用人工智能，如 ChatGPT，但没有必要为这一问题创建解决方案。如果您选择使用生成式人工智能，则必须遵守 COMAP 人工智能使用政策。这将导致一份额外的人工智能使用报告，您必须将其添加到 PDF 解决方案文件的末尾，并且不计入解决方案的 25 页总页数限制中。

提供的文件：

- ***Wimbledon_featured_matches.csv*** - 2023 年温布尔登网球赛第二轮之后的男子单打比赛数据集。
- ***data_dictionary.csv*** - 数据集说明。
- ***data_examples*** - 帮助理解所提供数据的示例。

术语表

大满贯网球大满贯是指在一个日历年度内赢得一个项目的全部四项主要锦标赛。四项大满贯赛事是澳大利亚网球公开赛、法国网球公开赛、温布尔登网球公开赛和美国网球公开赛，每项赛事为期两周。

关键术语/概念词汇表：

- **得分**：[3]
  - **比赛：**五局三胜制（温布尔登网球公开赛的男子组比赛）
  - **一局：一组**对局；6 局为一局，但棋手必须以两局的优势获胜，直到 6 - 6 打平时再进行决胜局（见下文）。
  - **游戏：**积分收集；玩家达到 4 分时获胜，但必须以 2 分获胜。见下文 "游戏得分"。

- **比赛得分**：[3]
  - 0 分 = 爱
  - 1 分 = 15
  - 2 分 = 30
  - 3 分 = 40
  - 并列得分 = 全部（例如 "30 全部）
  - 40 - 40 = 两分（双方赢得的分数相同，至少各得 3 分）。
  - 服务器赢得一个两分点 = Ad-in（或 "优势进入"）。
  - 接球手赢得两分 = Ad-out

- **发球**：球员交替担任 "发球员"（击球的球员）和 "回击球员"。在职业网球比赛中，发球员往往占有很大优势。选手有两次发球机会，在每一分上将球送入发球区。如果两次发球都未能将球送入发球区，则为 "双误"，发球员将获得这一分。
  - **破发**--回球选手赢得一局比赛。
  - **破发点** - 如果回击者获胜，他们将赢得比赛的点。
  - **保发**--发球方赢得比赛。

- **决胜局**：每局比赛在一方赢得 6 局后结束，只要他们至少领先两局（即 6 - 4）。否则，比赛继续进行，直到出现 6 - 6 的平局。此时将进行平局决胜。温布尔登网球赛的决胜局为先得 7 分（必须以 2 分优势获胜），但在 5th 盘比赛中，决胜局为先得 10 分（必须以 2 分优势获胜）。

- **休息时间/场地两侧**：球员在第一场比赛后交换场地两侧，然后每两场比赛后交换场地两侧。从第 3rd 局开始，每次换边都允许休息 90 秒。在决胜局中，球员每六分换边一次。每局比赛结束后，球员还需休息至少 2 分钟。允许医疗暂停和一次卫生间休息。

参考资料

[1] Braidwood, J. (2023), Novak Djokovic has created a unique rival - is Wimbledon defeat the beginning of the end, The Independent, https://www.independent.co.uk/sport/tennis/novak-djokovic-wimbledon-final-carlos-alcaraz-b2376600.html.

[2] https://www.merriam-webster.com/dictionary/momentum

[3] Rivera, J. (2023)，《网球计分解释：A guide to understanding the rules terms & point system at Wimbledon, The Sporting News, https://www.sportingnews.com/us/tennis/news/tennis-scoring-explained-rules-system-points-terms/7uzp2evdhbd11obdd59p3p1cx。

