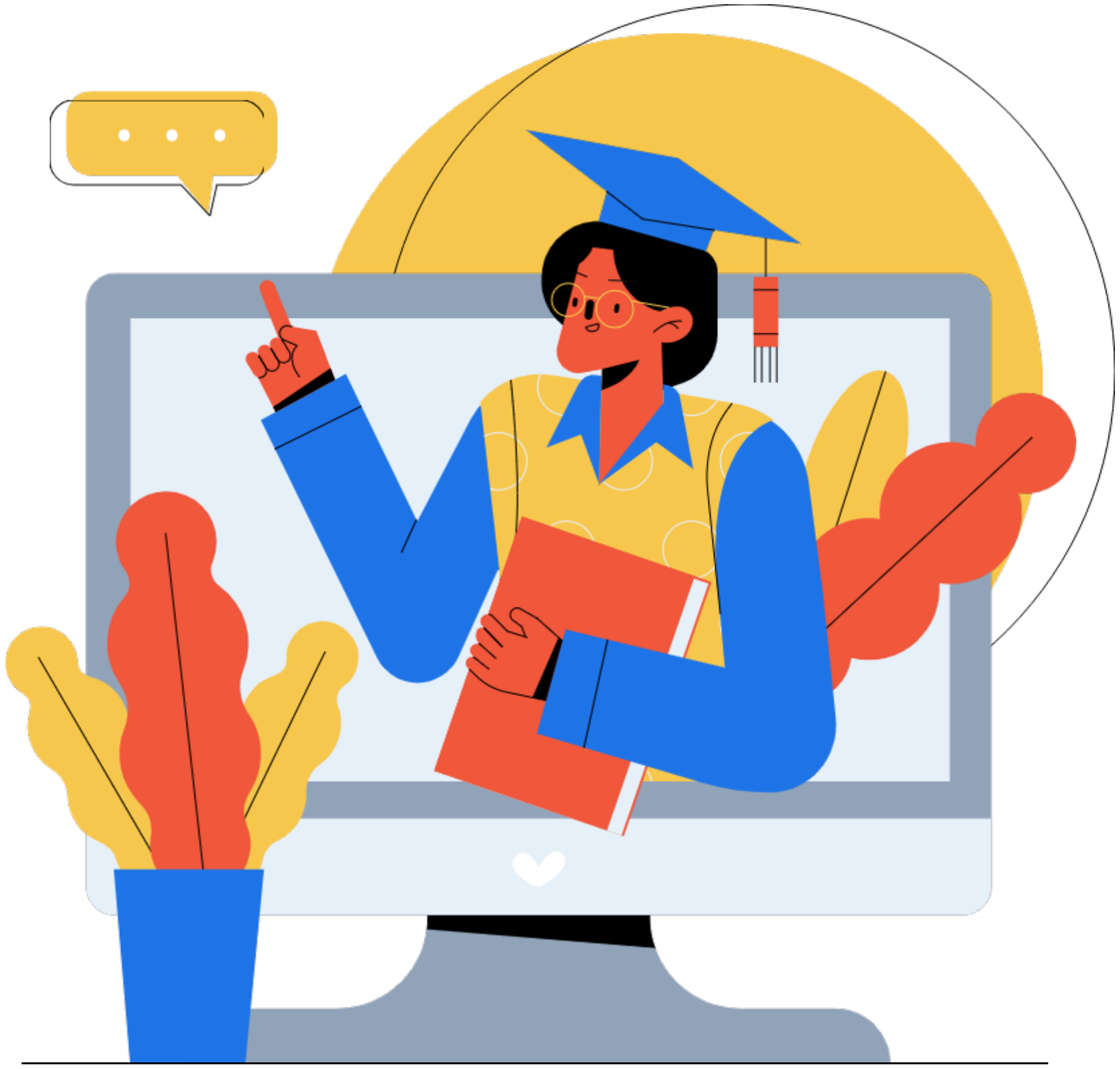
Coggle数据科学

参赛趋势 & 比赛汇总

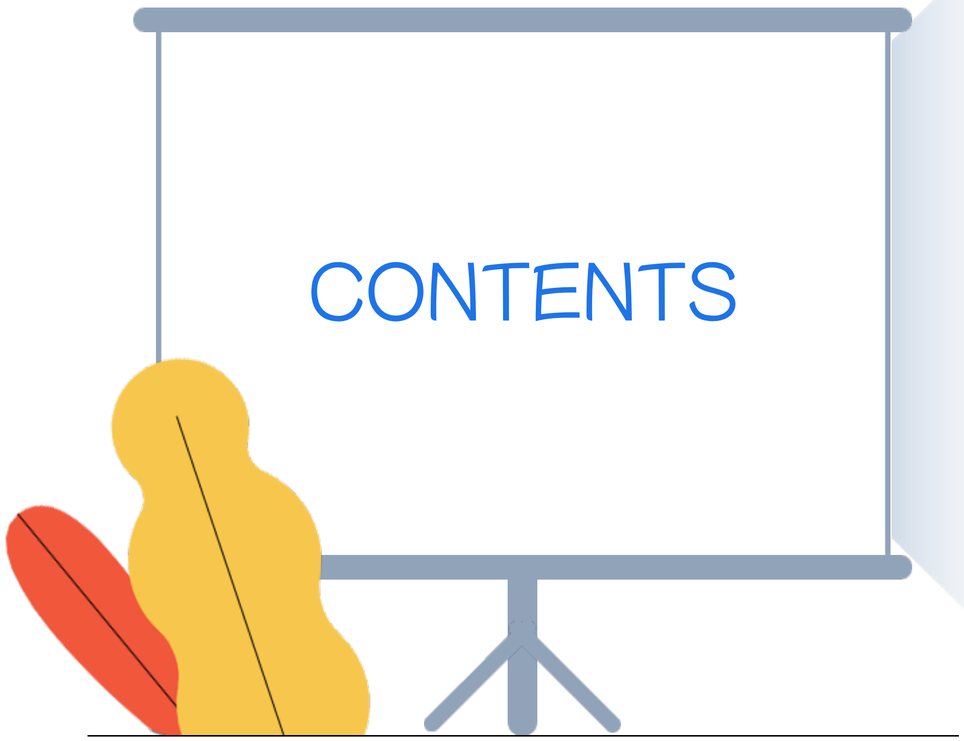
Kaggle年鉴

2022

Coggle数据科学



Coggle数据科学



PART 01

Kaggle年鉴介绍

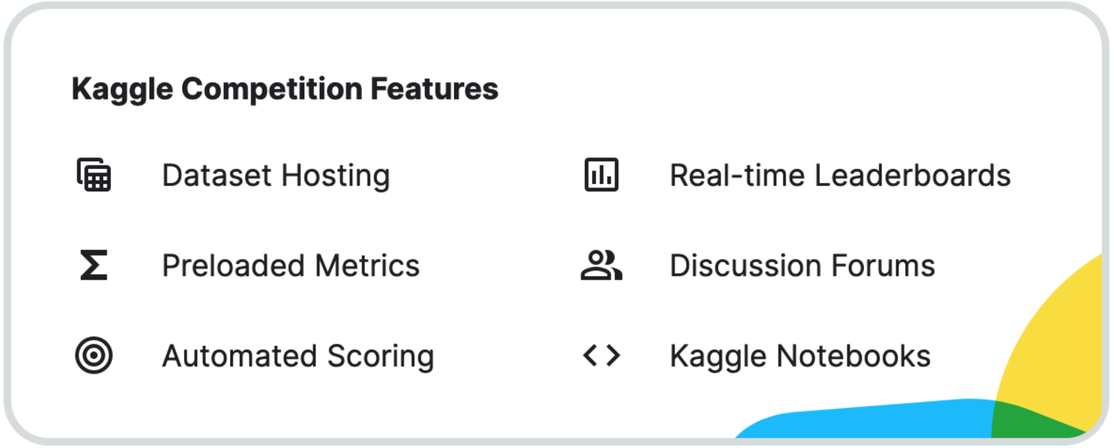
Part1 Kaggle年鉴介绍

Coggle数据科学

Kaggle比赛平台每年都会举办 N 多有价值的比赛，吸引了全球的用户参与。但作为参赛选手的你，你知道今年Kaggle最热门的比赛吗？今年最常见的比赛库是什么？

Kaggle年鉴内容（2022年度）：

- ✓ 比赛类型统计
- ✓ 参赛选手统计
- ✓ 模型&库统计
- ✓ 比赛内容统计



Part1 Kaggle年鉴介绍

Coggle数据科学

我们是「Coggle数据科学」

- ✓ 国内规模最大的Kaggle竞赛社群
- ✓ 全网最全的竞赛资讯汇总
- ✓ 每周不定期知识分享 & 学习活动

Coggle 愿景:

- ✓ Communication for Kaggle
- ✓ 让知识和技术改变世界

微信搜一搜

Coggle数据科学

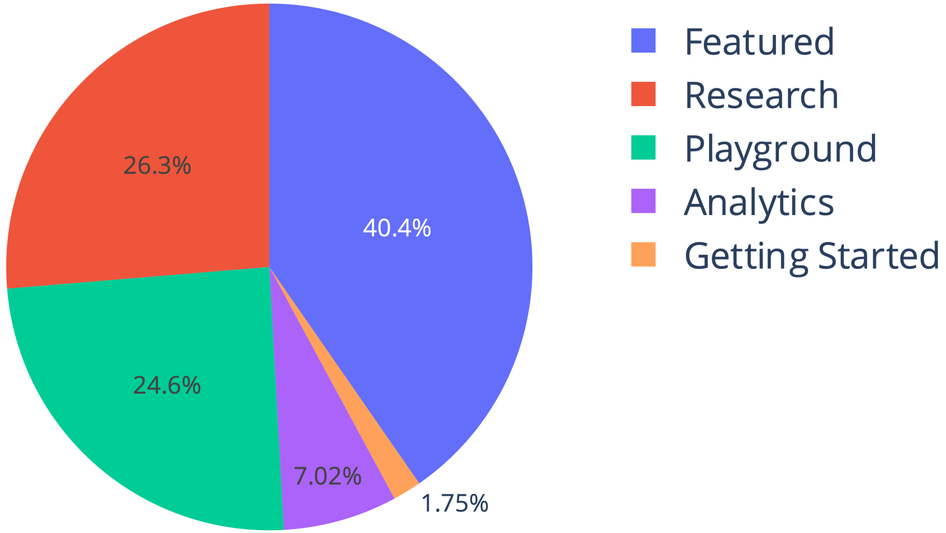
PART 02

比赛类型统计

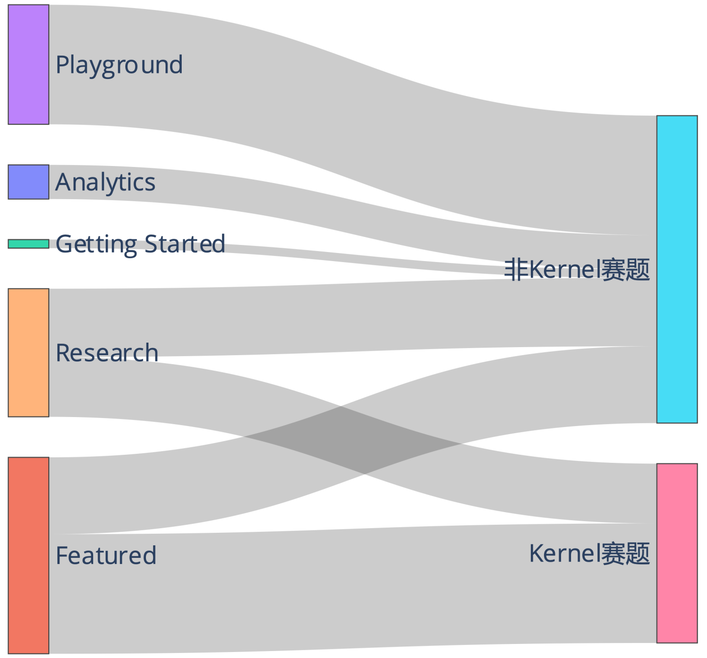
Part2 比赛类型统计

Coggle数据科学

2022年度举办57场比赛，共吸引了全球6万人次参加，总共提交方案84万次，总奖金162万美元。



- Kernel赛题：需通过Notebook提交的比赛
- 非Kernel赛题：通过Notebook & 文件提交的比赛



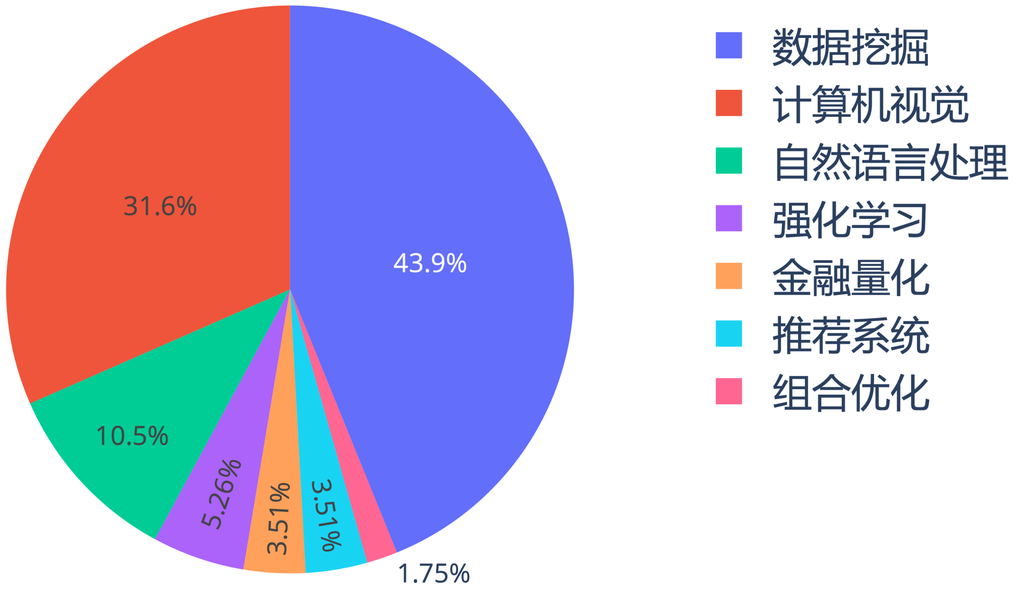
- Feature： 工业赛赛题，难度较大
- Research：学术赛题，难度较大
- Playground：练习赛，难度适中
- Analytics：数据分析赛
- Getting Started：入门赛，难度较低

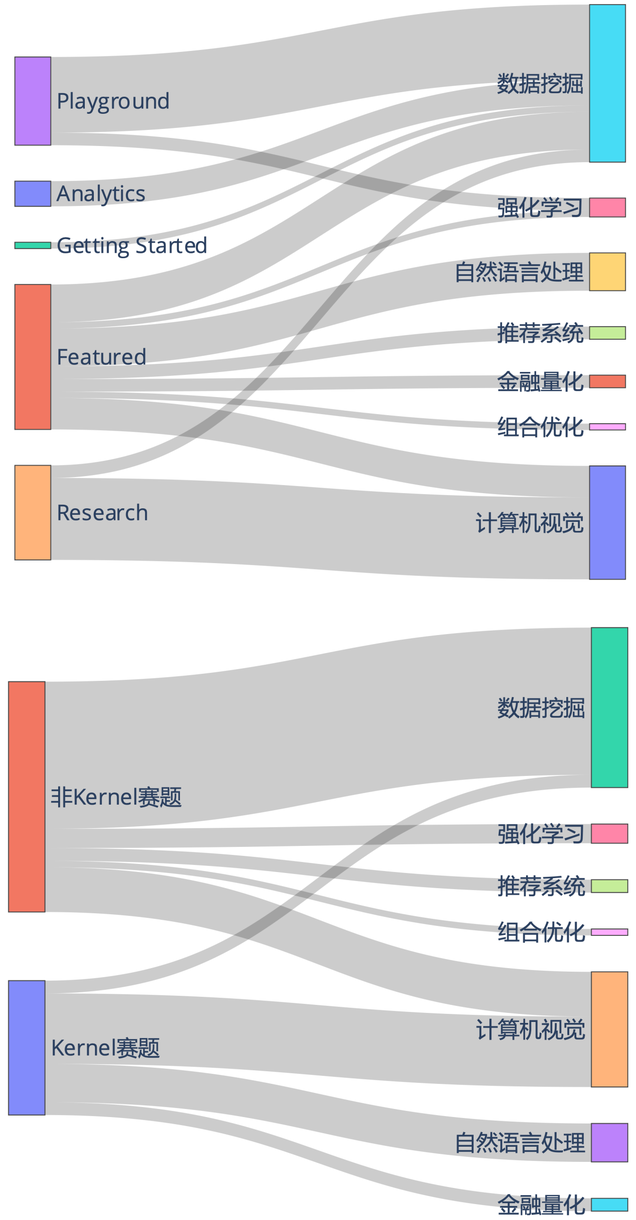
https://www.kaggle.com/docs/competitions
https://www.kaggle.com/docs/notebooks

Part2 比赛类型统计

Coggle数据科学

今年Kaggle比赛按照比赛任务可划分有7个方向，其中数据挖掘和计算机视觉占比较多。



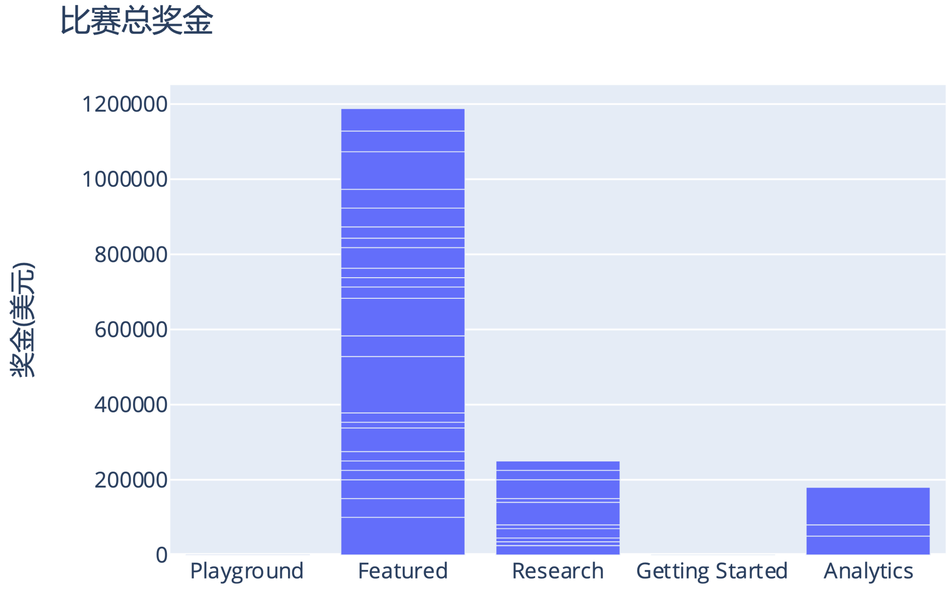


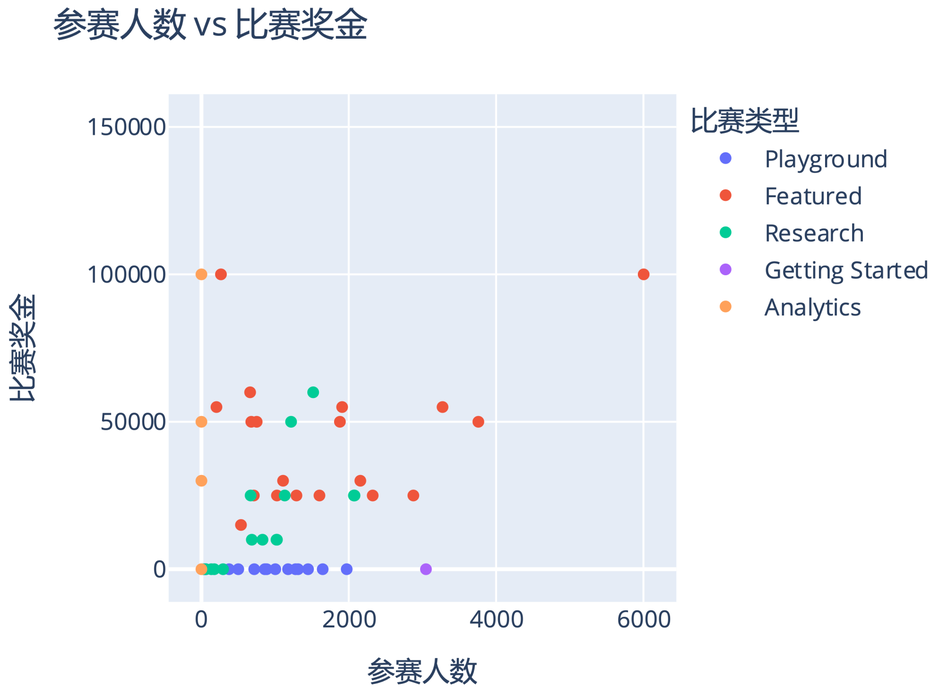
- ✓ 大部分数据挖掘赛题来自Playground，没有比赛积分和奖牌
- ✓ 部分视觉比赛是Kernel赛题，而所有文本赛题是Kernel赛题
- ✓ 今年金融量化比赛比往年多，但语音识别比赛比往年少
- ✓ 今年视觉赛题为语义分割赛题居多，纯分类赛题较少

Part2 比赛类型统计

Coggle数据科学

- ✓ Featured和Research类型的比赛奖金比较多
- ✓ Kernel赛题奖金比较多，因为Kernel赛题一般为Featured和Research类型
- ✓ 比赛奖金越多，参赛人数也越多
- ✓ Featured类型的比赛参赛人数最多，其次是Research类型的比赛

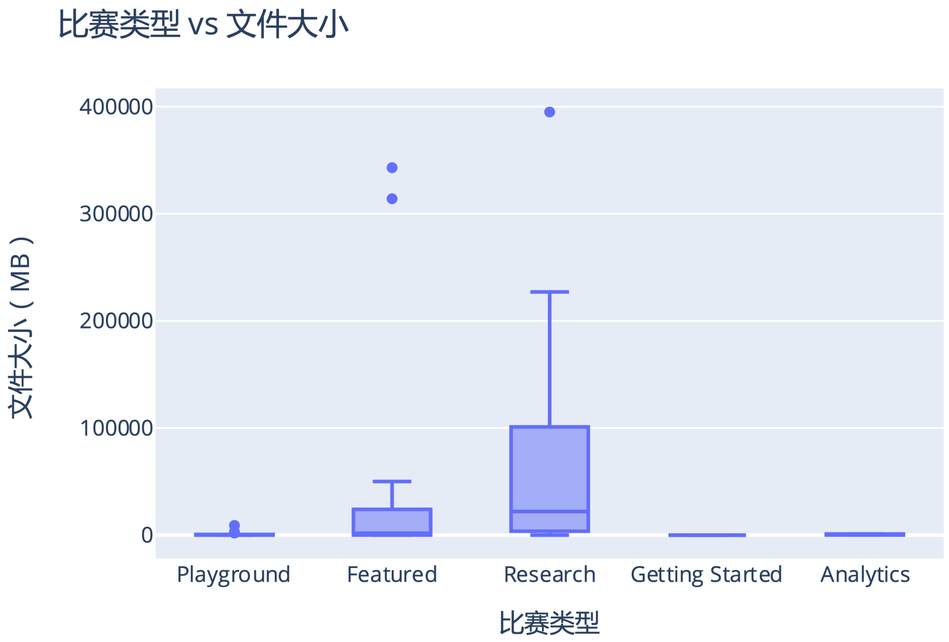


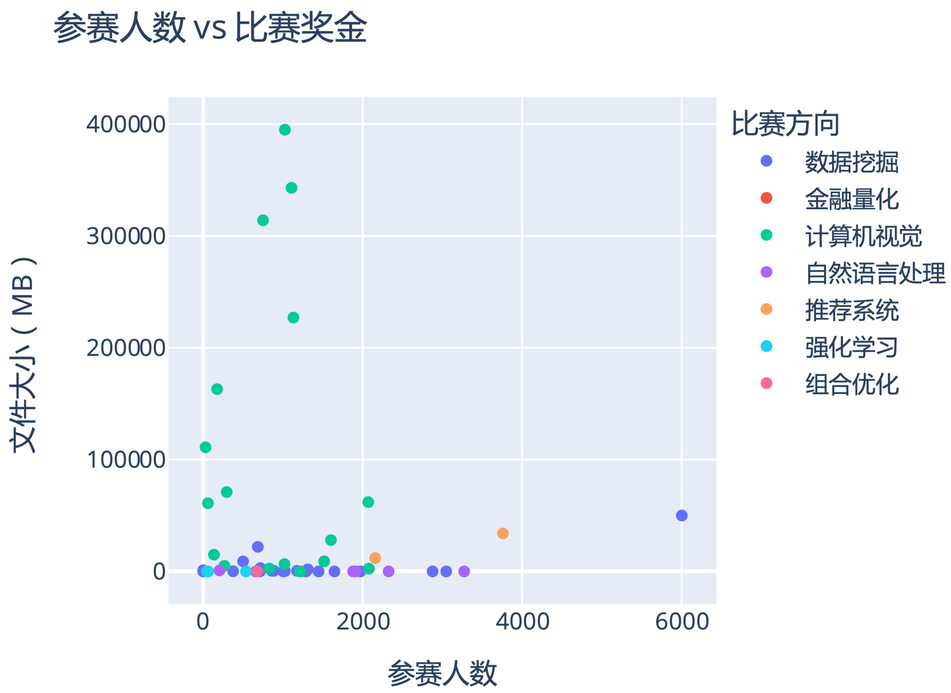


Part2 比赛类型统计

Coggle数据科学

- ✓ 大部分比赛文件在100GB以内，且入门类型比赛文件都比较小
- ✓ Research类型比赛文件最大，其次是Featured类型比赛
- ✓ 按照文件大小排序：计算机视觉 > 数据挖掘 > 自然语言处理
- ✓ 比赛文件越大，参赛人数越少





Part2 比赛类型统计

Kaggle年度热门比赛

比赛名称	比赛方向	参赛人数	提交次数
Ubiquant Market Prediction	金融量化	1646	16151
Santa 2022 - The Christmas Card Conundrum	组合优化	1312	11766
Kore 2022	强化学习	1003	9971
H&M Personalized Fashion Recommendations	推荐系统	860	7235
UW-Madison GI Tract Image Segmentation	计算机视觉	1176	8902
Feedback Prize - English Language Learning	自然语言处理	886	5984
American Express - Default Prediction	数据挖掘	1278	16346

Part2 比赛类型统计

Coggle数据科学

Tabular Playground Series 年度系列比赛

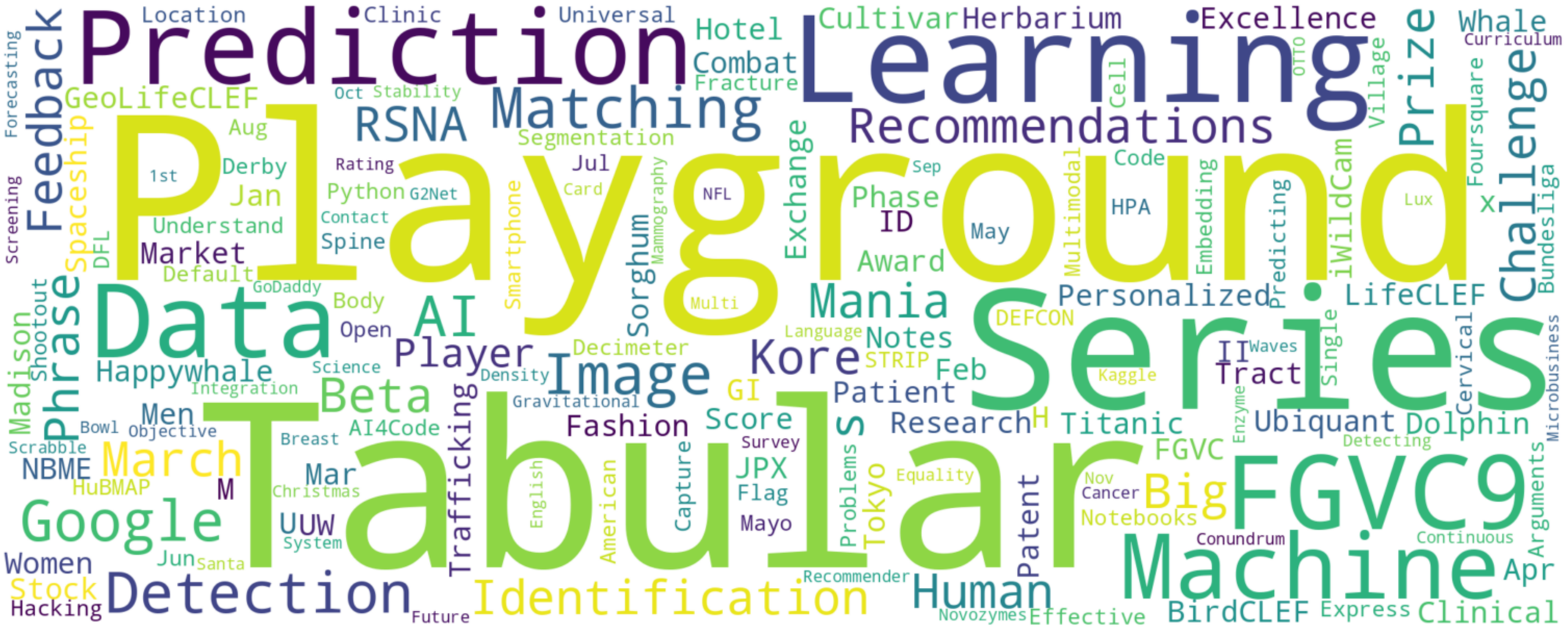
比赛名称	比赛任务	比赛难度	参赛人数	提交次数
Tabular Playground Series - Jan 2022	销量预测	★★	1646	16151
Tabular Playground Series - Feb 2022	多分类	★★	1312	11766
Tabular Playground Series - Mar 2022	时序回归	★★	1003	9971
Tabular Playground Series - Apr 2022	二分类	★★	860	7235
Tabular Playground Series - May 2022	二分类	★★	1176	8902
Tabular Playground Series - Jun 2022	缺失值填充	★★★	886	5984
Tabular Playground Series - Jul 2022	无监督聚类	★★★	1278	16346
Tabular Playground Series - Aug 2022	二分类	★★	1972	21790
Tabular Playground Series - Sep 2022	销量预测	★★	1447	13085
Tabular Playground Series - Oct 2022	二分类	★★	500	4659
Tabular Playground Series - Nov 2022	二分类	★★	717	7260

Part2 比赛类型统计

Coggle数据科学

2022年Kaggle比赛年度关键词：Playground

从2021年开始Kaggle月会举办Playground Series比赛，比赛数据都为结构化数据，但每次任务不同。



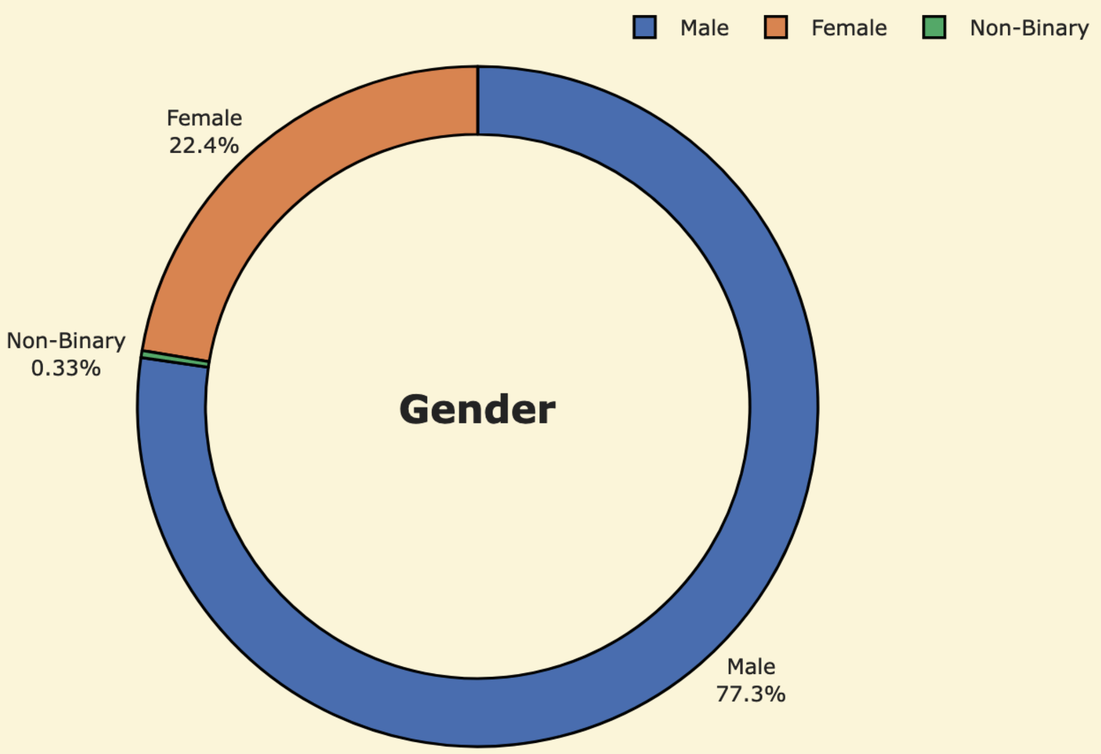
PART 03

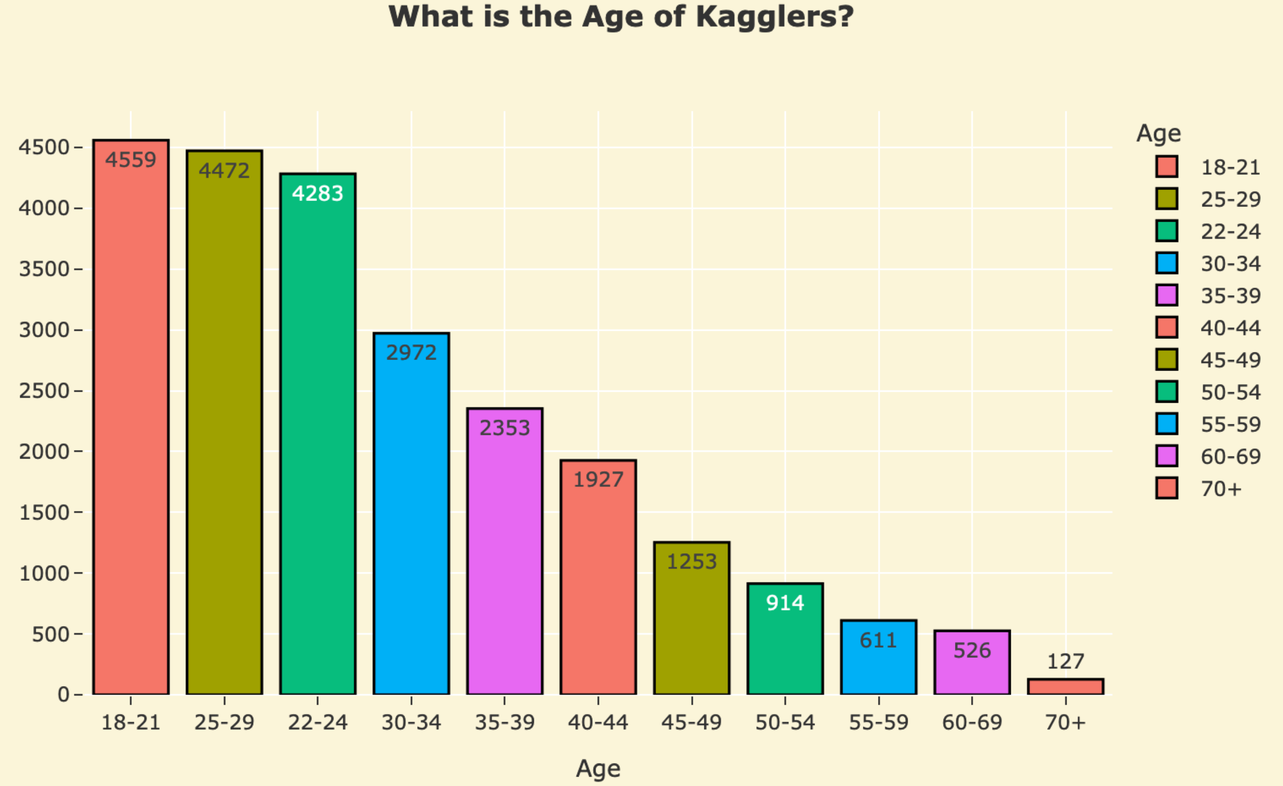
参赛选手统计

Part3 参赛选手统计

Coggle数据科学

- ✓ Kaggle平台上男性选手占比为77%，和「Coggle」公众号性别占比相同
- ✓ 大部分的Kaggle用户在40岁，且20 – 30之间年轻人比较多





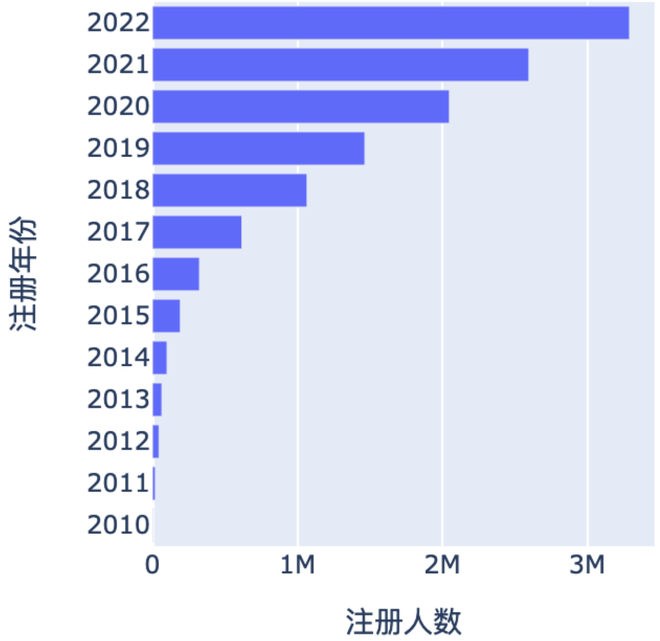
Part3 参赛选手统计

Coggle数据科学

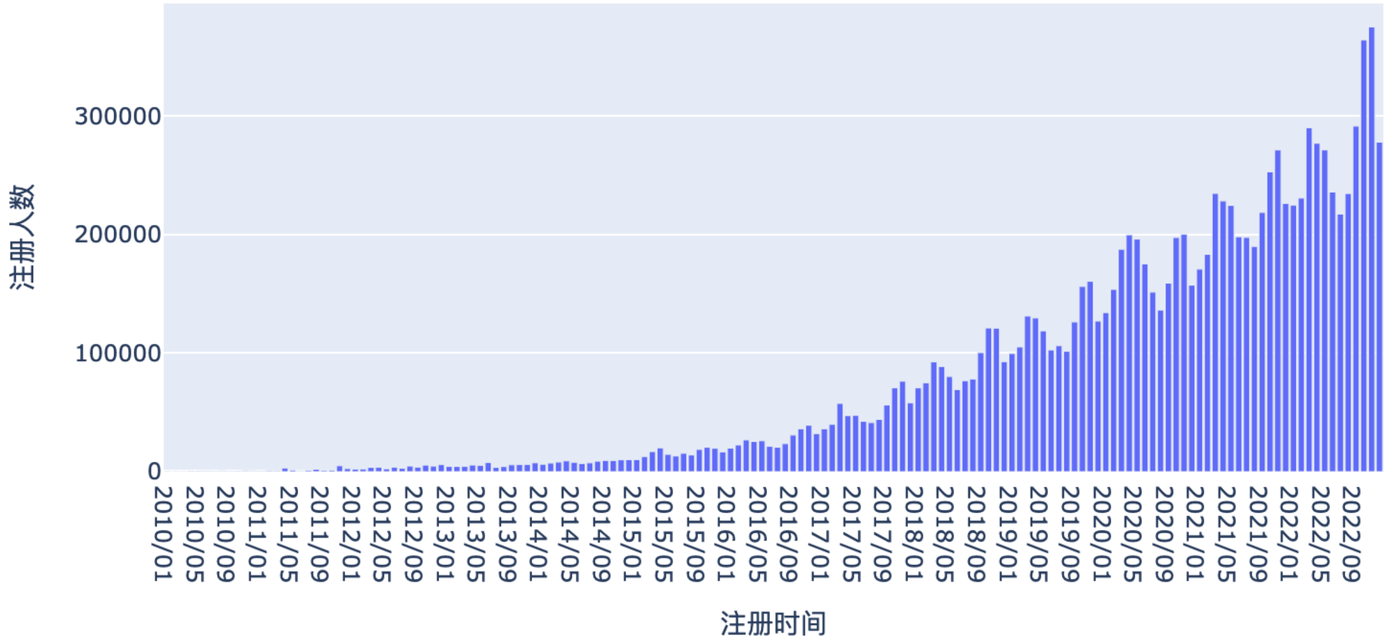
- ✓ Kaggle累计注册用户已经超过1100万
- ✓ Kaggle逐渐成为数据集 & 模型 & 代码运行平台

- Kaggle Dataset可以免费存储100GB私有文件
- Kaggle Notebook可以支持CPU / GPU / TPU运行
  - 不限制CPU Notebook运行时间
  - 每周30 – 40小时免费GPU / TPU运行时间