帮助理解数据集的示例

例 1：第 5 行

专栏	价值	说明
match_id	*“2023-wimbledon-1301”*	1301 "中的 "3 "表示第三轮比赛，"01 "表示该轮列出的第一场比赛。
逝去时间	*"0:01:31"*	比赛第一分开始后 1 分 31 秒，发球得分。
point_no、*game_no*、*set_no*（"no "是数字的缩写）。	4, 1, 1	所打的一分是比赛第 1st 盘第 1st 局的第 4th 分。
p1_sets, *p2_sets*, *p1_games*, *p2_games*	0, 0, 0, 0	由于这是比赛的第一局，双方都还没有赢过一局或一盘。
p1_score, *p2_score*	15, 30	打分时的比分是 15（1 号选手）比 30（2 号选手）。因此，棋手 1 赢得了之前的一分，棋手 2 赢得了两分。
服务器	1	球员 1（阿尔卡拉斯）在此点上发球。
serve_no	1	这一分是在第一个发球局进行的，这意味着阿尔卡拉斯的第一个发球局就在比赛中。
point_victor	1	阿尔卡拉斯赢得这一分（1 号选手）。
p1_points_won, p2_points_won	2, 2	1 号球员（阿尔卡拉斯）获得胜利，因此他在 本场比赛中的总得分现在是 2 分（之前是 1 分）。对于 玩家 2 的数值仍为 2，因为玩家 2 输掉了这一分。
game_victor, *set_victor*	0, 0	阿尔卡拉斯赢得这一分后，双方的比分为 30 - 30（各得 2 分），因此双方都没有赢得这一分（都 = 0）。
U - AC 栏		请允许我们确定是如何赢得这一分的：
p1_winner	1	阿尔卡拉斯以一记 "不可触碰 "的击球赢得了这一分。
p1_ace	0	该球不是发球（因为 = 0）。
获胜者拍摄类型	F	这一球是正手击球（而不是反手击球）。
p2_net_pt	1	球员 2（Jarry）在这一分中将自己定位在球网附近的某处。
p2_net_pt_won	0	由于阿尔卡拉斯赢得了这一分，虽然在这一分中 Jarry 在网前，但这个值为 0。

AH - AM 栏	全部 = 0	即使棋手 2 赢了这一分，比赛也不会结束，因此这一分不是 "破发点"，这些都是 0。
p1_distance_run, *p2_distance_run*	51.108, 75.631	每位球员在这一点上的跑动距离（以米为单位）。
集会次数	13	双方球员在这一分内投中的球数总和。
时速、*发球宽度*、*发球深度*、*回球深度*	130、BW、CTL、D	阿尔卡拉兹（发球员）在回球员的 "身体/宽度 "处（我们之前看到是一发）击出了一记 130 分的发球，并且接近表示进球或出球的界线。Jarry（回球员）在球场 "深处 "回球(所以靠近球场的另一端）。

例 2：第 8 - 12 行

第一局的最后四分说明了平分（"deuce"）和优势（"ad"）的概念。每一行都是比赛的后续时间点

。

行	专栏	价值	说明
第 8 行	*p1_score*, *p2_score*	40, 40	比分为 40 - 40，即双方各赢得 3 分（也称 "两分"）。
	point_victor	1	阿尔卡拉斯赢得第 7 点（第 8 行）。
第 9 行	*p1_score*, *p2_score*	公元 40 年	由于阿尔卡拉斯赢得了前一分（第 7 分），第 8 分的得分是 现在阿尔卡拉斯为 "AD"，雅里为 "40"，这意味着阿尔卡拉斯又赢了一分，并有可能在下一分赢得比赛。
	point_victor	2	Jarry（2 号选手）赢得第 8 点（第 9 行）。
第 10 行	*p1_score*, *p2_score*	40, 40	比分恢复为 40 - 40（"两分"），这意味着每位选手赢得的分数与之前相同，尽管现在是每人 4 分。
	point_victor	1	阿尔卡拉斯赢得第 9 点（第 10 行）。
第 11 行	*p1_score*, *p2_score*	公元 40 年	阿尔卡拉斯赢得第 9 点，再次占据优势。
	point_victor	1	阿尔卡拉斯赢得第 10 点（第 11 行），这意味着他赢得了比赛（现在又得了 2 点）。
第 12 行	*游戏编号*	2	这是第二场比赛的第一分。
	p1_games	1	阿尔卡拉斯赢得了第一场比赛。

例 3：第 51 行

比赛的 51st 点说明了 "破发点"--非发球方（回发球方）有机会赢得比赛的点。

行	专栏	价值	说明
第 51 行	*p1_score*, *p2_score*	40, 30	比分是 40 - 30，这意味着 1 号选手（阿尔卡拉斯）领先。
	服务器	2	Jarry（2 号选手）发球。
	p1_break_pt	1	如果阿尔卡拉斯赢得这一分，他将赢得比赛；由于他没有发球，这是一个 "破发点"。
	point_victor	1	阿尔卡拉斯赢得一分（也赢得了比赛）。
	p1_break_pt_won	1	阿尔卡拉斯赢得了这一局，并且在这一分上没有发球。

在 COMAP 竞赛中使用大型语言模型和生成式人工智能工具

这项政策的出台是由于大型语言模型（LLM）和生成式人工智能辅助技术的兴起。该政策旨在为团队、顾问和评委提供更大的透明度和指导。该政策适用于学生工作的各个方面，从模型的研究和开发（包括代码创建）到书面报告。由于这些新兴技术发展迅速，COMAP 将适时完善本政策。