每年新注册人数



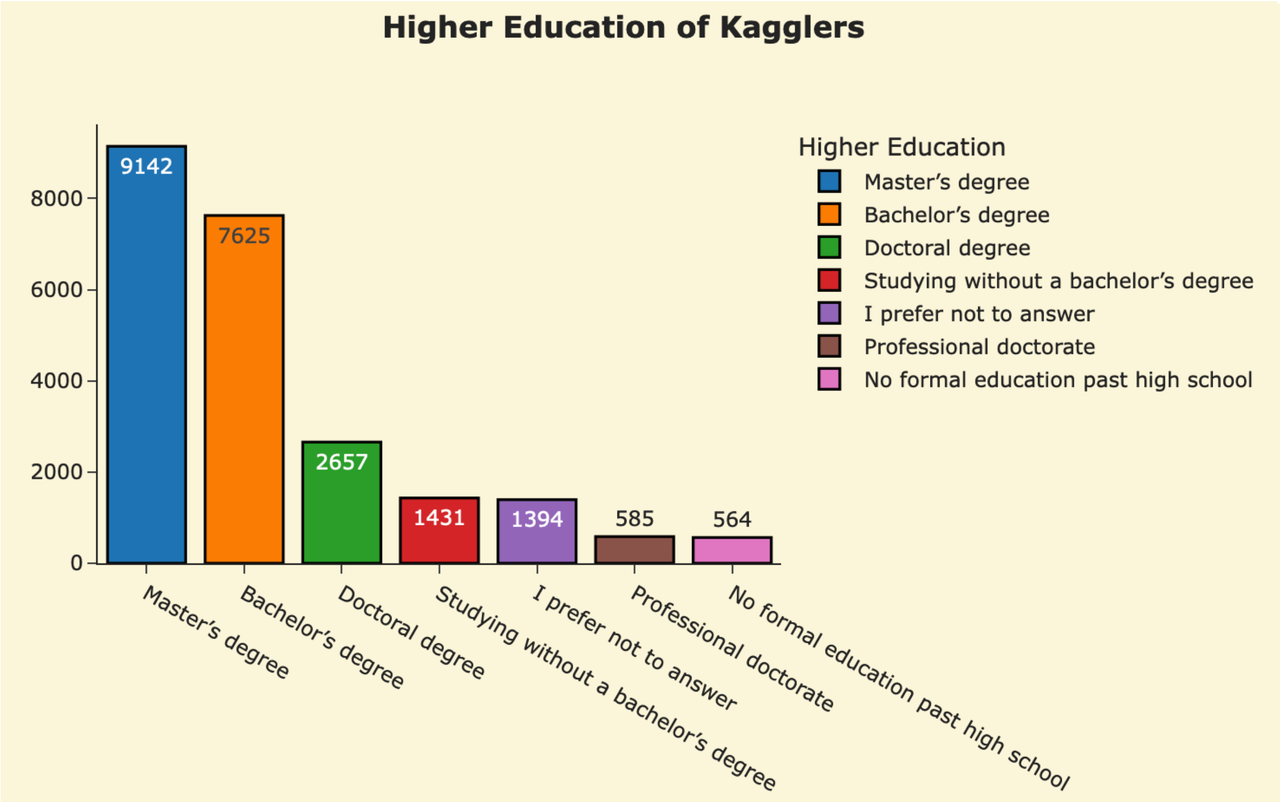
每月新注册人数

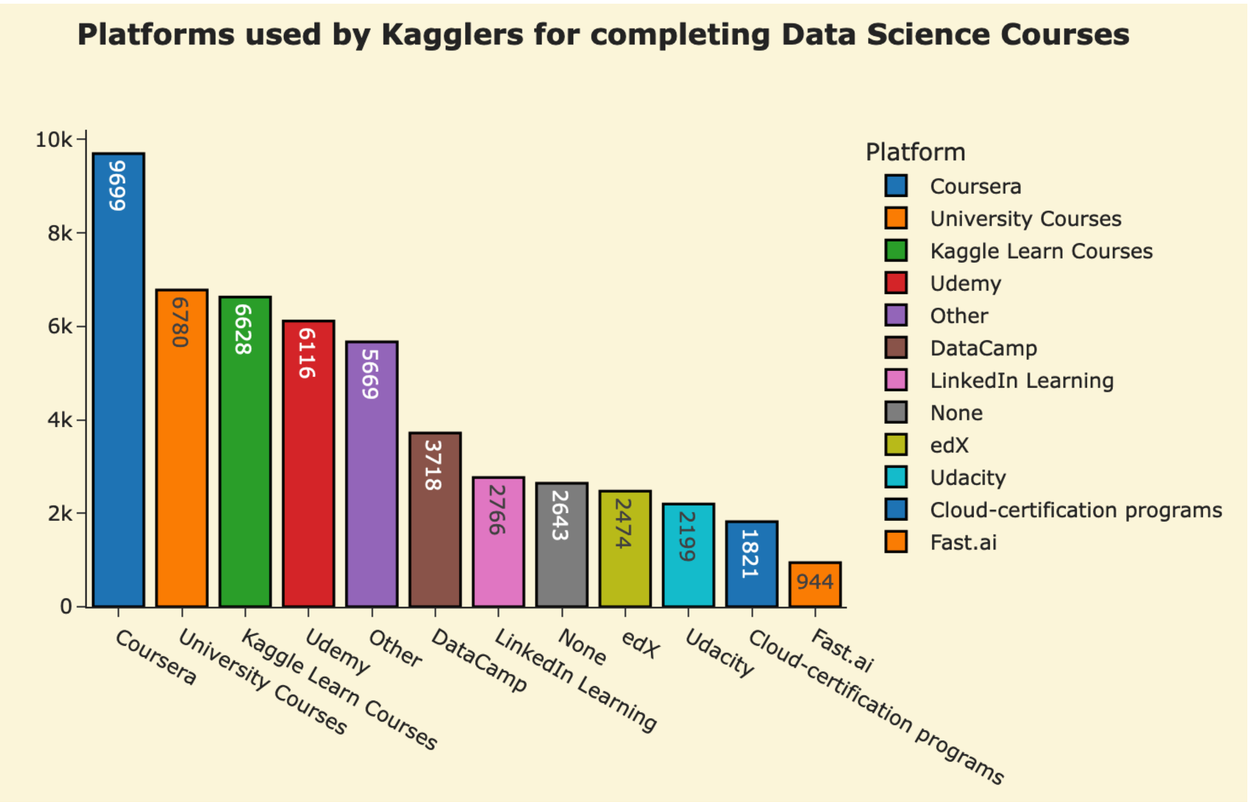


Part3 参赛选手统计

Coggle数据科学

- ✓ 40%左右的Kaggle用户拥有硕士学历，高学历占比较多
- ✓ Coursera是Kaggle用户最偏爱的在线学习平台





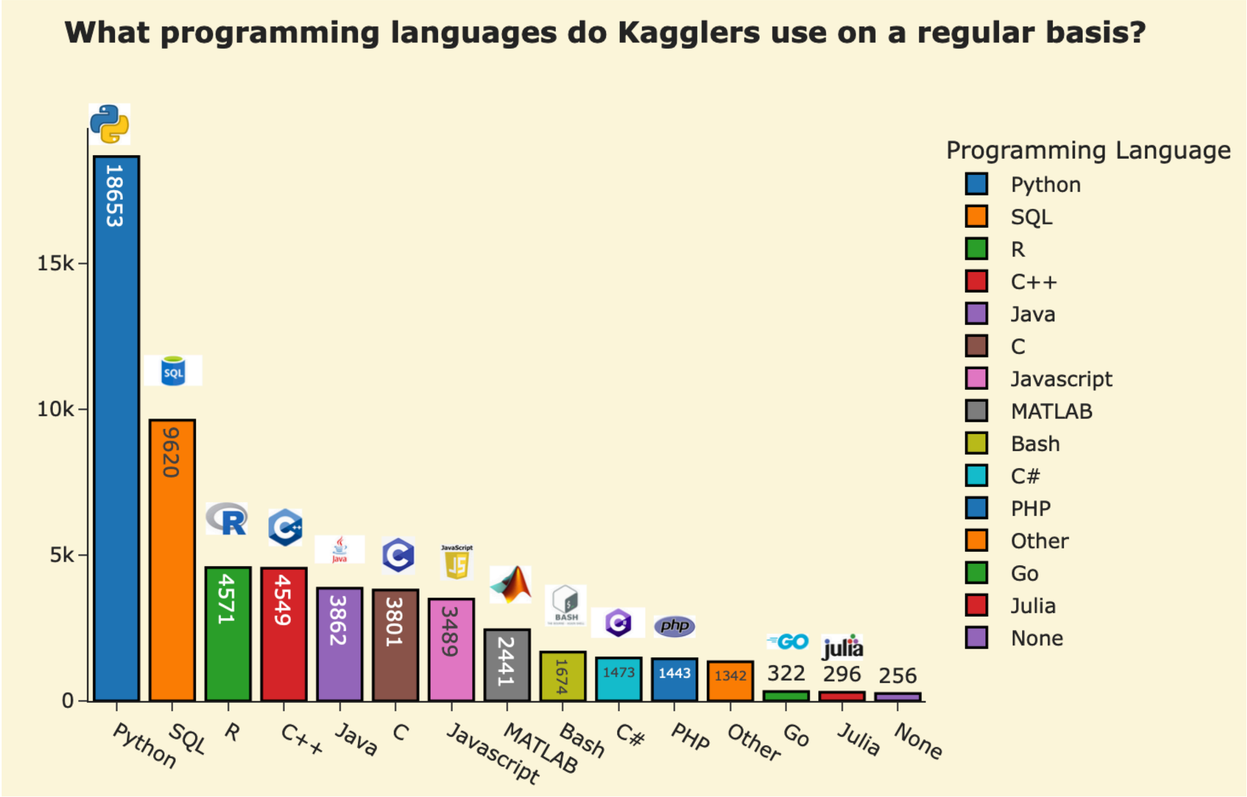
PART 04

参赛工具统计

Part4 参赛工具统计

Coggle数据科学

- ✓ Python是第一编程语言，其次是SQL和R
- ✓ 具统计Kaggle上R语言的Notebook在逐渐变少



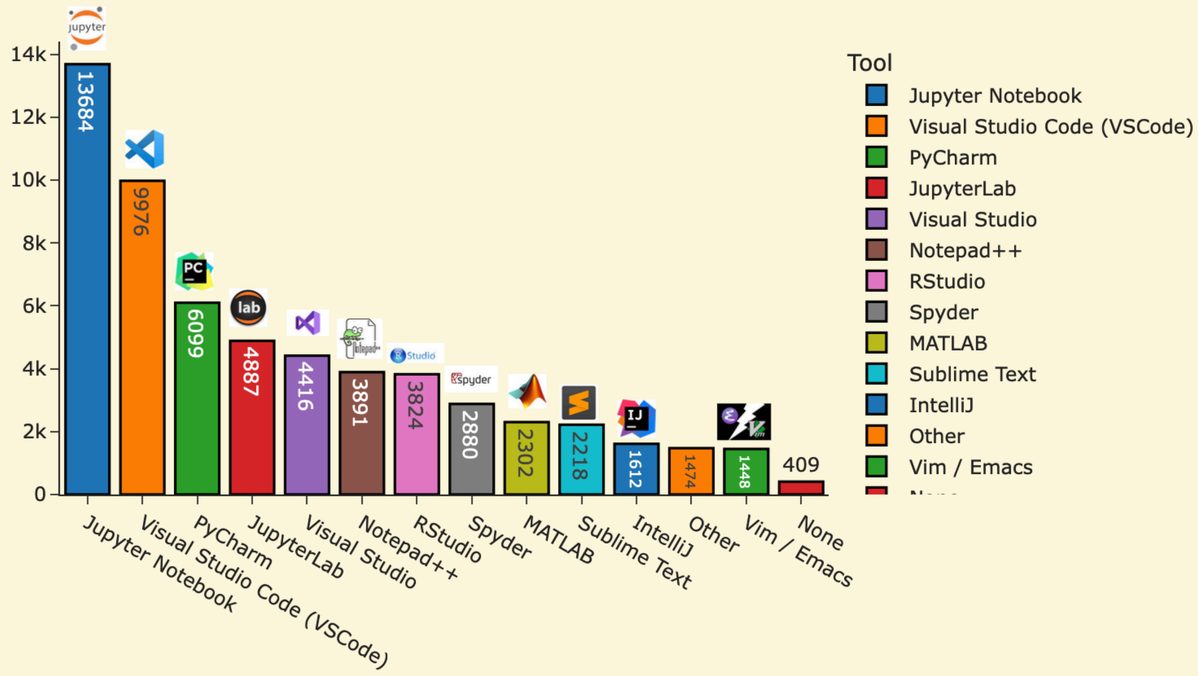
Part4 参赛工具统计

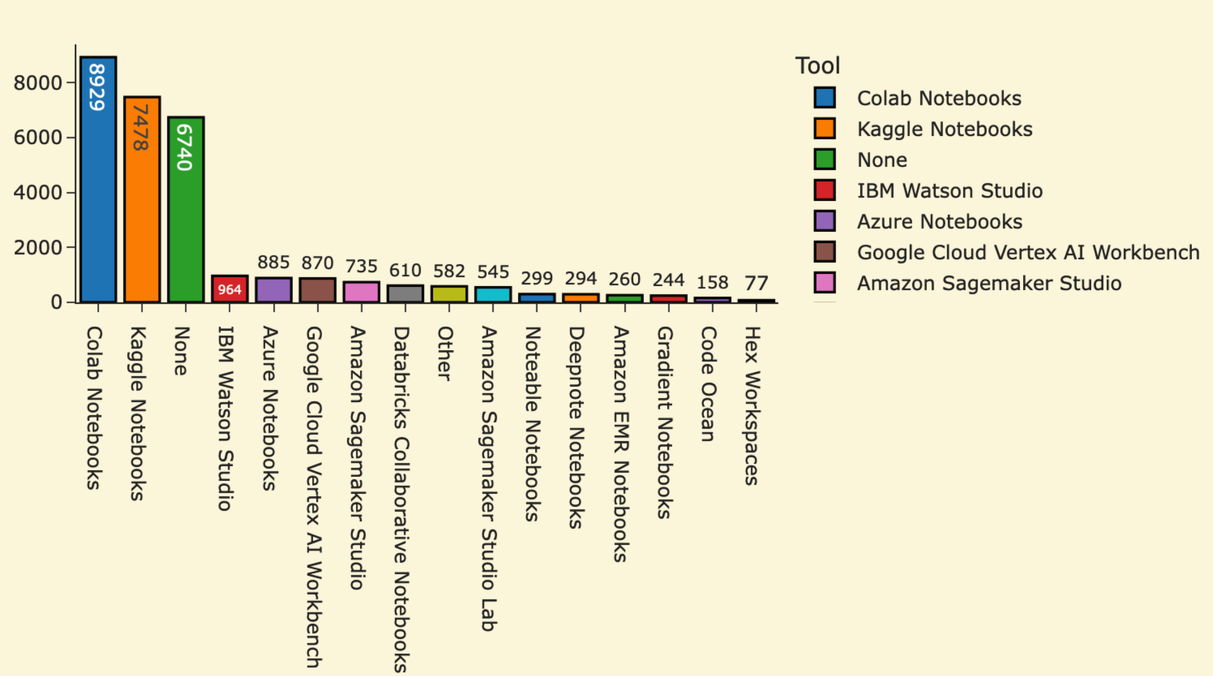
Coggle数据科学

- ✓ Jupyter Notebook是最受欢迎的IDE，其次是VSCode
- ✓ Colab和Kaggle是最受欢迎的Notebook平台

Colab是谷歌的免费在线Notebook平台，支持CPU / GPU和TPU运行，但需要开代理使用。

- Colab可以读取谷歌云盘文件
- Colab的GPU不限制每周运行时间

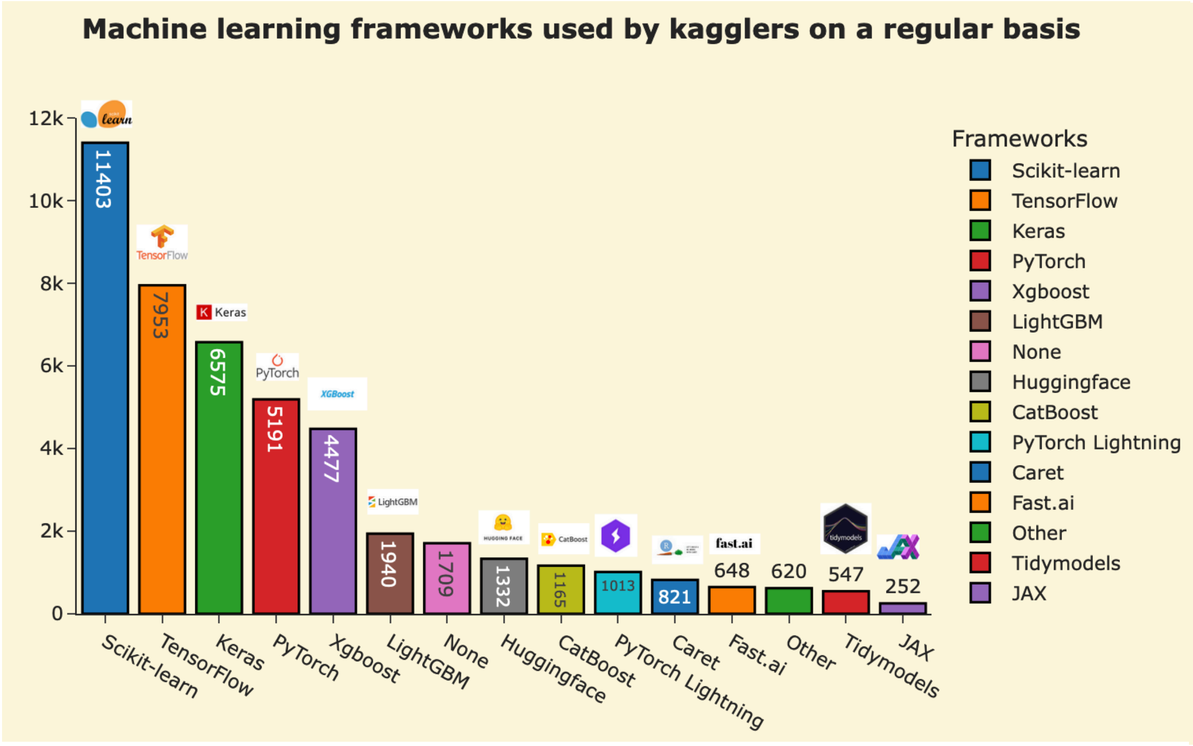


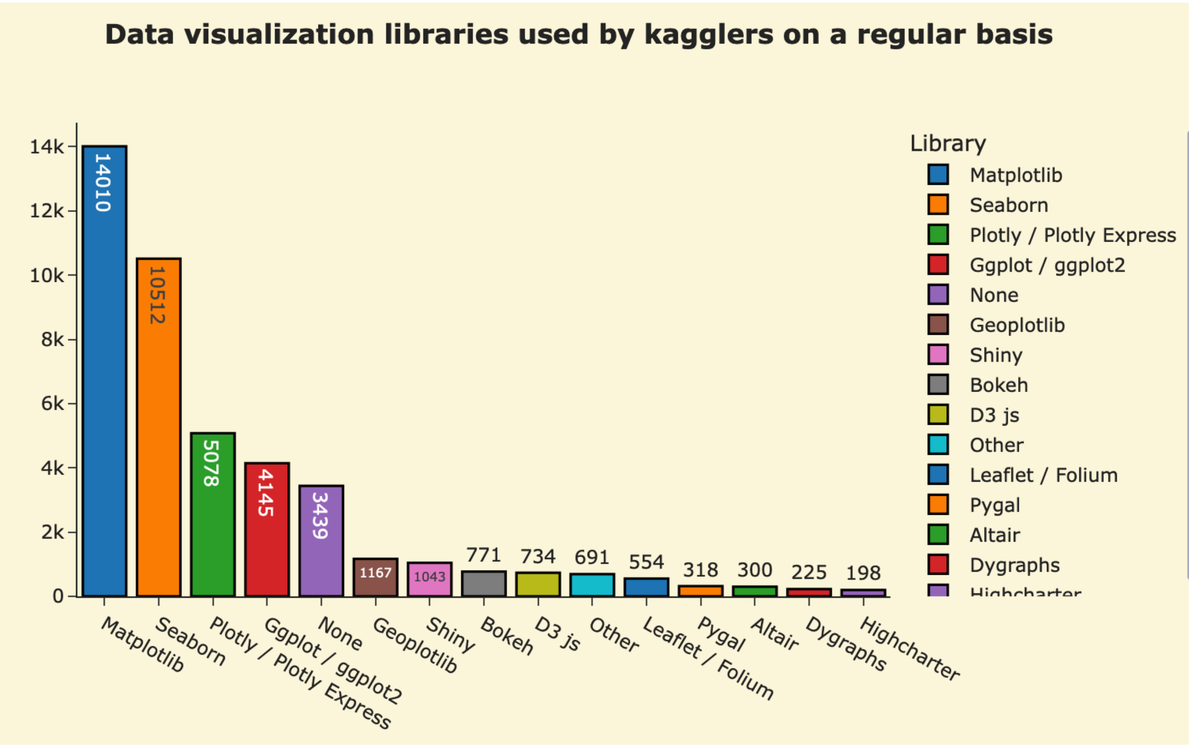


Part4 参赛工具统计

Coggle数据科学

- ✓ 机器学习库流行排序：scikit-learn、XGBoost、LightGBM、Catboost、Caret
- ✓ 深度学习库流行排序：TensorFlow、Keras、Pytorch、Jax
- ✓ 可视化库流行排序：Matplotlib、Seaborn、Plotly、ggplot2





PART 05

比赛内容汇总

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Jan 2022

Practice your ML skills on this approachable dataset!

赛题任务：有两家（虚构的）独立连锁店销售 Kaggle 商品，它们希望成为所有 Kaggle 产品的官方渠道。我们决定看看 Kaggle 社区是否可以帮助我们确定哪些连锁店未来的销售额会最好。

是否Kernel赛题：否

赛题数据大小：1.7 MB

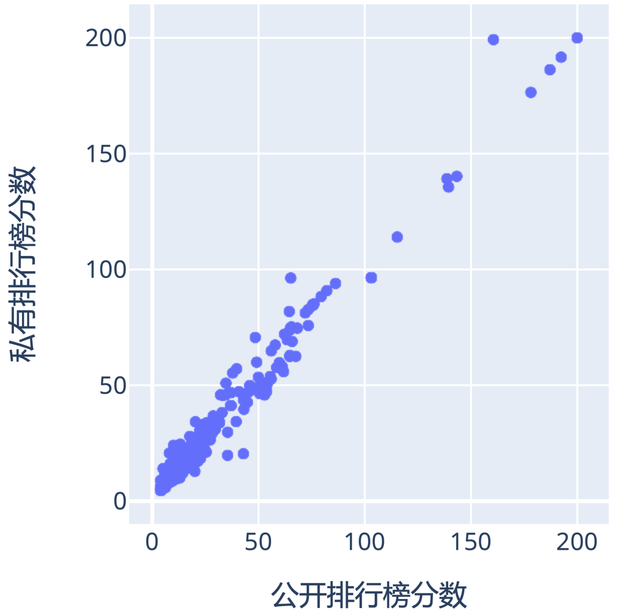
赛题类型：Playground、数据挖掘、时序回归

评价指标：SMAPE

报名人数/提交次数：1646 / 16151

赛题难度：⭐⭐

排行榜 SMAPE 得分（越低越好）



- ✓ 第1名：方案
- ✓ 第5名：方案
- ✓ 第16名：方案

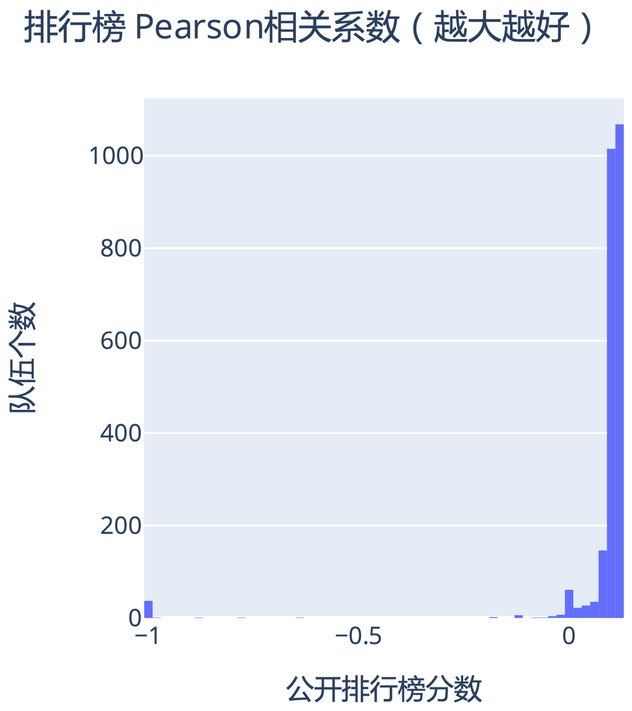
Part5 比赛内容汇总

Coggle数据科学

赛题名称：Ubiquant Market Prediction

Make predictions against future market data

赛题任务：Ubiquant是国内领先的量化对冲基金，总部位于中国。他们成立于2012年，依托国际数学和计算机科学人才以及尖端技术来推动量化金融市场投资。在本次比赛中，您将构建一个预测投资回报率的模型。根据历史价格训练和测试您的算法。热门条目将尽可能准确地解决这个现实世界的数据科学问题。



是否Kernel赛题：是

赛题数据大小：

赛题类型：Featured、金融量化

评价指标：Pearson 相关系数

报名人数/提交次数：2893 / 4159

赛题难度：⭐⭐⭐⭐

- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Happywhale - Whale and Dolphin Identification

Identify whales and dolphins by unique characteristics

赛题任务：Happywhale 是一个研究合作和公民科学网络平台。它的使命是通过高质量的保护科学和教育提高全球对海洋环境的理解和关心。在本次比赛中，您将开发一个模型，根据鲸鱼和海豚的自然标记的独特但通常很微妙的特征来匹配个体鲸鱼和海豚。

是否Kernel赛题：否

赛题数据大小：62GB

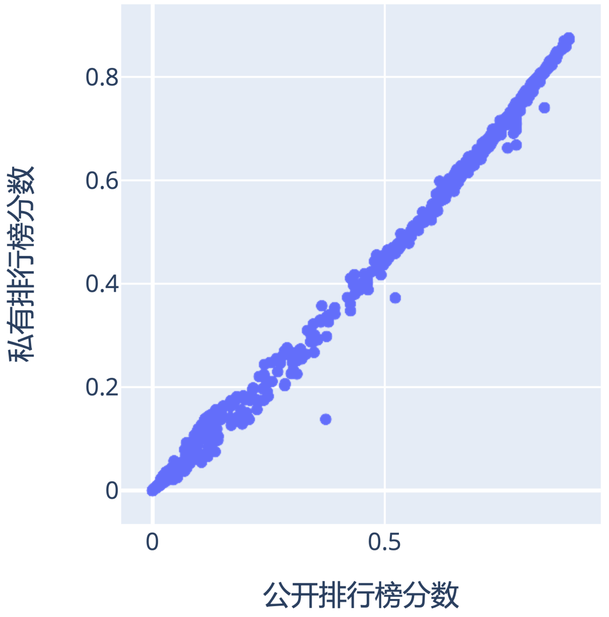
赛题类型：Research、计算机视觉、细粒度分类

评价指标：Mean Average Precision @ 5

报名人数/提交次数：1588 / 39284

赛题难度：⭐⭐⭐⭐

排行榜 MAP@5 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Tabular Playground Series - Feb 2022]

Practice your ML skills on this approachable dataset!

赛题任务：比赛任务是使用来有数据压缩和数据丢失的基因组分析技术的数据对10种不同的细菌种类进行分类。

是否Kernel赛题：否

赛题数据大小：1.8GB

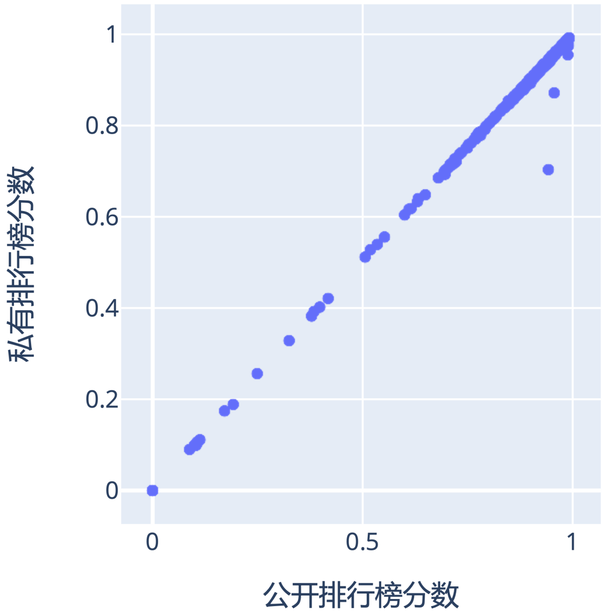
赛题类型：Playground、数据挖掘、多分类

评价指标：准确率

报名人数/提交次数：1255 / 11766

赛题难度：⭐⭐

排行榜 准确率 得分（越大越好）



✓ 第1名：方案，代码

✓ 第2名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：NBME - Score Clinical Patient Notes

Identify Key Phrases in Patient Notes from Medical Licensing Exams

赛题任务：在本次比赛中，您将确定患者笔记中的特定临床概念。具体来说，您将开发一种自动化方法，将临床概念从考试规则（例如，“食欲减退”）映射到医学生写的临床患者笔记中表达这些概念的各种方式。

是否Kernel赛题：是

赛题数据大小：35MB

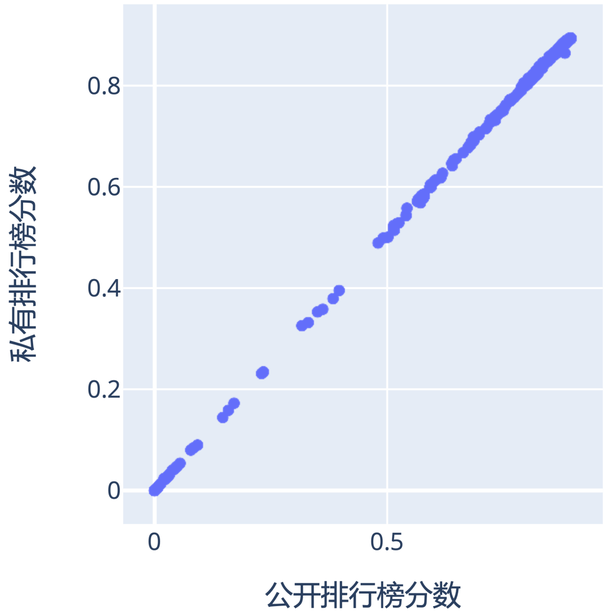
赛题类型：Featured、自然语言处理、信息抽取

评价指标：F1值

报名人数/提交次数：1471 / 28049

赛题难度：⭐⭐⭐

排行榜 F1值 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：H&M Personalized Fashion Recommendations

Provide product recommendations based on previous purchases

赛题任务：H&M 在线商店为购物者提供了广泛的产品选择供他们浏览。H&M 集团邀请您根据以往交易数据以及客户和产品元数据开发产品推荐。可用的元数据涵盖从简单数据（例如服装类型和客户年龄）到产品描述中的文本数据，再到服装图像中的图像数据。

是否Kernel赛题：否

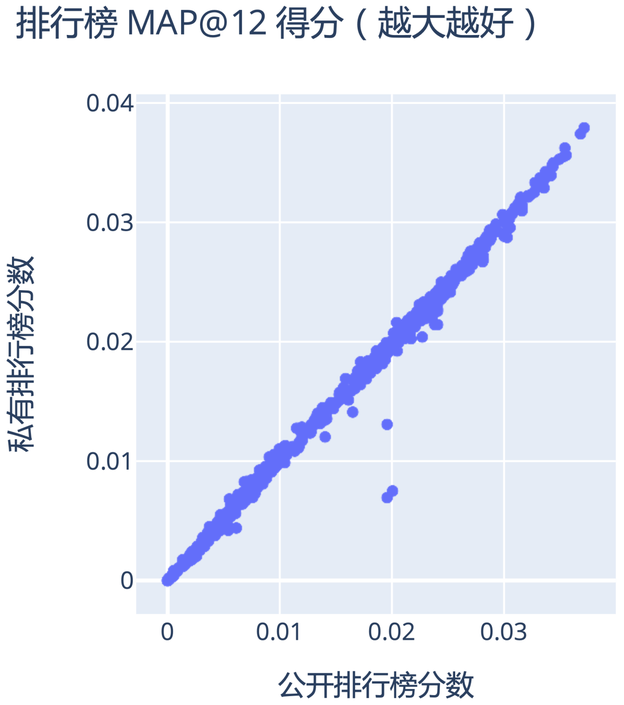
赛题数据大小：34GB

赛题类型：Featured、推荐系统、多模态

评价指标：Mean Average Precision @ 12

报名人数/提交次数：2952 / 38854

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Herbarium 2022 - FGVC9

Identify plant species of the Americas from herbarium specimens

赛题任务：在植物学中，“植物群”是对在某个地理区域发现的植物的完整描述。今年的竞赛数据集旨在识别北美的植物群，北美植物群数据集包含 15,501 种植物的 105 万张图像，这些植物占北美记录的分类单元的 90% 以上。

是否Kernel赛题：否

赛题数据大小：163GB

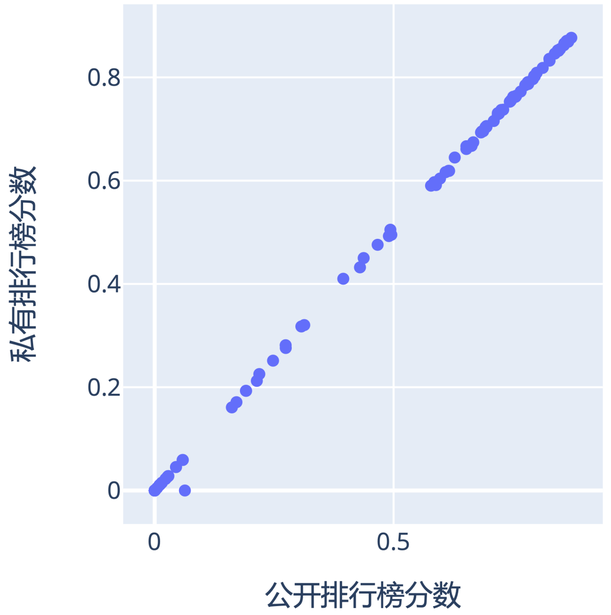
赛题类型：Research、计算机视觉、细粒度分类

评价指标：F1值

报名人数/提交次数：184 / 1534

赛题难度：⭐⭐⭐

排行榜 F1 得分（越大越好）



✓ 第1名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：BirdCLEF 2022

Identify bird calls in soundscapes

赛题任务：当前处理大型生物声学数据集的方法涉及对每个记录进行手动注释，这需要大量的时间。在本次比赛中，您将使用机器学习技能通过声音识别鸟类种类。您将开发一个模型，该模型可以处理连续的音频数据，然后通过声学识别物种。

是否Kernel赛题：是

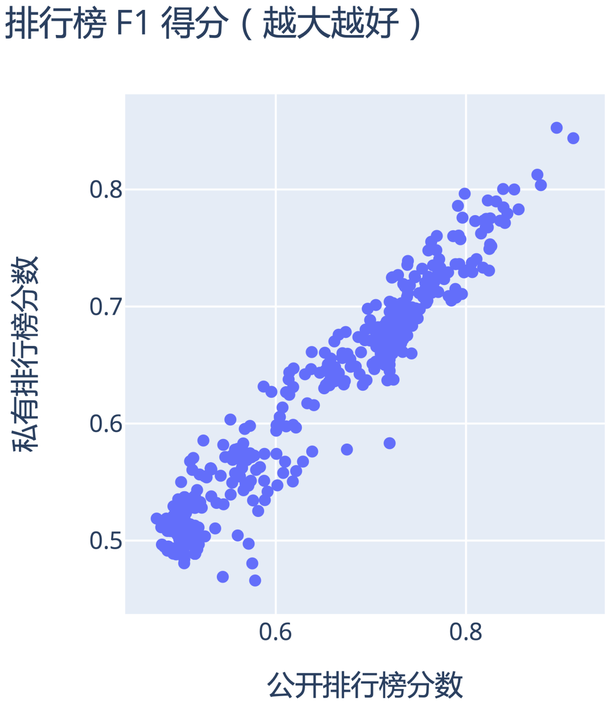
赛题数据大小：6.6GB

赛题类型：Research、语音识别

评价指标：F1值

报名人数/提交次数：1019 / 23352

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案，代码
- ✓ 第2名：方案，代码
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：March Machine Learning Mania 2022 - Women's

Predict the 2022 College Women's Basketball Tournament

赛题任务：在第八届年度NACC中，今年比赛试图预测今年美国女子大学篮球锦标赛的结果。赛题提供了 NCAA 历史比赛的数据，并鼓励您使用其他来源的公开可用数据。在这个两阶段比赛的第一阶段，参赛者将针对之前的比赛构建和测试他们的模型。在第二阶段，参与者将预测 2022 年锦标赛的结果。

是否Kernel赛题：否

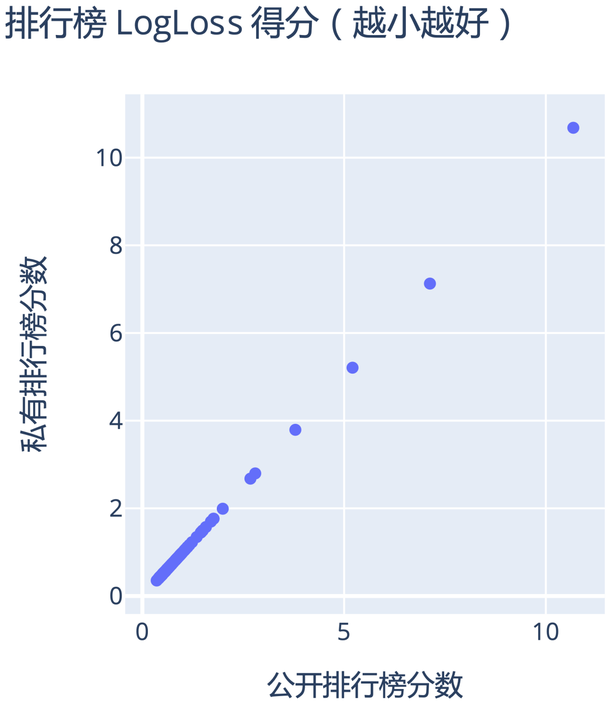
赛题数据大小：25MB

赛题类型：Featured、数据挖掘

评价指标：LogLoss

报名人数/提交次数：711 / 1203

赛题难度：⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：March Machine Learning Mania 2022 - Men's

Predict the 2022 College Men's Basketball Tournament

赛题任务：在第八届年度NACC中，今年比赛试图预测今年美国男子大学篮球锦标赛的结果。赛题提供了 NCAA 历史比赛的数据，并鼓励您使用其他来源的公开可用数据。在这个两阶段比赛的第一阶段，参赛者将针对之前的比赛构建和测试他们的模型。在第二阶段，参与者将预测 2022 年锦标赛的结果。

是否Kernel赛题：否

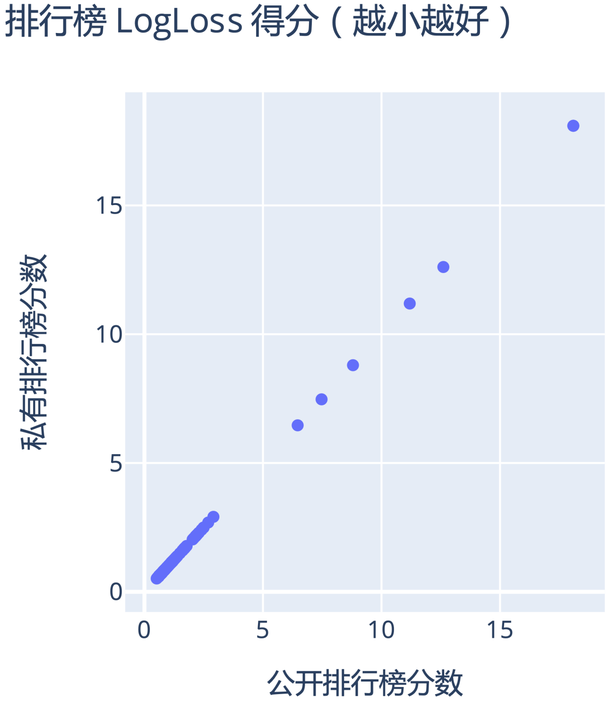
赛题数据大小：238

赛题类型：Featured、数据挖掘

评价指标：LogLoss

报名人数/提交次数：1025 / 1681

赛题难度：⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第5名：方案

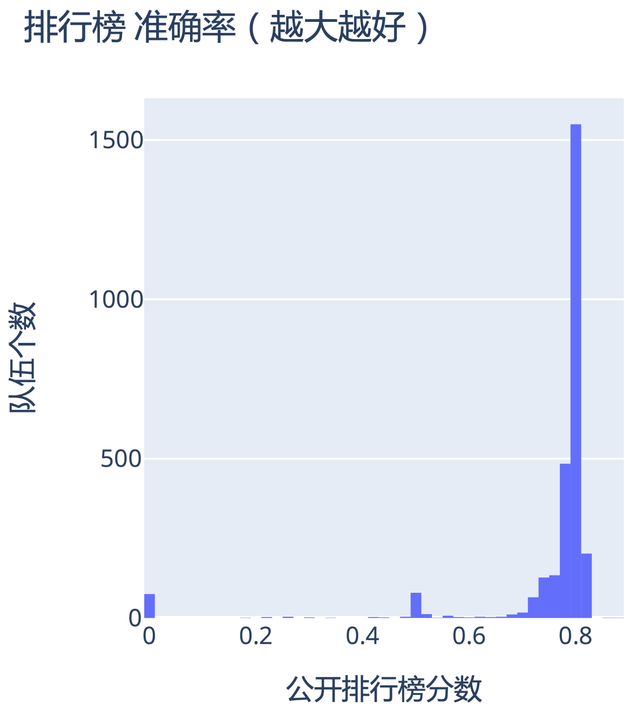
Part5 比赛内容汇总

Coggle数据科学

赛题名称：Spaceship Titanic

Predict which passengers are transported to an alternate dimension

赛题任务：泰坦尼克号宇宙飞船是一个月前发射的星际客轮。这艘船载有近 13,000 名乘客，开始了它的处女航，将移民从我们的太阳系运送到围绕附近恒星运行的三个新宜居系外行星。为了帮助救援人员和找回失踪的乘客，您面临的挑战是使用从飞船损坏的计算机系统中恢复的记录来预测哪些乘客被异常运送。



是否Kernel赛题：否

赛题数据大小：1MB

赛题类型：Playground、数据挖掘、二分类

评价指标：准确率

报名人数/提交次数：3048 / 24068

赛题难度：⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Mar 2022

Practice your ML skills on this approachable dataset!

赛题任务：赛题挑战是预测美国大都市十二小时的交通流量。数据集中的时间序列标有位置坐标和方向。

是否Kernel赛题：否

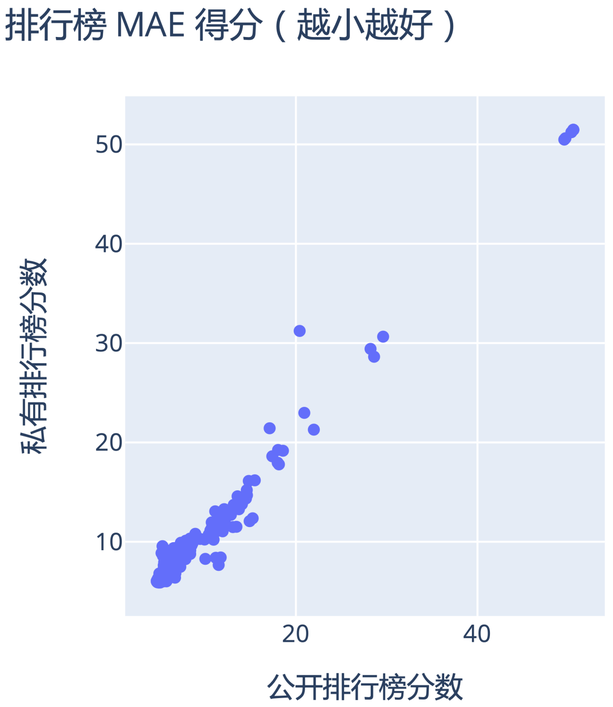
赛题数据大小：31MB

赛题类型：Playground、数据挖掘、时序回归

评价指标：MAE

报名人数/提交次数：1003 / 9971

赛题难度：⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：GeoLifeCLEF 2022 - LifeCLEF 2022 x FGVC9

Location-based species presence prediction

赛题任务：本次比赛的目的是预测植物和动物物种的定位。赛题提供了来自法国和美国的 17000 种物种的 160 万地理定位观测值。赛题目标是对于测试集中的每个 GPS 位置，返回一组应包含真实观察到的物种的候选物种。

是否Kernel赛题：否

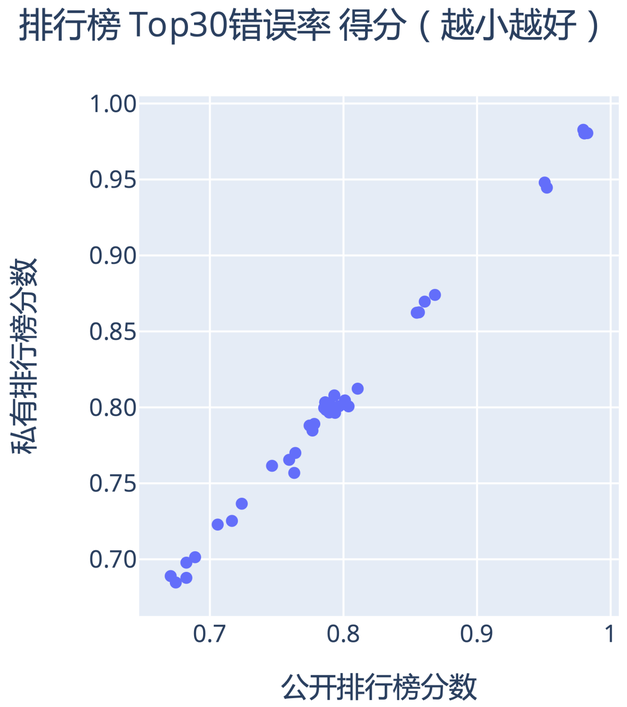
赛题数据大小：61GB

赛题类型：Research、计算机视觉

评价指标：Top30 错误率

报名人数/提交次数：59 / 242

赛题难度：⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案

Part5 比赛内容汇总

赛题名称：Excellence in Research Award (Phase II)

WiDS Datathon Further Examines the Impacts of Climate Change

赛题任务：今年的 WiDS Datathon 第一阶段侧重于通过预测使用情况提高建筑能效，减轻气候变化影响的重要方法。在第二阶段中，我们将研究气候变化对多个领域的影响。

是否Kernel赛题：否

赛题数据大小：66MB

赛题类型：Analytics

评价指标：分析结果评分

报名人数/提交次数：

赛题难度：⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Sorghum -100 Cultivar Identification - FGVC 9

Identify crop varietals

赛题任务：在植物育种的背景下，栽培品种或亲本系通常是已知的。但传感器数据可用于确定种植中可能发生错误。例如，种子可能在播种前被贴错标签，或者播种机可能被卡住，导致种子在田间不均匀地播撒。赛题数据可用于开发和评估各种植物类型，在本次比赛中，我们关注图片中显示的植物是什么品种?