参赛团队必须公开、诚实地说明他们对人工智能工具的所有使用情况。团队及其提交的材料越透明，其工作就越有可能得到他人的充分信任、赞赏和正确使用。这些信息的披露有助于了解智力工作的发展，也有助于对贡献给予适当的肯定。如果不对人工智能工具的作用进行公开、明确的引用和参考，有问题的段落和作品更有可能被认定为剽窃并取消资格。

解决这些问题并不需要使用人工智能工具，但允许负责任地使用这些工具。COMAP 认识到 LLM 和生成式人工智能作为生产力工具的价值，它们可以帮助团队准备提交材料；例如，为结构生成初步想法，或者在总结、转述、语言润色等方面。在模型开发的许多任务中，人类的创造力和团队合作是必不可少的，而依赖人工智能工具则会带来风险。因此，我们建议在使用这些技术进行模型选择和构建、协助创建代码、解释数据和模型结果以及得出科学结论等任务时谨慎行事。

必须指出的是，LLM 和生成式人工智能有其局限性，无法取代人类的创造力和批判性思维。如果团队选择使用 LLM，COMAP 建议他们意识到这些风险：

- 客观性：以前发表的含有种族主义、性别歧视或其他偏见的内容可能会出现在 LLM 生成的文本中，一些重要的观点可能无法体现。
- 准确性：LLM 可能会产生 "幻觉"，即生成错误的内容，尤其是在其领域之外或在处理复杂或模糊的主题时。它们可能会生成语言上可信但科学上不可信的内容，可能会弄错事实，还可能生成不存在的引文。有些 LLM 只针对特定日期前发布的内容进行训练，因此呈现的内容并不完整。
- 语境理解：LLM 无法将人类的理解力应用到文本的上下文中，尤其是在处理成语表达、讽刺、幽默或隐喻性语言时。这会导致生成的内容出现错误或曲解。
- 训练数据：LLM 需要大量高质量的训练数据才能达到最佳性能。然而，在某些

领域或语言中，此类数据可能并不容易获得，从而限制了任何输出结果的实用性。

团队指南

参赛队必须

1. **请在报告中明确说明使用了 LLM 或其他人工智能工具，**包括使用了哪种模型以及用 于何种目的。请使用内联引文和参考文献部分。同时，在 25 页的解决方案后附上人工智能使用报告（如下所述）。
2. **核实**语言模型生成的内容和任何引文的**准确性、有效性和适当性，**并纠正任何错误或不一致之处。
3. **按照此处提供的指导提供引文和参考文献。**仔细检查引文，确保其准确无误，并正确引用。
4. 由于法学硕士可能会从其他来源复制大量文字，因此**要注意抄袭的可能性**。检查原始资料来源，确保您没有剽窃他人的作品。

COMAP 将采取适当行动
当我们发现提交材料可能是在未披露使用此类工具的情况下
编写的。

引用和参考文献说明

无论团队选择使用什么工具，都要仔细考虑如何记录和引用。各种风格指南都开始纳入人工智能工具的引用和参考政策。请使用内联引文，并在 25 页解决方案的参考文献部分列出所有使用过的人工智能工具。

无论团队是否选择使用人工智能工具，主要解决方案报告的页数仍限制在 25 页以内。如果团队选择使用人工智能，请在报告末尾添加一个新部分，标题为 "人工智能使用报告"。这一新部分没有页数限制，也不会算作 25 页解决方案的一部分。

举例说明（*并非*详尽无遗，请根据具体情况进行调整）：

关于人工智能使用情况的报告

1. OpenAI *ChatGPT*（2023 年 11 月 5 日版本，ChatGPT-4）
   Query1: <*插入您输入到人工智能工具中的准确措辞*>
   Output: <插入*人工智能*工具*的完整输出结果*>。

2. OpenAI *Ernie*（2023 年 11 月 5 日版本，Ernie 4.0）
   查询 1：<插入*随后输入到人工智能工具中的任何内容的准确措辞*> 输出：<插入*第二个*查询*的完整输出内容*>。

3. Github *CoPilot*（2024 年 2 月 3 日版本）
   Query1：<请*输入您输入人工智能工具的准确措辞*> Output：<请输入*人工智能*工具*的完整输出结果*>。

4. 谷歌*巴德*（2024 年 2 月 2 日版）
   查询：<请输入*查询的准确措辞*> 输出：<请输入*人工智能*工具*的完整输出信息*> 查询：<请输入*查询的准确措辞*> 输出：<请输入*人工智能工具的完整输出信息*