是否Kernel赛题：否

赛题数据大小：71GB

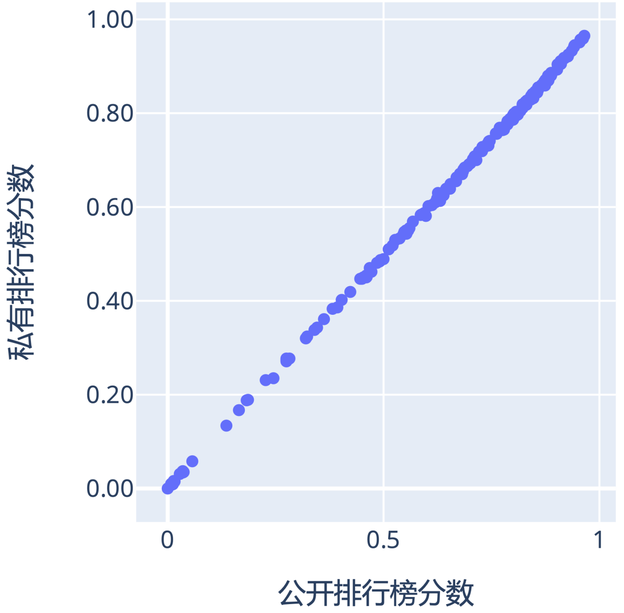
赛题类型：Research、计算机视觉、多分类

评价指标：准确率

报名人数/提交次数：294 / 3904

赛题难度：⭐⭐⭐

排行榜 准确率 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Hotel-ID to Combat Human Trafficking 2022 - FGVC9](#)

Recognizing hotels to aid Human trafficking investigations

赛题任务：人口贩运的受害者经常在旅馆房间里被拍到。识别图片中的酒店对这些人口贩运调查至关重要，即使图像中没有受害者，酒店识别通常也是一项具有挑战性的细粒度视觉识别任务。在本次比赛中，参赛者的任务是识别来自 TraffickCam 数据集中的酒店，这些图像基于具有已知酒店 ID 的大型训练图像库。

是否Kernel赛题：是

赛题数据大小：15GB

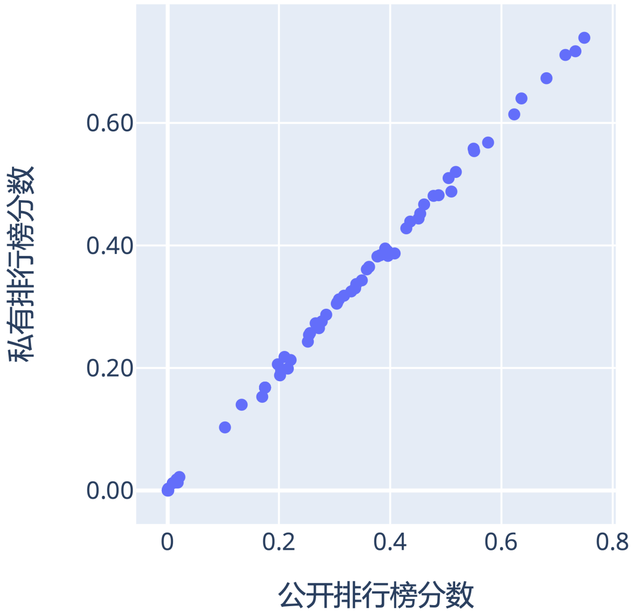
赛题类型：Research、计算机视觉、细粒度分类

评价指标：Mean Average Precision @ 5

报名人数/提交次数：135 / 1712

赛题难度：⭐⭐⭐

排行榜 MAP@5 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案

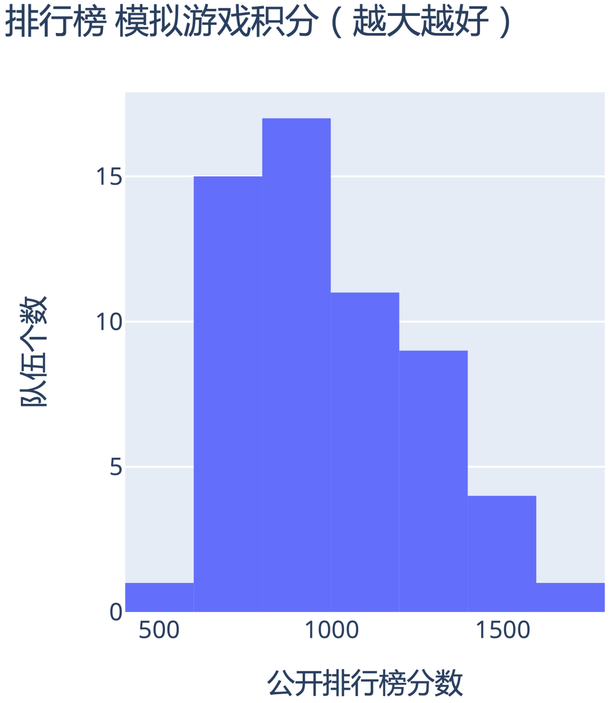
Part5 比赛内容汇总

Coggle数据科学

赛题名称：Kore 2022 - Beta

Collect the maximum amount of Kore against your opponents

赛题任务：在这款回合制模拟游戏中，您将控制一支小型宇宙飞船舰队。当你从太空深处开采稀有矿物“kore”时，你将它传送回你的家乡。在每场比赛中，四名玩家将竞争从棋盘上收集最多的矿物。谁在 400 回合结束时拥有最多的矿物，或者在此之前从棋盘上消灭所有对手，最终将成为赢家。



是否Kernel赛题：是

赛题数据大小：

赛题类型：强化学习

评价指标：模拟游戏积分

报名人数/提交次数：61 / 734

赛题难度：⭐⭐⭐⭐⭐

✓ 第1名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：iWildCam 2022 - FGVC9

Count the number of animals in a sequence of images

赛题任务：Camera Traps 可自动收集大量图像数据。全世界的生态学家都使用相机来监测动物物种的生物多样性和种群密度。为了估计相机数据中物种个数和种群密度，生态学家不仅需要知道看到了哪些物种，还需要知道看到了每个物种的数量。今年的 iWildCam 比赛将完全专注于动物计数。

是否Kernel赛题：否

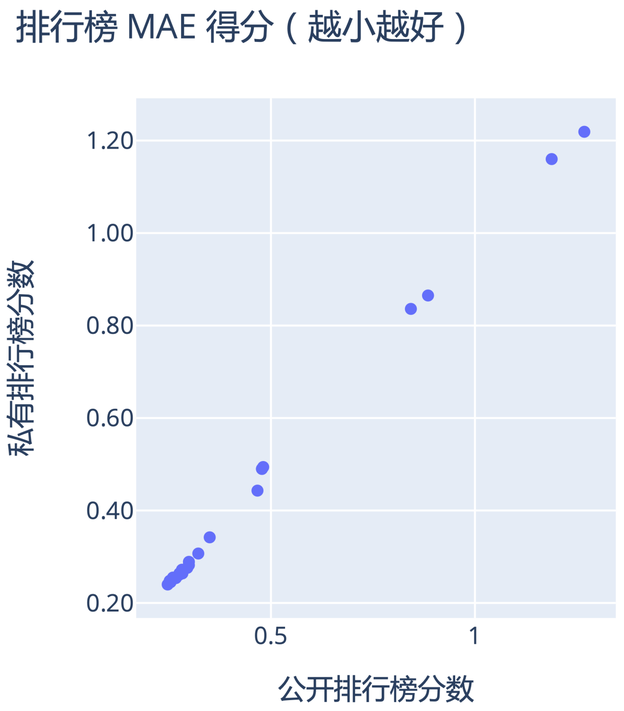
赛题数据大小：111GB

赛题类型：Research、计算机视觉、物体计数

评价指标：MAE

报名人数/提交次数：29 / 300

赛题难度：⭐⭐⭐⭐



✓ 第1名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：U.S. Patent Phrase to Phrase Matching

Help Identify Similar Phrases in U.S. Patents

赛题任务：专利在授予前要经过严格的审查程序，并且由于美国的创新历史跨越两个多世纪和 1100 万项专利。在本次比赛中，您将在一个语义相似性数据集上训练您的模型，以通过匹配专利文件中的关键短语来提取相关信息。在专利检索和审查过程中确定短语之间的语义相似性对于确定一项发明是否已经存在。

是否Kernel赛题：是

赛题数据大小：2MB

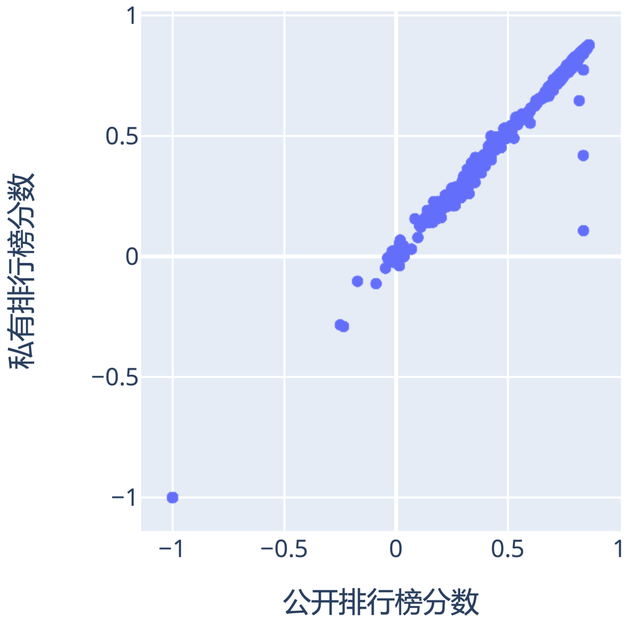
赛题类型：Featured、自然语言处理、信息抽取

评价指标：Pearson 相关系数

报名人数/提交次数：2325 / 42902

赛题难度：⭐⭐⭐⭐

排行榜 Pearson 相关系数 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案，代码
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Apr 2022

Practice your ML skills on this approachable dataset!

赛题任务：赛题提供了数千个 60 秒的生物传感器数据序列，这些数据来自数百名参与者，他们可能处于两种可能的活动状态中的任何一种。你能从传感器数据中确定参与者处于什么状态吗?

是否Kernel赛题：否

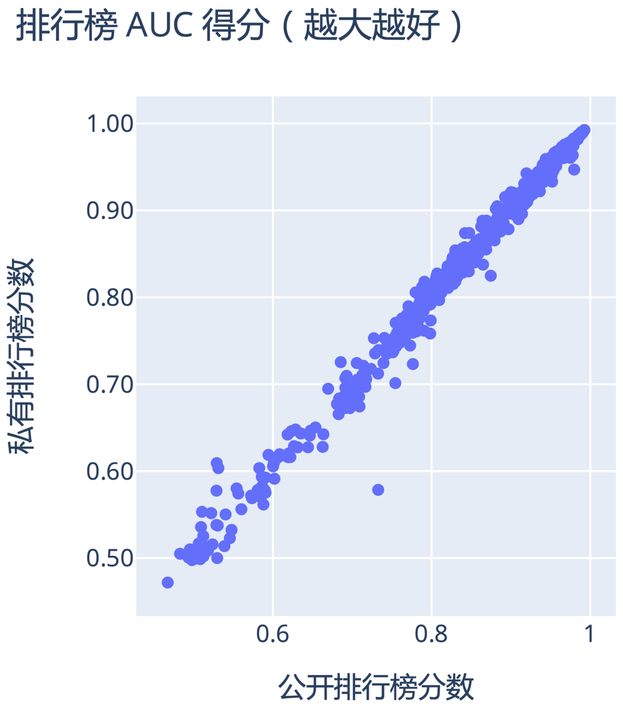
赛题数据大小：591MB

赛题类型：Playground、数据挖掘、二分类

评价指标：AUC

报名人数/提交次数：860 / 7235

赛题难度：⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案，代码
- ✓ 第3名：方案，代码
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Image Matching Challenge 2022

Register two images from different viewpoints

赛题任务：我们最好的相机往往是手机，我们可能会拍下地标性建筑，或许我们也能够创建一个更完整的三维视图。从图像重建 3D 对象和建筑物的过程称为运动结构 (SfM)。在本次比赛中，您将创建一个机器学习算法，从不同的角度配准两张图像。

是否Kernel赛题：是

赛题数据大小：2.6GB

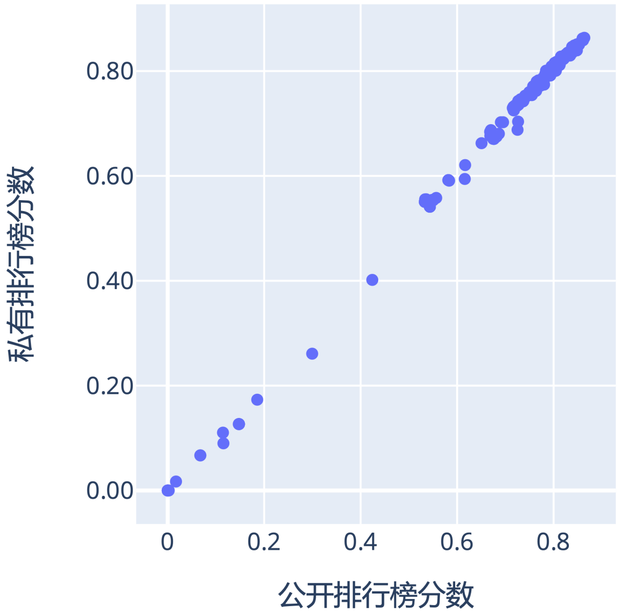
赛题类型：Research、计算机视觉、SfM

评价指标：平均准确率

报名人数/提交次数：829 / 14170

赛题难度：⭐⭐⭐⭐⭐

排行榜 平均准确率 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案，代码
- ✓ 第5名：方案

Part5 比赛内容汇总

赛题名称：JPX Tokyo Stock Exchange Prediction

Explore the Tokyo market with your data science skills

赛题任务：本次比赛将使用日本金融数据，让散户投资者能够最大程度地分析市场。比赛将涉及从符合预测条件的股票（约 2,000 只股票）中构建投资组合。具体来说，每个参与者将股票从最高到最低的预期回报进行排名，并根据前 200 只股票和后 200 只股票之间的回报差异进行评估。

是否Kernel赛题：是

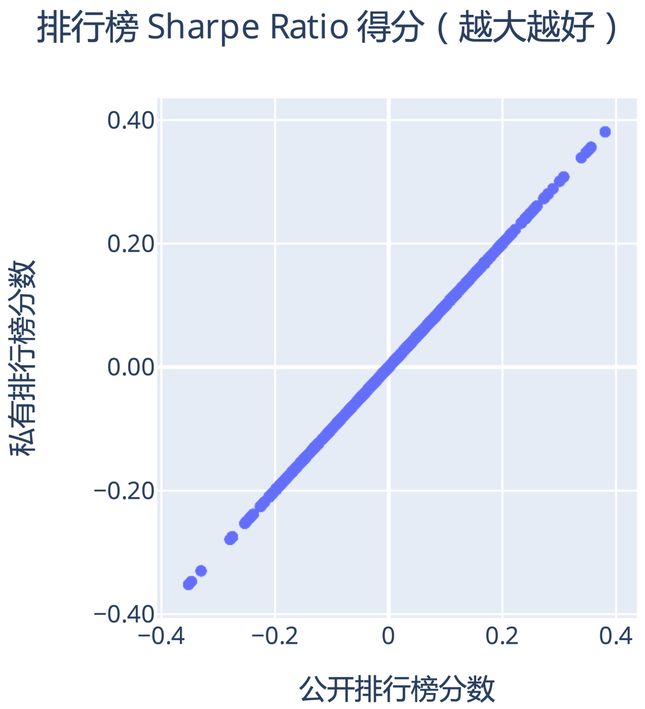
赛题数据大小：1.3GB

赛题类型：Featured，金融量化

评价指标：Sharpe Ratio

报名人数/提交次数：2033 / 1572

赛题难度：⭐⭐⭐⭐



- ✓ 第2名：方案
- ✓ 第4名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Kore 2022

Use a fleet of spaceships to mine minerals before your opponents

赛题任务：在这款回合制模拟游戏中，您将控制一支小型宇宙飞船舰队。你们并不是唯一有这个目标的文明。在每场比赛中，两名玩家将竞争从棋盘上收集最多的科雷。谁在 400 回合结束时拥有最大的 kore 缓存 - 或者在此之前从棋盘上消灭所有对手 - 将成为赢家！

是否Kernel赛题：是

赛题数据大小：

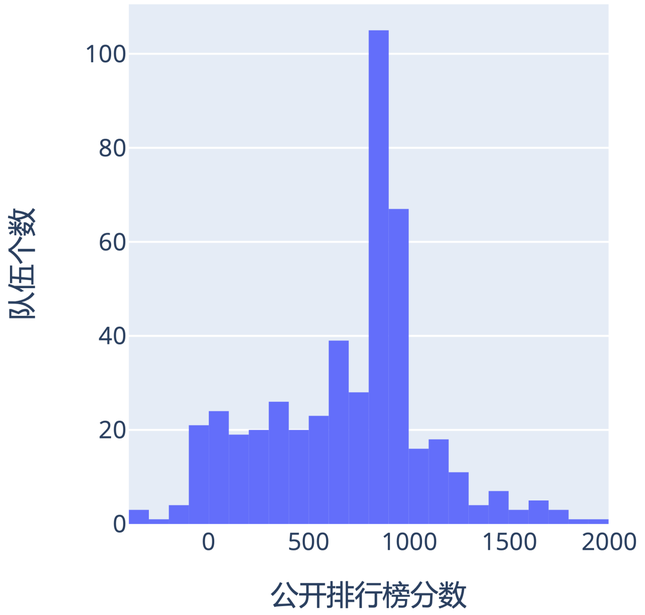
赛题类型：Featured，强化学习

评价指标：

报名人数/提交次数：537 / 12037

赛题难度：⭐⭐⭐⭐⭐

排行榜 模拟游戏积分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：UW-Madison GI Tract Image Segmentation

Track healthy organs in medical scans to improve cancer treatment

赛题任务：在本次比赛中您将创建一个深度学习模型，在 MRI 扫描中自动分割胃和肠。 MRI扫描来自实际的癌症患者，他们在放射治疗期间的不同日子进行了 1-5 次 MRI 扫描。您将基于这些扫描的数据集构建您的算法，以提出创造性的深度学习解决方案，帮助癌症患者获得更好的护理。

是否Kernel赛题：是

赛题数据大小：2.5GB

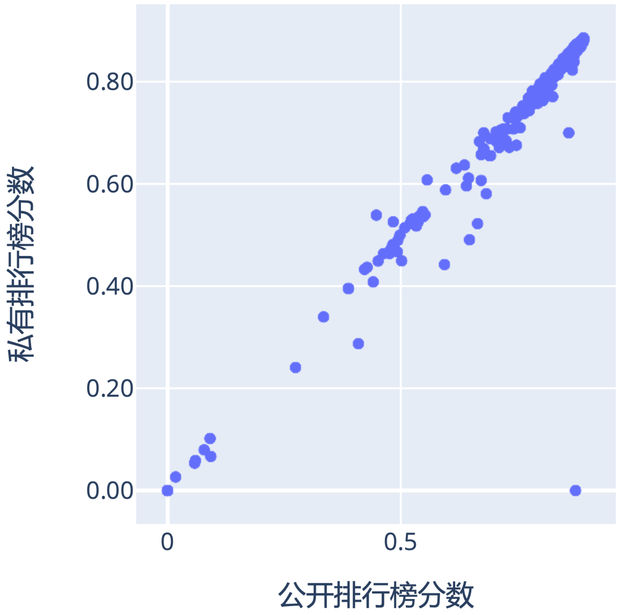
赛题类型：Research，计算机视觉、语义分割

评价指标：Dice系数 和 3D Hausdorff 距离

报名人数/提交次数：2078 / 40956

赛题难度：⭐⭐⭐⭐

排行榜 Dice系数 和 Hausdorff 得分越大越好)



- ✓ 第1名：方案，代码
- ✓ 第2名：方案
- ✓ 第3名：方案，代码
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Foursquare - Location Matching

Match point of interest data across datasets

赛题任务：当您寻找附近的餐馆或计划在未知区域出差时，商业兴趣点（POI）的大规模数据集可能包含丰富的真实世界信息。在本次比赛中，您将匹配兴趣点。使用包含超过 150 万个地点条目的数据集，这些条目经过大量更改以包括噪声、重复、无关或不正确的信息，您将生成一种算法来预测哪些地点条目代表相同的兴趣点。

是否Kernel赛题：是

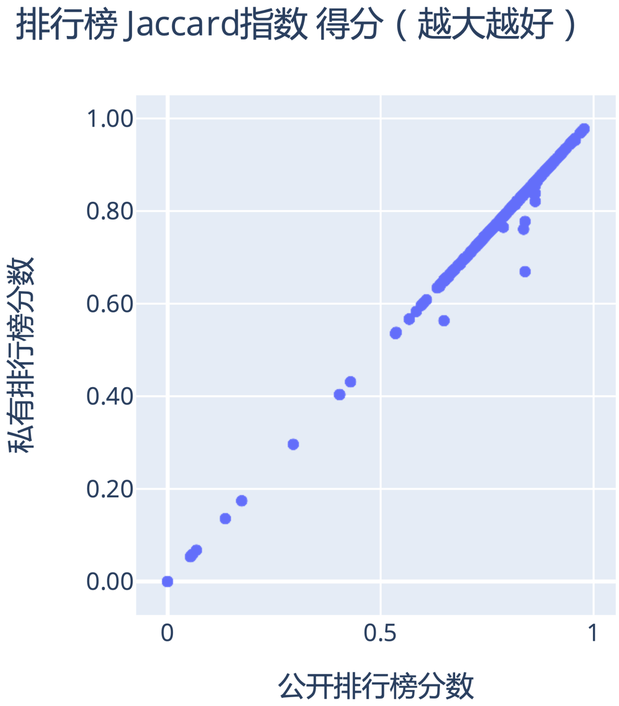
赛题数据大小：

赛题类型：Featured，数据挖掘

评价指标：Jaccard 指数

报名人数/提交次数：1290 / 22050

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案，代码
- ✓ 第2名：方案，代码
- ✓ 第3名：方案
- ✓ 第4名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Tabular Playground Series - May 2022]

Practice your ML skills on this approachable dataset!

赛题任务：本次赛题数据包含许多不同的交互特征。本次竞赛是探索识别和利用这些交互特征完成分类任务。

是否Kernel赛题：否

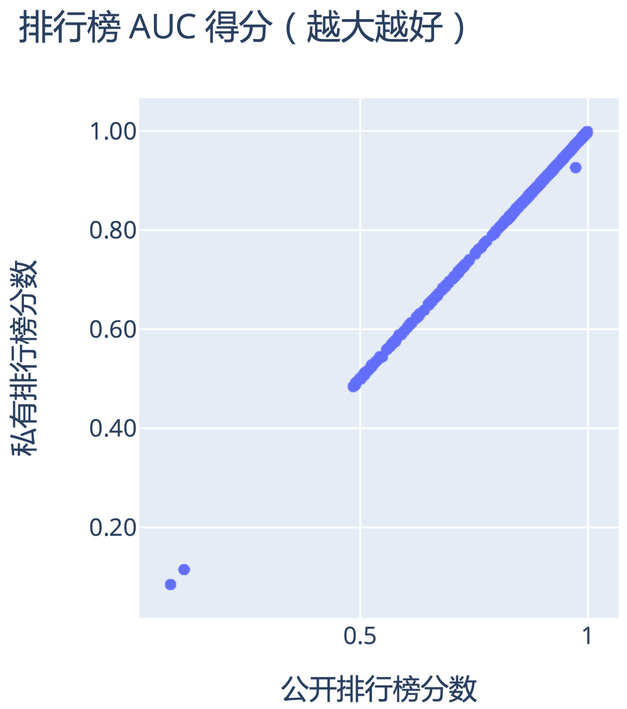
赛题数据大小：574MB

赛题类型：Playground、数据挖掘、二分类

评价指标：AUC

报名人数/提交次数：1176 / 8902

赛题难度：⭐⭐



- ✓ 第1名：方案，代码
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Google Smartphone Decimeter Challenge 2022

Improve GNSS positioning and navigation accuracy on smartphones

赛题任务：比赛的目标是将智能手机的位置计算到分米甚至厘米的分辨率，这可以实现需要车道级精度的服务，您在本次比赛中的努力可能会影响如何解释这些更困难的数据。

是否Kernel赛题：否

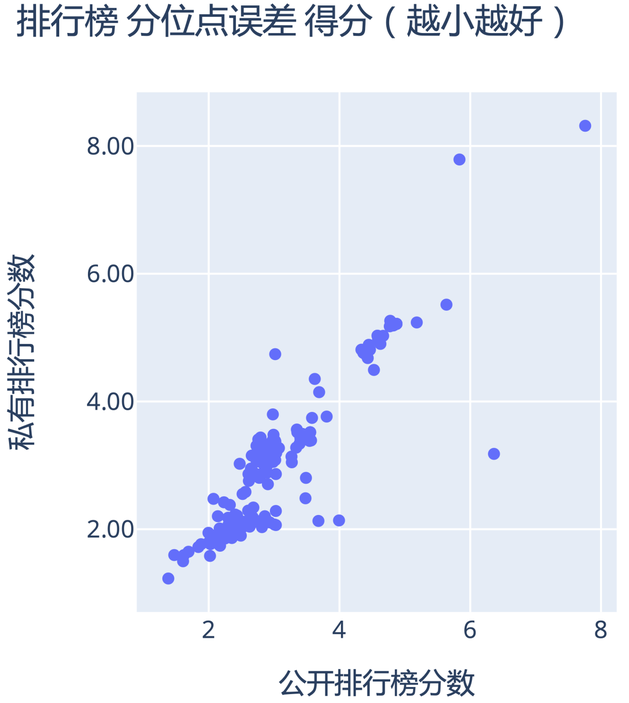
赛题数据大小：22GB

赛题类型：Research、数据挖掘

评价指标：50% 和 95% 分位点误差

报名人数/提交次数：684 / 10270

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第3名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Google AI4Code Understand Code in Python Notebooks

Predict the relationship between code and comments

赛题任务：本次比赛的目的是了解 Python Notebook 中代码和注释之间的关系。您面临的挑战是根据代码单元的顺序重建给定笔记本中单元格的顺序，展示对哪种自然语言引用哪种代码的理解。比赛从 Kaggle 收集了大约 160,000 个公共 Python 笔记本的数据集，参与者使用这个数据集来完成建模。

是否Kernel赛题：是

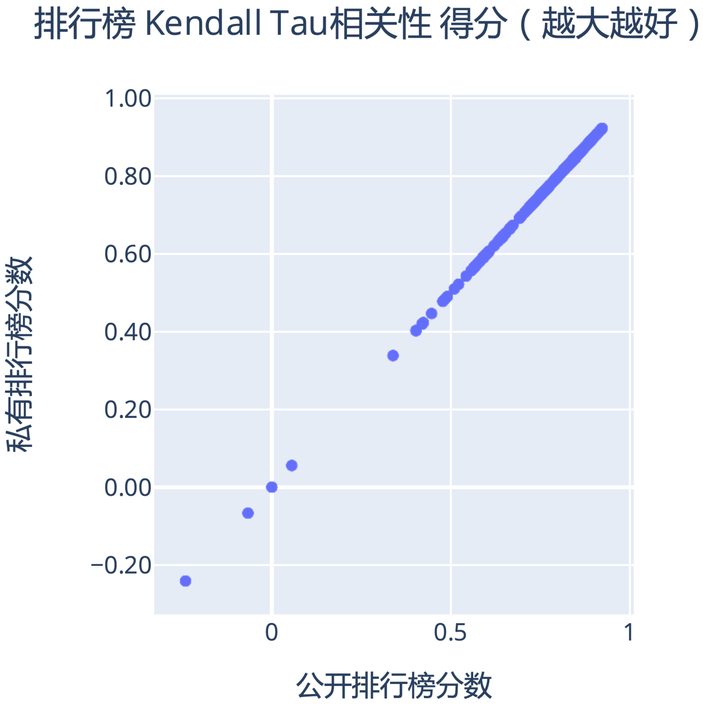
赛题数据大小：2GB

赛题类型：Featured、自然语言处理

评价指标： Kendall Tau 相关性

报名人数/提交次数：

赛题难度：⭐⭐⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案，代码
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Feedback Prize - Predicting Effective Arguments]

Rate the effectiveness of argumentative writing elements from students

赛题任务：本次比赛的目的是将学生写作中的议论文元素分类为“有效”、“适当”或“无效”。您将创建一个模型，该模型根据代表美国 6 至 12 年级人口的数据进行训练，以最大程度地减少偏差。

是否Kernel赛题：是

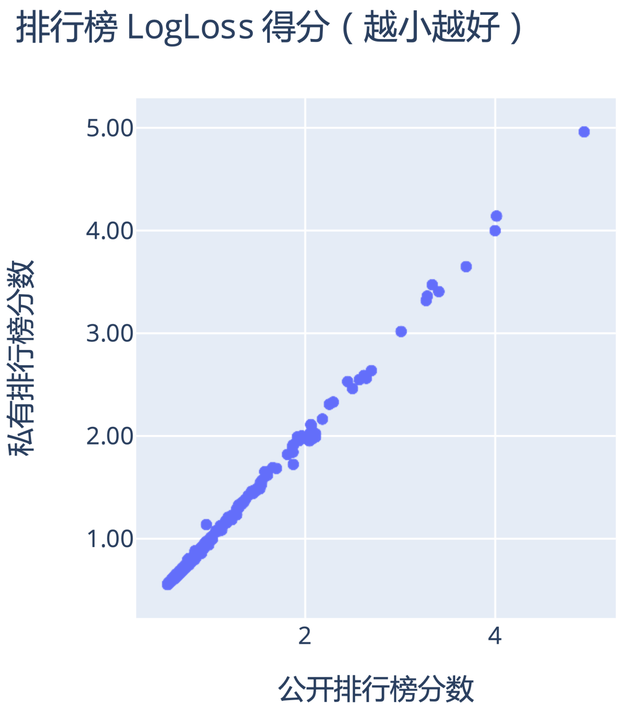
赛题数据大小：20MB

赛题类型：Featured、自然语言处理

评价指标：LogLoss

报名人数/提交次数：1910 / 29139

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案，代码
- ✓ 第2名：方案，代码，代码
- ✓ 第3名：方案，代码
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：American Express - Default Prediction

Predict if a customer will default in the future

赛题任务：信用违约预测是管理消费贷款业务风险的核心。信用违约预测允许贷方优化贷款决策，从而带来更好的客户体验和稳健的商业经济。在本次比赛中，您将运用您的机器学习技能来预测信用违约。您将利用一个工业规模的数据集来构建一个机器学习模型。

是否Kernel赛题：否

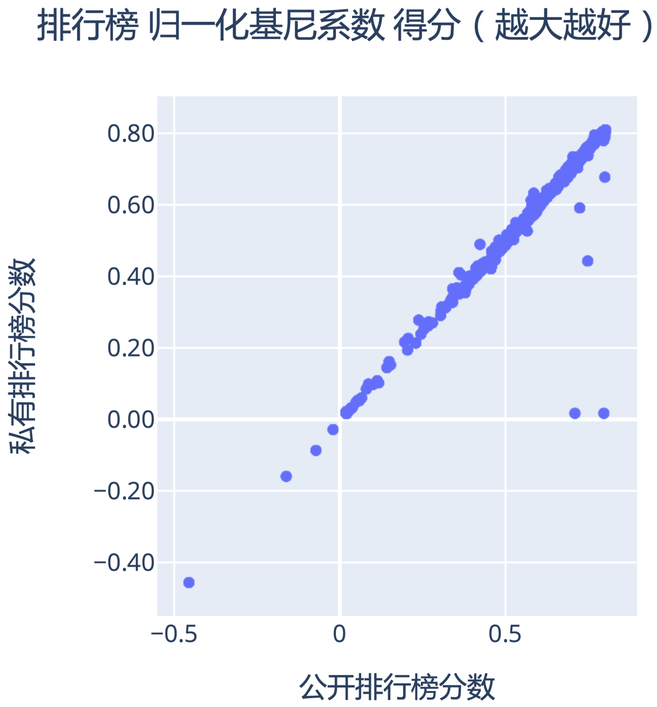
赛题数据大小：50GB

赛题类型：Featured、数据挖掘

评价指标：归一化基尼系数

报名人数/提交次数：6003 / 90058

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Jun 2022

Practice your ML skills on this approachable dataset!

赛题任务：在本次赛题数据中包含了一些缺失值，而参赛选手的任务是对缺失值进行填充

是否Kernel赛题：否

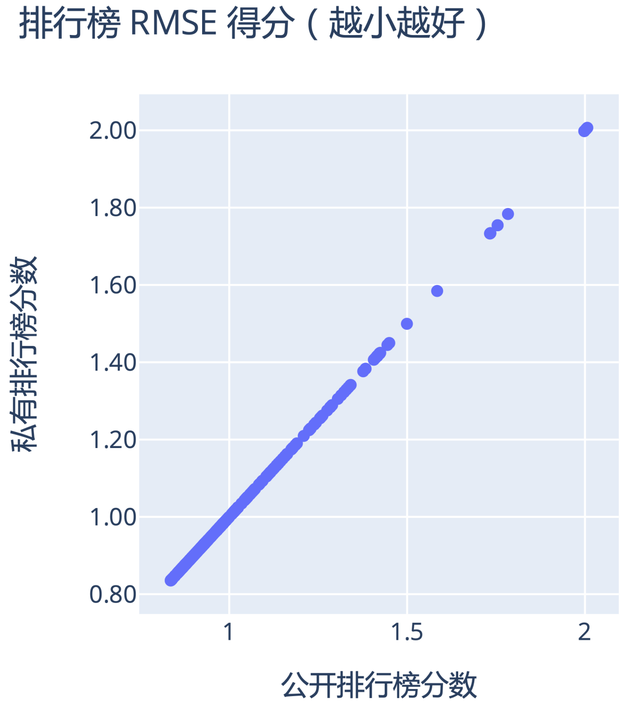
赛题数据大小：44MB

赛题类型：Playground、数据挖掘

评价指标：RMSE

报名人数/提交次数：886 / 5984

赛题难度：⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案，代码
- ✓ 第4名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：HuBMAP + HPA - Hacking the Human Body

Segment multi-organ functional tissue units

赛题任务：人类生物分子图谱计划正在努力创建细胞水平的人类参考图谱。在本次比赛中，您将识别和分割五个人体器官的功能组织单位。您将使用组织切片图像的数据集构建模型，并尽可能准确地对器官进行识别。

是否Kernel赛题：是

赛题数据大小：9GB

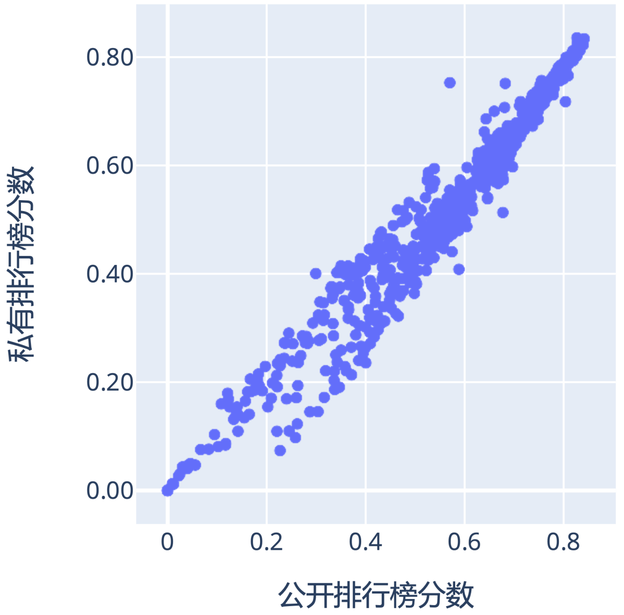
赛题类型：Research、计算机视觉、语义分割

评价指标：Dice 系数

报名人数/提交次数：1517 / 39568

赛题难度：⭐⭐⭐

排行榜 Dice系数 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案，代码
- ✓ 第3名：方案
- ✓ 第4名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Jul 2022

Practice your ML skills on this approachable dataset!

赛题任务：在结构化表格中包含了分组数据，需要参赛选手完成数据聚类过程。

是否Kernel赛题：否

赛题数据大小：44MB

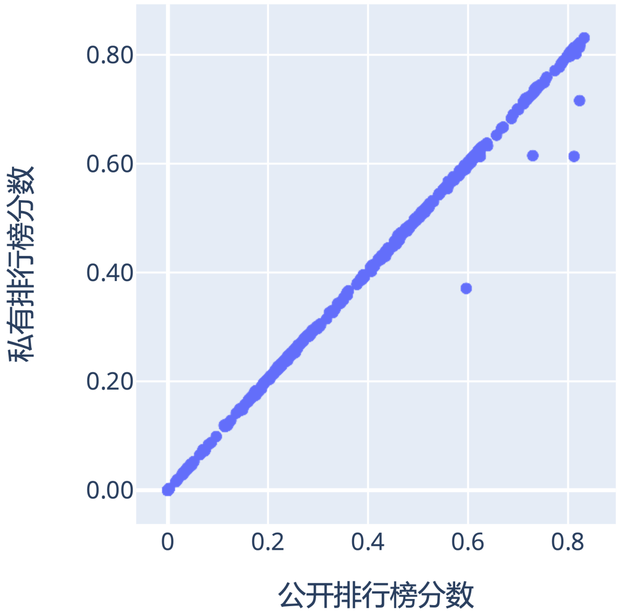
赛题类型：Playground、数据挖掘

评价指标： Adjusted Rand Index

报名人数/提交次数：1278 / 16346

赛题难度：⭐⭐⭐

排行榜 Adjusted Rand Index 得分（越大越好）



✓ 第1名：方案，代码

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Mayo Clinic - STRIP AI

Image Classification of Stroke Blood Clot Origin

赛题任务：本次比赛的目标是对缺血性中风中的血凝块起源进行分类。使用整张幻灯片数字病理图像，您将构建一个模型来区分两种主要的急性缺血性中风和病因亚型。参赛选手的模型将使医疗保健提供者能够更好地识别致命中风中血栓的起源，使医生更容易开出最好的中风后治疗方案，并降低第二次中风的可能性。

是否Kernel赛题：否

赛题数据大小：395GB

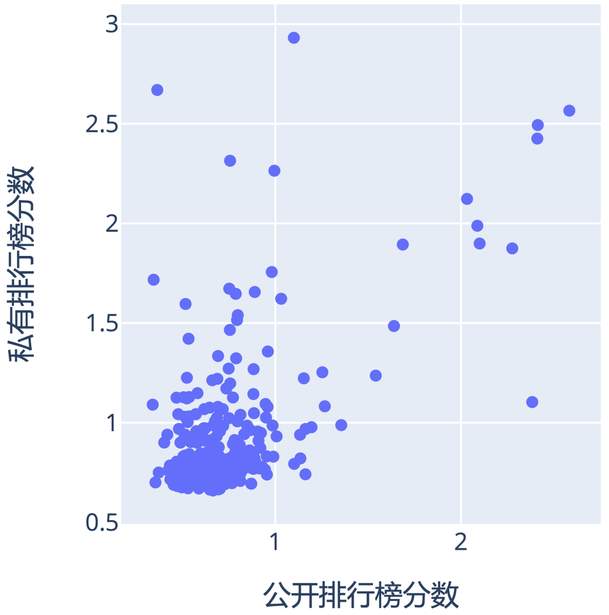
赛题类型：Research、计算机视觉、图像分类

评价指标：LogLoss

报名人数/提交次数：1025 / 6980

赛题难度：⭐⭐⭐⭐

排行榜 LogLoss 得分（越小越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Google Universal Image Embedding

Create image representations that work across many visual domains

赛题任务：图像表示是计算机视觉应用程序的重要组成部分。传统上，图像嵌入学习的研究重点是每个领域的模型。在本次比赛中，开发的模型有望检索与给定查询图像相关的数据库图像。我们数据集中的图像包含各种对象类型，例如服装、艺术品、地标、家具、包装商品等。

是否Kernel赛题：是

赛题数据大小：

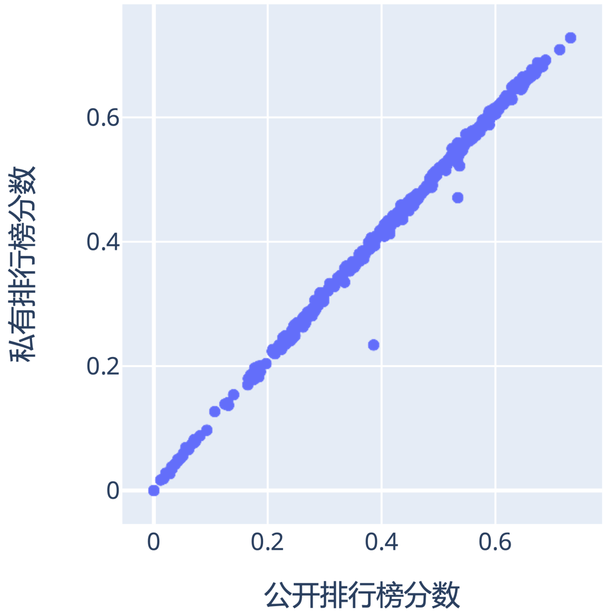
赛题类型：Research、计算机视觉

评价指标：Mean Precision @ 5

报名人数/提交次数：1217 / 20984

赛题难度：⭐⭐⭐⭐⭐

排行榜 Mean Precision@5 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案，代码
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：RSNA 2022 Cervical Spine Fracture Detection

Identify cervical fractures from scans

赛题任务：仅在美国，每年就有超过150万例脊柱骨折发生，成人脊柱骨折的影像学诊断现在几乎完全是通过计算机断层扫描，而不是射线照片。快速检测和确定任何椎骨骨折的位置对于防止创伤后神经功能恶化和瘫痪至关重要。在此挑战赛中，您将尝试开发机器学习模型，对椎骨的骨折进行检测和定位。

是否Kernel赛题：是

赛题数据大小：343GB

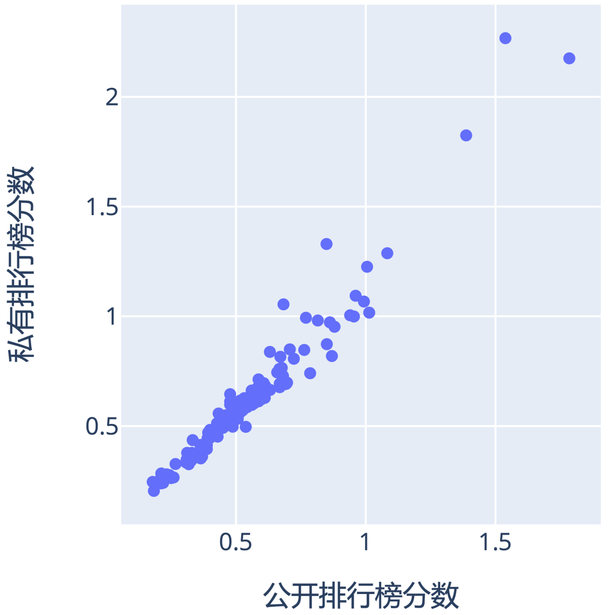
赛题类型：Featured、计算机视觉

评价指标：LogLoss

报名人数/提交次数：1108 / 12871

赛题难度：⭐⭐⭐⭐

排行榜 LogLoss 得分（越小越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案，代码
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：DFL - Bundesliga Data Shootout

Identify plays based upon video footage

赛题任务：您将检测原始德甲比赛中的橄榄球（英式足球）传球（包括界外球和传中）时间。您将开发一个计算机视觉模型，可以自动对长视频记录中的这些事件进行分类。

是否Kernel赛题：是

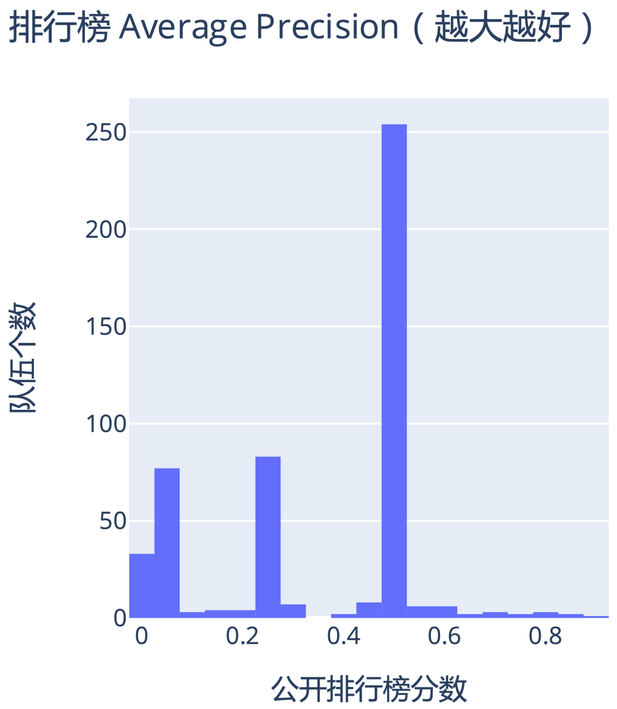
赛题数据大小：3.7GB

赛题类型：Featured、计算机视觉、事件监测

评价指标：Average Precision

报名人数/提交次数：530 / 500

赛题难度：⭐⭐⭐⭐⭐



Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Tabular Playground Series - Aug 2022]

Practice your ML skills on this approachable dataset!

赛题任务：给定产品属性，预测故障的概率。

是否Kernel赛题：否

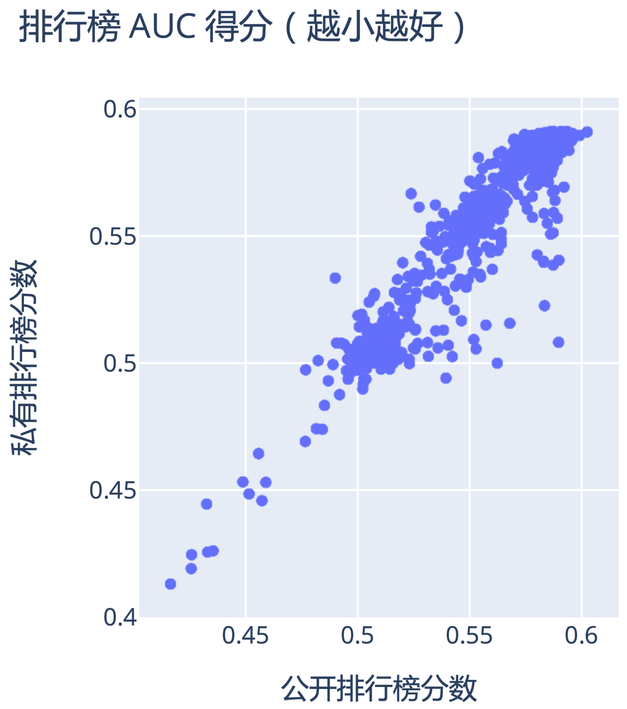
赛题数据大小：7MB

赛题类型：Playground、数据挖掘

评价指标：AUC

报名人数/提交次数：1972 / 21790

赛题难度：⭐⭐



✓ 第1名：[方案]

Part5 比赛内容汇总

Coggle数据科学

赛题名称：AI Village Capture the Flag @ DEFCON

Hack AI! Collect flags by evading, poisoning, stealing, and fooling AI/ML

赛题任务：CTF与常规的Kaggle竞赛在流程上存在不同。在每个CTF挑战期间，参赛者将与存储在输入目录中的 API 端点或代码/对象进行交互。挑战成功完成后，该挑战将返回一个标志。

是否Kernel赛题：否

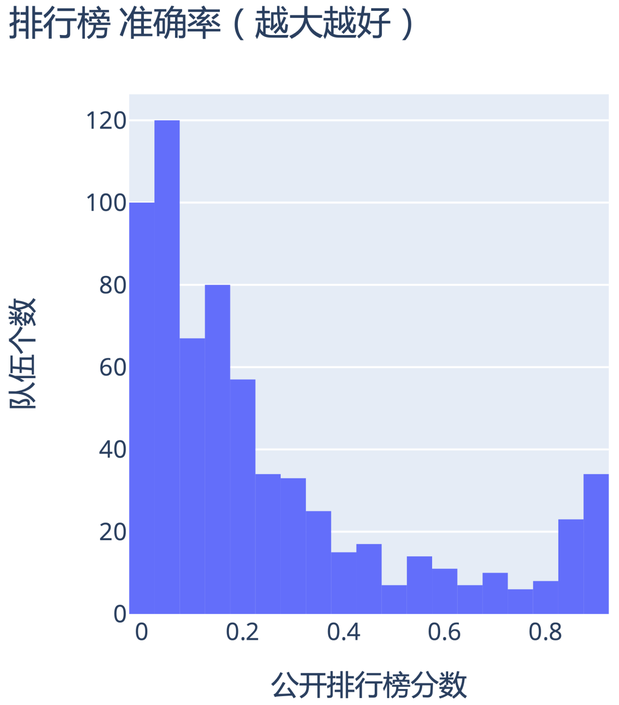
赛题数据大小：212MB

赛题类型：Research、数据挖掘

评价指标：准确率

报名人数/提交次数：668 / 4235

赛题难度：⭐⭐⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

赛题名称：Big Data Derby 2022

Analyze horse racing data to improve the health of the horse and strategy of competition

赛题任务：本次比赛的目标是分析赛马战术、起草策略和路径效率。分析结果将帮助赛马主人、驯马师和兽医更好地了解马匹的性能和福利如何相互配合。通过更好的数据分析，马的福利可以显着改善。

是否Kernel赛题：是

赛题数据大小：977MB

赛题类型：Analytics

评价指标：

报名人数/提交次数：

赛题难度：⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Open Problems - Multimodal Single-Cell Integration

Predict how DNA, RNA & protein measurements co-vary in single cells

赛题任务：本次比赛的目的是预测蛋白质测量值如何在单细胞中变化。您的工作将有助于加速跨细胞状态层映射遗传信息的方法的创新。

是否Kernel赛题：否

赛题数据大小：28GB

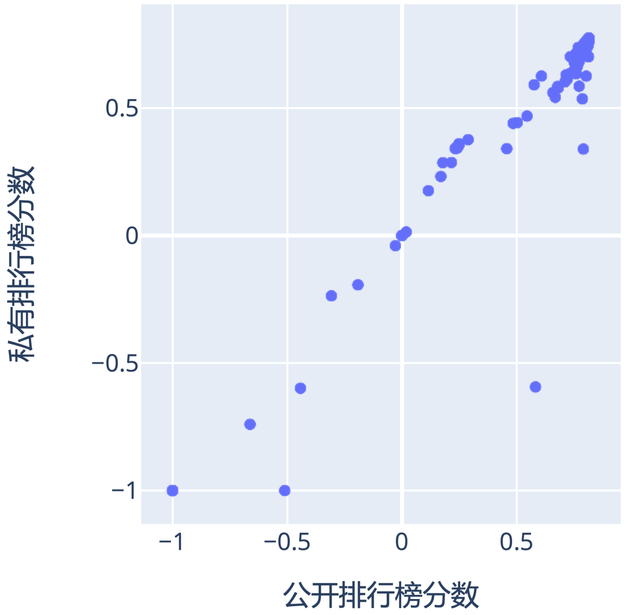
赛题类型：Featured

评价指标：Pearson相关系数

报名人数/提交次数：1602 / 27149

赛题难度：⭐⭐⭐⭐

排行榜 Pearson相关系数 得分（越大越好）



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Feedback Prize - English Language Learning

Evaluating language knowledge of ELL students from grades 8-12

赛题任务：比赛目的是评估 8 至 12 年级英语学习者的语言能力。

是否Kernel赛题：是

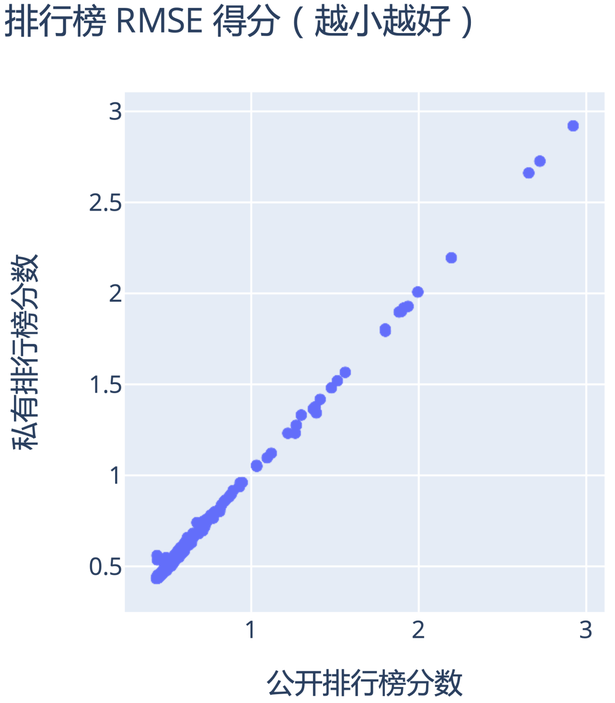
赛题数据大小：9MB

赛题类型：Featured、自然语言处理

评价指标：RMSE

报名人数/提交次数：3273 / 49503

赛题难度：⭐⭐⭐⭐



- ✓ 第1名：方案
- ✓ 第2名：方案
- ✓ 第3名：方案
- ✓ 第4名：方案
- ✓ 第5名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Tabular Playground Series – Sep 2022]

Practice your ML skills on this approachable dataset!

赛题任务：预测图书的未来销量。

是否Kernel赛题：否

赛题数据大小：5MB

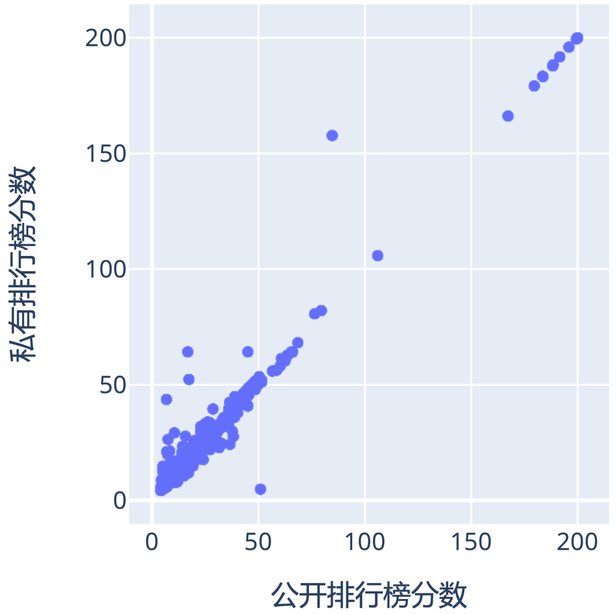
赛题类型：Playground、数据挖掘

评价指标：SMAPE

报名人数/提交次数：1447 / 13085

赛题难度：⭐⭐

排行榜 SMAPE 得分（越小越好）



✓ 第2名：方案

✓ 第3名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Novozymes Enzyme Stability Prediction

Help identify the thermostable mutations in enzymes

赛题任务：酶是在生物体的化学反应中充当催化剂的蛋白质。本次比赛的目标是预测酶变体的热稳定性。

是否Kernel赛题：否

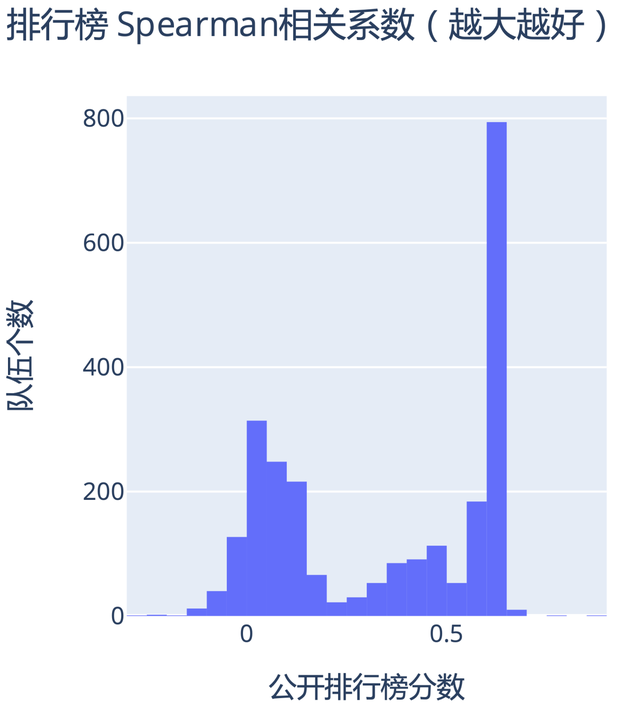
赛题数据大小：16MB

赛题类型：Featured、数据挖掘、生物医学

评价指标：Spearman相关系数

报名人数/提交次数：2463 /

赛题难度：⭐⭐



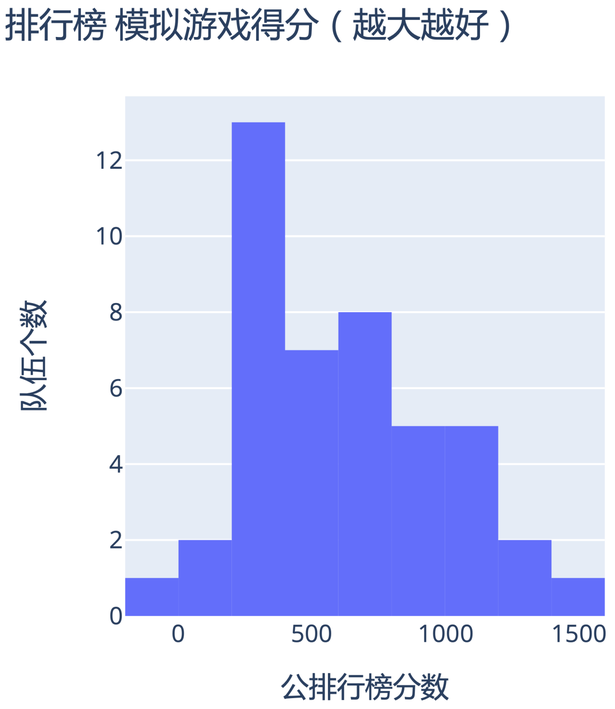
Part5 比赛内容汇总

Coggle数据科学

赛题名称：Lux AI 2022 - Beta

Terraform mars and help test season 2 of the Lux AI Challenge!

赛题任务：Lux AI Challenge 是一项竞赛，参赛者设计代理来解决 1v1 场景中与其他竞争对手的多变量优化、资源收集和分配问题。除了优化之外，成功的智能体还必须能够分析对手并制定适当的策略以取得优势。



是否Kernel赛题：是

赛题数据大小：

赛题类型：Playground、强化学习

评价指标：模拟游戏积分

报名人数/提交次数：44 / 109

赛题难度：⭐⭐⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Oct 2022

Practice your ML skills on this approachable dataset!

赛题任务：根据火箭联盟比赛的历史数据来预测每支球队在接下来的 10 秒内得分的概率。

是否Kernel赛题：否

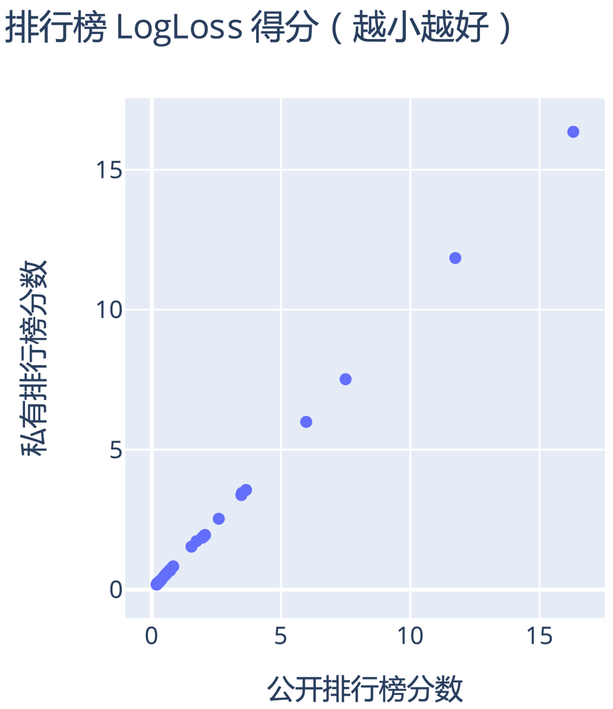
赛题数据大小：9GB

赛题类型：Playground、数据挖掘

评价指标：LogLoss

报名人数/提交次数：500 / 4659

赛题难度：⭐⭐



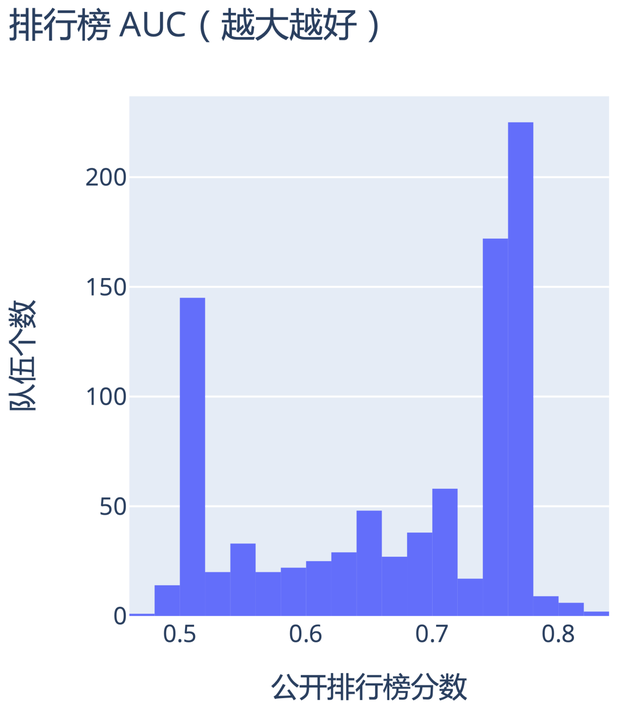
Part5 比赛内容汇总

Coggle数据科学

赛题名称：G2Net Detecting Continuous Gravitational Waves

Help us detect long-lasting gravitational-wave signals!

赛题任务：本次比赛的目标是寻找连续的引力波信号。您将开发一个足够灵敏的模型，以检测噪声数据中快速旋转的中子星发出的微弱但持久的信号。



是否Kernel赛题：否

赛题数据大小：227GB

赛题类型：Research、时序信号

评价指标：AUC

报名人数/提交次数：1131 / 27458

赛题难度：⭐⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：2022 Kaggle Machine Learning & Data Science Survey

The most comprehensive dataset available of ML and data science

赛题任务：今年Kaggle再次发起年度数据科学调查挑战赛，我们想更加深入了解社区逐年发生的变化。

是否Kernel赛题：是

赛题数据大小：26MB

赛题类型：Analytics

评价指标：

报名人数/提交次数：

赛题难度：⭐⭐

- ✓ 第1名：方案
- ✓ 第2名：方案

Part5 比赛内容汇总

赛题名称：NFL Big Data Bowl 2023

Help evaluate linemen on pass plays

赛题任务：在今年的NFL比赛中，参赛选手可以访问 NFL 的 Next Gen Stats 数据，包括球员追踪、比赛、比赛和球员信息。本次比赛希望挖掘NFL传球过程中存在的规律。

是否Kernel赛题：否

赛题数据大小：965MB

赛题类型：Analytics

评价指标：

报名人数/提交次数：

赛题难度：⭐⭐

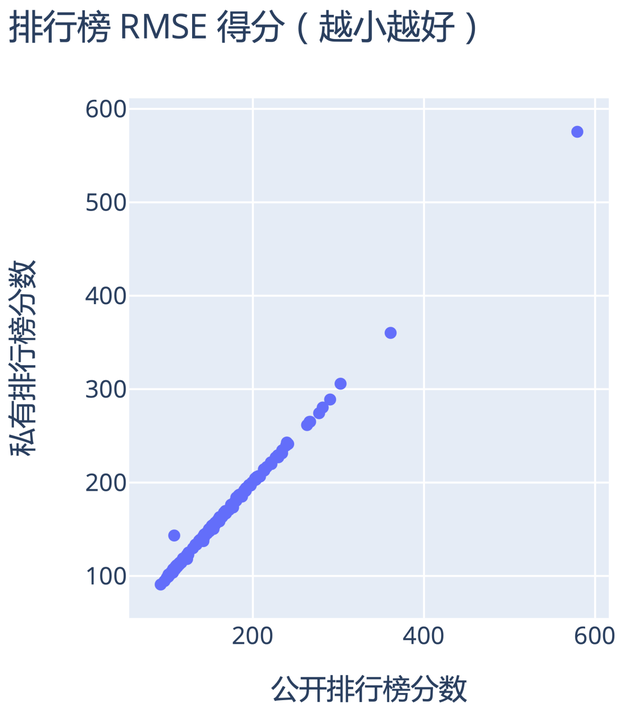
Part5 比赛内容汇总

Coggle数据科学

赛题名称：Scrabble Player Rating

Predict players' ratings based on Woogles.io gameplay

赛题任务：你是 Kaggle 拼字游戏大师吗？在以 Woogles.io 数据为特色的第二届比赛中，参赛者需要根据 Scrabble 游戏玩法预测玩家的评分。



是否Kernel赛题：否

赛题数据大小：122MB

赛题类型：RMSE

评价指标：

报名人数/提交次数：

赛题难度：⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：OTTO – Multi-Objective Recommender System

Build a recommender system based on real-world e-commerce sessions

赛题任务：本次比赛的目标是预测电子商务点击、购物车添加和订单。您将基于用户会话中的先前事件构建多目标推荐系统。

是否Kernel赛题：否

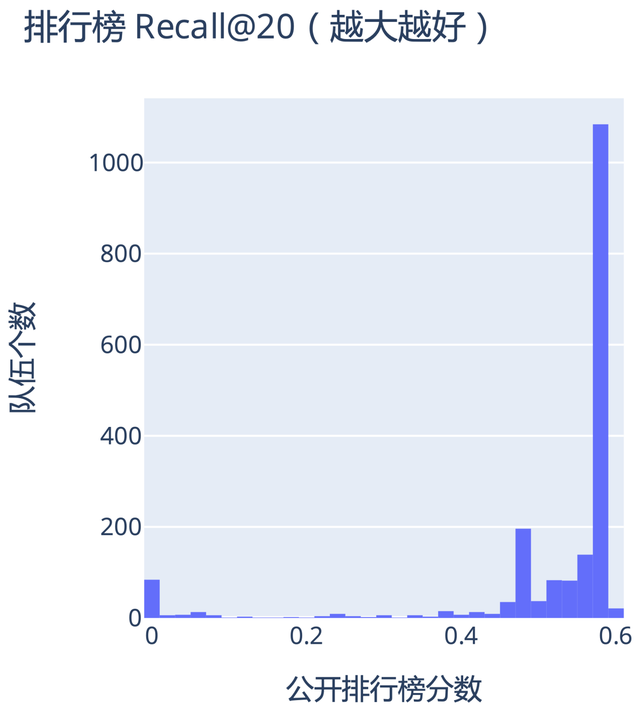
赛题数据大小：12GB

赛题类型：Featured、推荐系统

评价指标：Recall@20

报名人数/提交次数：2157 / 13836

赛题难度：⭐⭐⭐⭐



Part5 比赛内容汇总

Coggle数据科学

赛题名称：Tabular Playground Series - Nov 2022

Practice your ML skills on this approachable dataset!

赛题任务：对给定的数据构建二分类模型

是否Kernel赛题：否

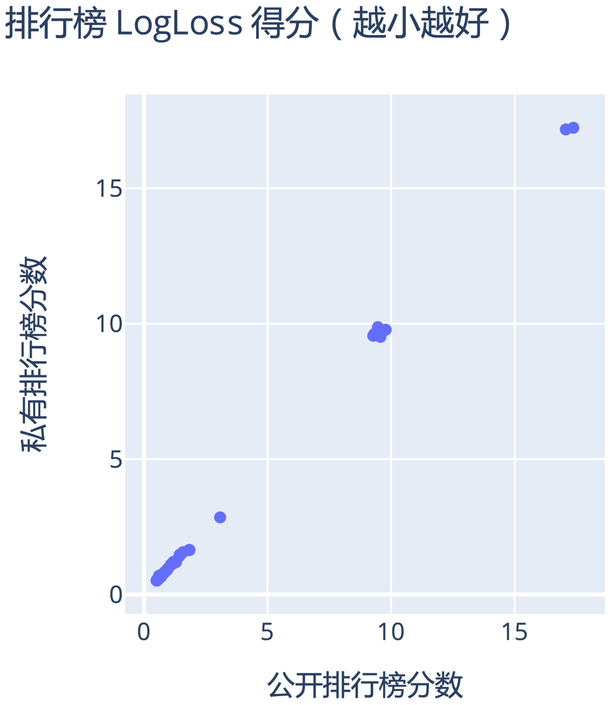
赛题数据大小：3GB

赛题类型：Playground、数据挖掘

评价指标：LogLoss

报名人数/提交次数：717 / 7260

赛题难度：⭐⭐



✓ 第1名：方案

✓ 第3名：方案

Part5 比赛内容汇总

Coggle数据科学

赛题名称：RSNA Mammography Breast Cancer Detection

Find breast cancers in screening mammograms

赛题任务：本次比赛的目标是识别乳腺癌。您将使用从定期筛查中获得的筛查性乳房 X 线照片来训练您的模型。

是否Kernel赛题：是

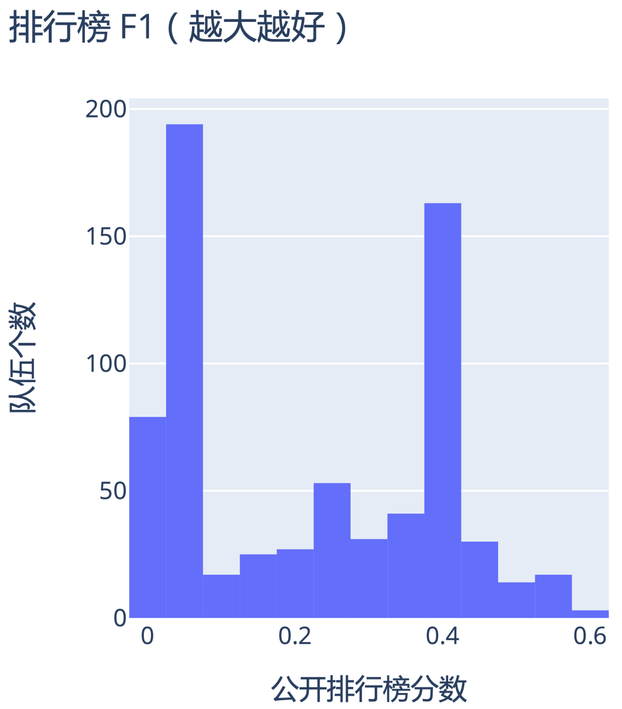
赛题数据大小：314GB

赛题类型：Featured、计算机视觉、图像分类

评价指标：F1

报名人数/提交次数：750 / 7893

赛题难度：⭐⭐⭐



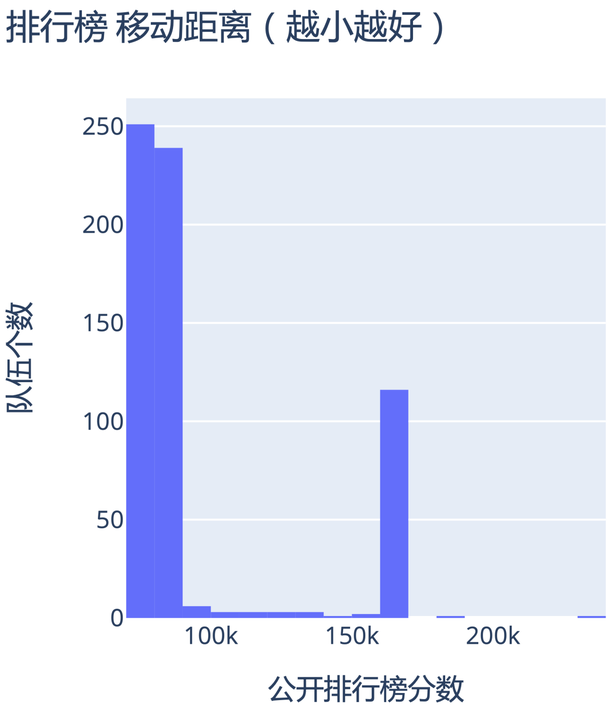
Part5 比赛内容汇总

Coggle数据科学

赛题名称：[Santa 2022 - The Christmas Card Conundrum]

Optimize the configuration space for printing an image

赛题任务：圣诞老人的精灵们每年都依赖同一个供应商来印刷一年一度的圣诞贺卡。赛题人数是确定制作今年圣诞贺卡的最佳方式，方法是选择移动机械臂和更改打印颜色的最有效路径来制作今年的图像。



是否Kernel赛题：否

赛题数据大小：10MB

赛题类型：Featured、组合优化

评价指标：移动距离

报名人数/提交次数：

赛题难度：⭐⭐⭐⭐

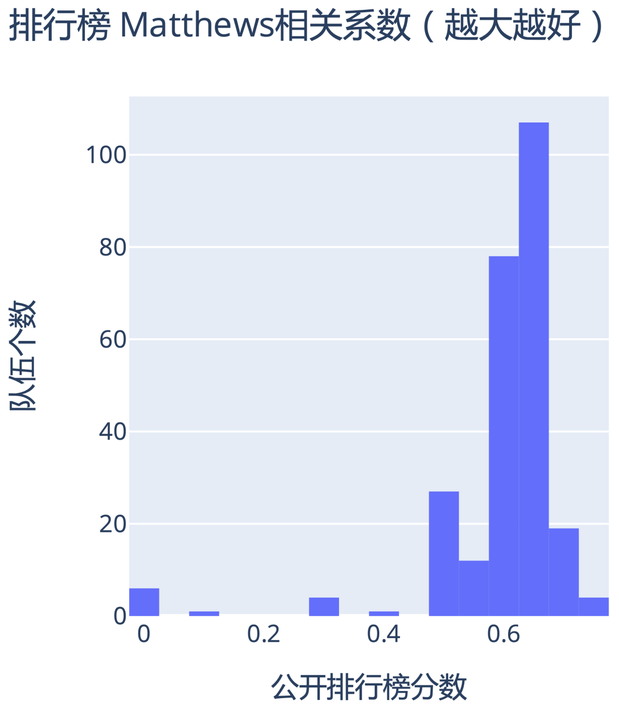
Part5 比赛内容汇总

Coggle数据科学

赛题名称：1st and Future - Player Contact Detection

Detect Player Contacts from Sensor and Video Data

赛题任务：本次比赛的目标是检测球员在 NFL 橄榄球比赛中经历的外部接触的时间。您将使用视频和玩家跟踪数据来识别接触时刻，以帮助提高玩家安全性。



是否Kernel赛题：是

赛题数据大小：5GB

赛题类型：Featured、计算机视觉、事件监测

评价指标：Matthews相关系数

报名人数/提交次数：265 / 1498

赛题难度：⭐⭐⭐⭐

Part5 比赛内容汇总

Coggle数据科学

赛题名称：Learning Equality - Curriculum Recommendations

Enhance learning by matching K-12 content to target topics

赛题任务：将教育内容与课程中的特定主题相匹配的过程。

是否Kernel赛题：是

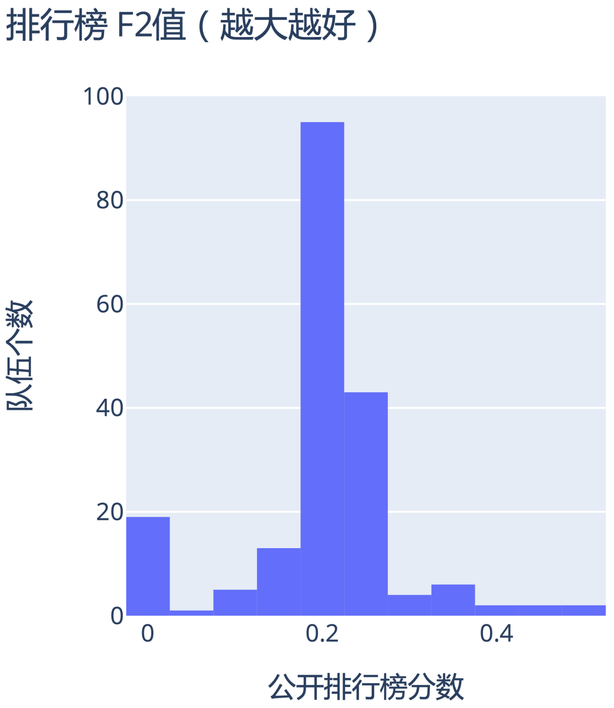
赛题数据大小：891MB

赛题类型：Featured、自然语言处理

评价指标：F2值

报名人数/提交次数：204 / 1427

赛题难度：⭐⭐⭐⭐



Part5 比赛内容汇总

Coggle数据科学

赛题名称：GoDaddy - Microbusiness Density Forecasting

Forecast Next Month's Microbusiness Density

赛题任务：本次比赛的目标是预测给定地区每月的微型企业密度。您将开发一个根据美国县级数据训练的准确模型。

是否Kernel赛题：是

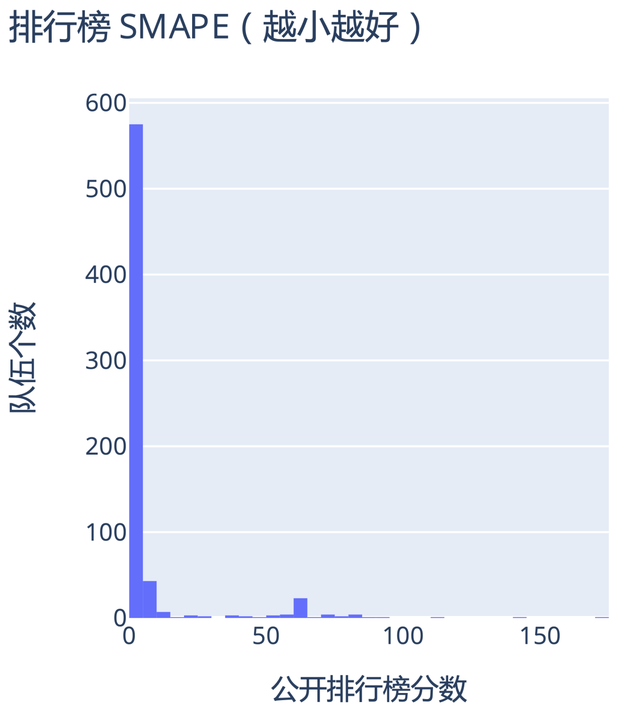
赛题数据大小：10MB

赛题类型：Featured、数据挖掘

评价指标：SMAPE

报名人数/提交次数：661 / 4082

赛题难度：⭐⭐⭐



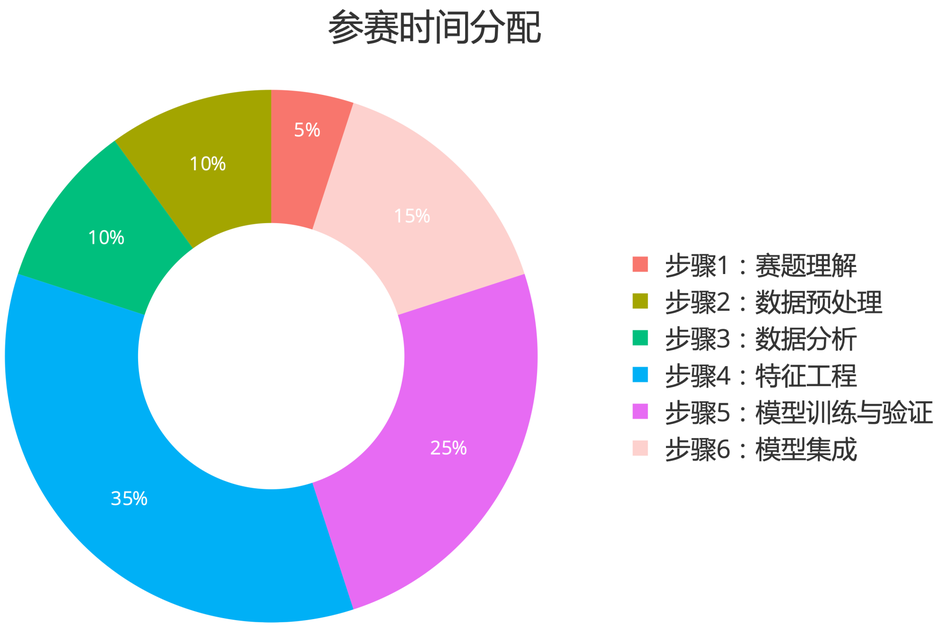
PART 06

Kaggle学习路径

Part6 Kaggle学习路径

【参赛建议】：数据挖掘类型

- ✓ 赛题难度：入门、进阶赛题居多
- ✓ 参赛建议：适合小白入门，对机器配置要求低
- ✓ 常见赛题方向：二分类、多分类、回归、时序预测
- ✓ 必备Python库：Pandas、Sklearn、XGBoost、LightGBM、CatBoost
- ✓ 常见模型：树模型和集成学习居多



入门练习赛题：

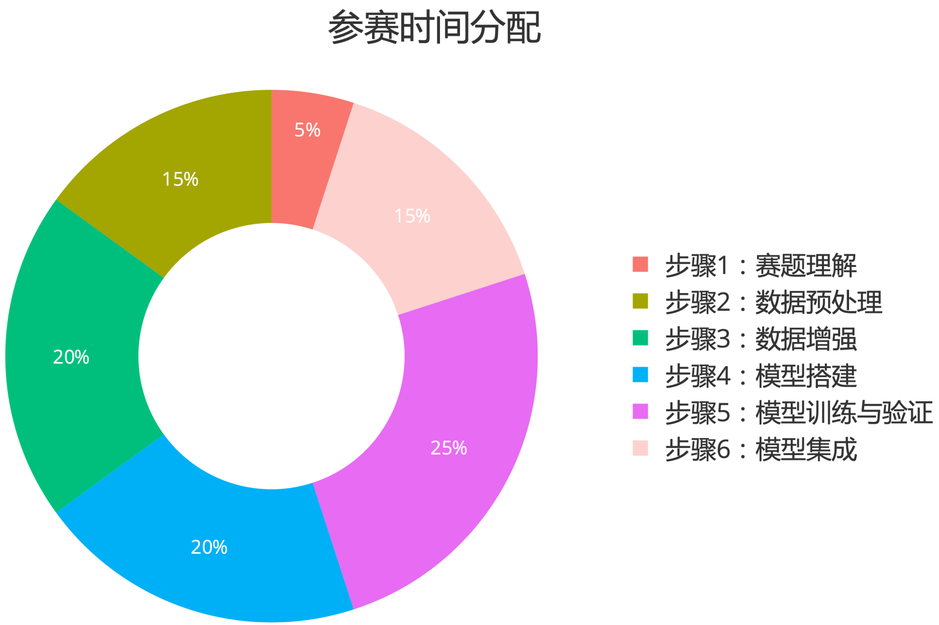
- ✓ Titanic - Machine Learning from Disaster
- ✓ House Prices - Advanced Regression Techniques
- ✓ Spaceship Titanic

Part6 Kaggle学习路径

Coggle数据科学

【参赛建议】：计算机视觉类型

- ✓ 赛题难度：进阶和较难居多，需要GPU支持
- ✓ 参赛建议：适合学习深度学习入门，建议以分类赛题入门
- ✓ 常见赛题方向：图像多分类、细粒度分类、语义分割
- ✓ 必备Python库：Pytorch、TensorFlow、timm
- ✓ 常见模型：CNN模型、transformer模型



入门练习赛题：

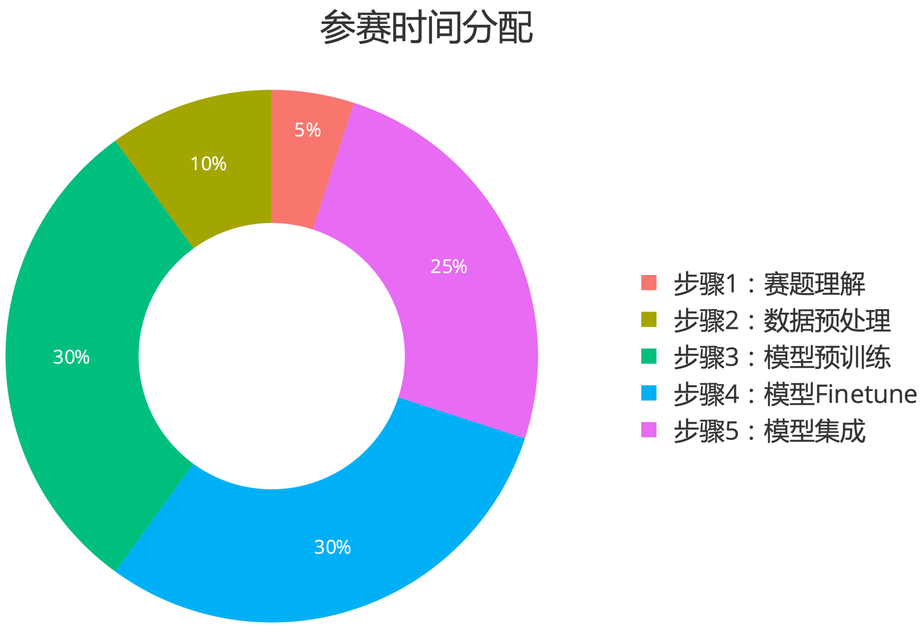
- ✓ Natural Language Processing with Disaster Tweets

Part6 Kaggle学习路径

Coggle数据科学

【参赛建议】：自然语言处理类型

- ✓ 赛题难度：进阶比赛
- ✓ 参赛建议：适合学习或了解BERT的同学
- ✓ 常见赛题方向：文本分类、实体抽取、文本匹配
- ✓ 必备Python库：NLTK、gensim、Pytorch、transformers
- ✓ 常见模型：BERT模型和相关变种



入门练习赛题：

- ✓ Facial Keypoints Detection

Part6 Kaggle学习路径

【学习资料】：Pandas

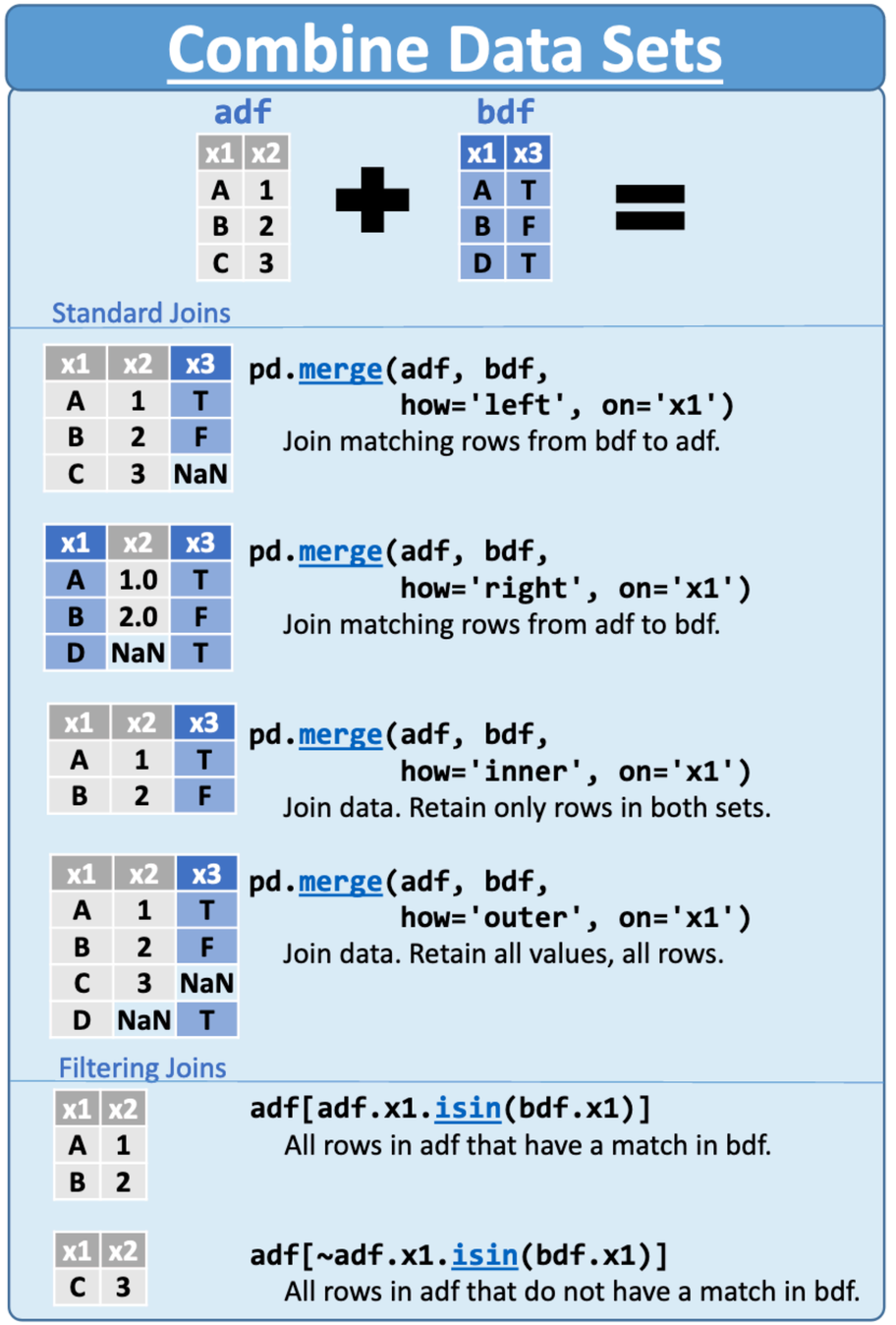
Pandas可以快读读取结构化数据，并进行分析和统计

✓ 官网文档，使用案例

✓ 学习难度：⭐⭐⭐

✓ 学习要点：

- ✓ 【基础】使用Pandas读取数据、索引数据、对数据进行排序
- ✓ 【进阶】使用Pandas进行统计、分组统计 & 计数、交叉表
- ✓ 【进阶】使用Pandas进行特征编码（文本、类别、数值、列表）
- ✓ 【进阶】使用Pandas进行时序数据处理、相关性计算
- ✓ 【深入】使用Numba、Cython或多线性加速计算



Part6 Kaggle学习路径

Coggle数据科学

【学习资料】：scikit-learn

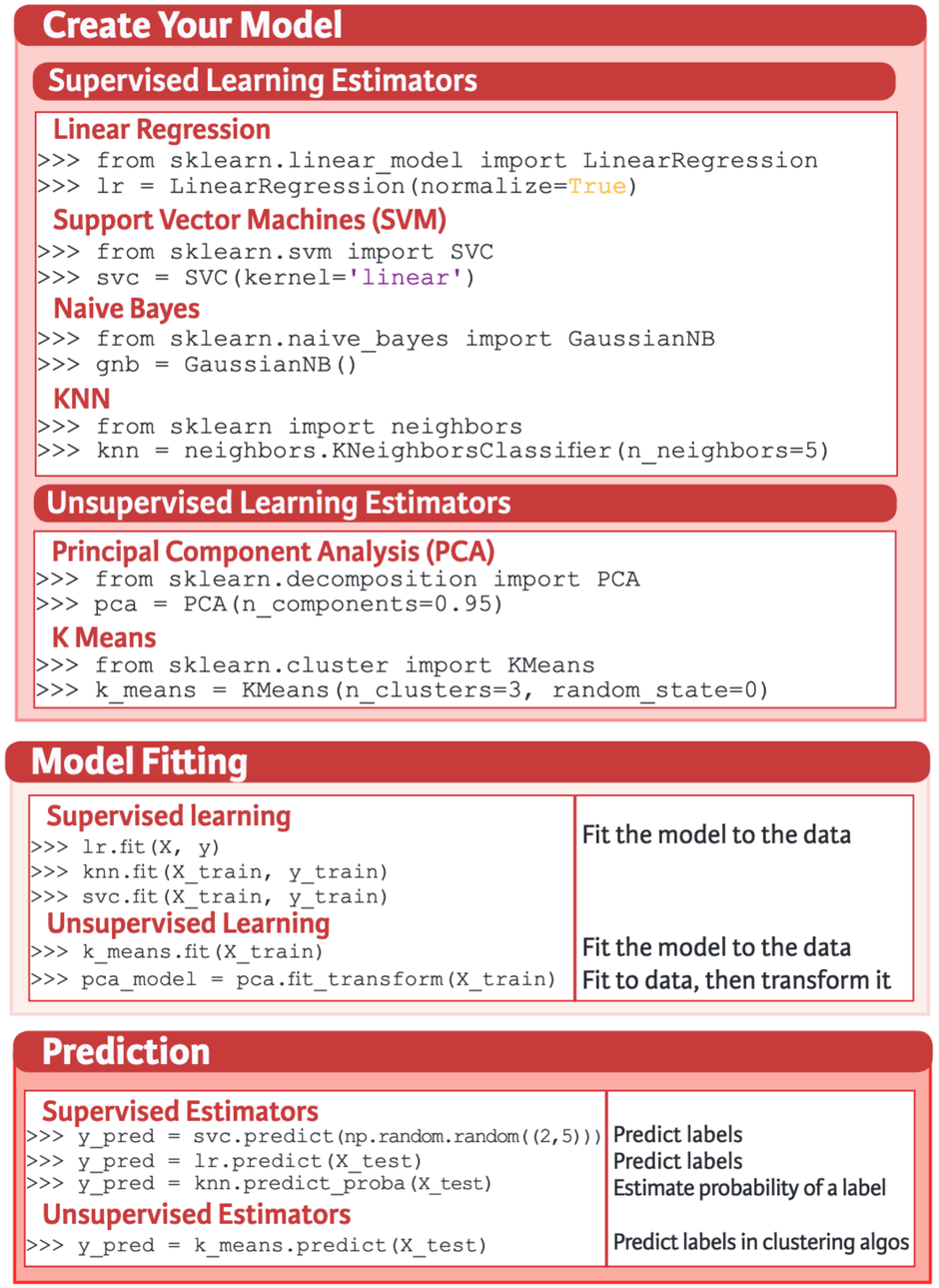
scikit-learn（sklearn）包含了众多机器学习模型、数据划分和评价方法

✓ 官网文档，使用案例

✓ 学习难度：⭐⭐⭐⭐

✓ 学习要点：

- ✓ 【基础】掌握sklearn各种分类、回归、聚类、降维方法的使用
- ✓ 【基础】掌握sklearn数据划分、特征编码的使用
- ✓ 【进阶】掌握sklearn特征筛选方法的使用
- ✓ 【进阶】掌握sklearn模型调参方法的使用
- ✓ 【深入】对比多个sklearn模型的优缺点，并进行模型集成



Part6 Kaggle学习路径

Coggle数据科学

【学习资料】：Matplotlib / Seaborn

Matplotlib包含了众多基础的可视化方法，比较灵活可配置。

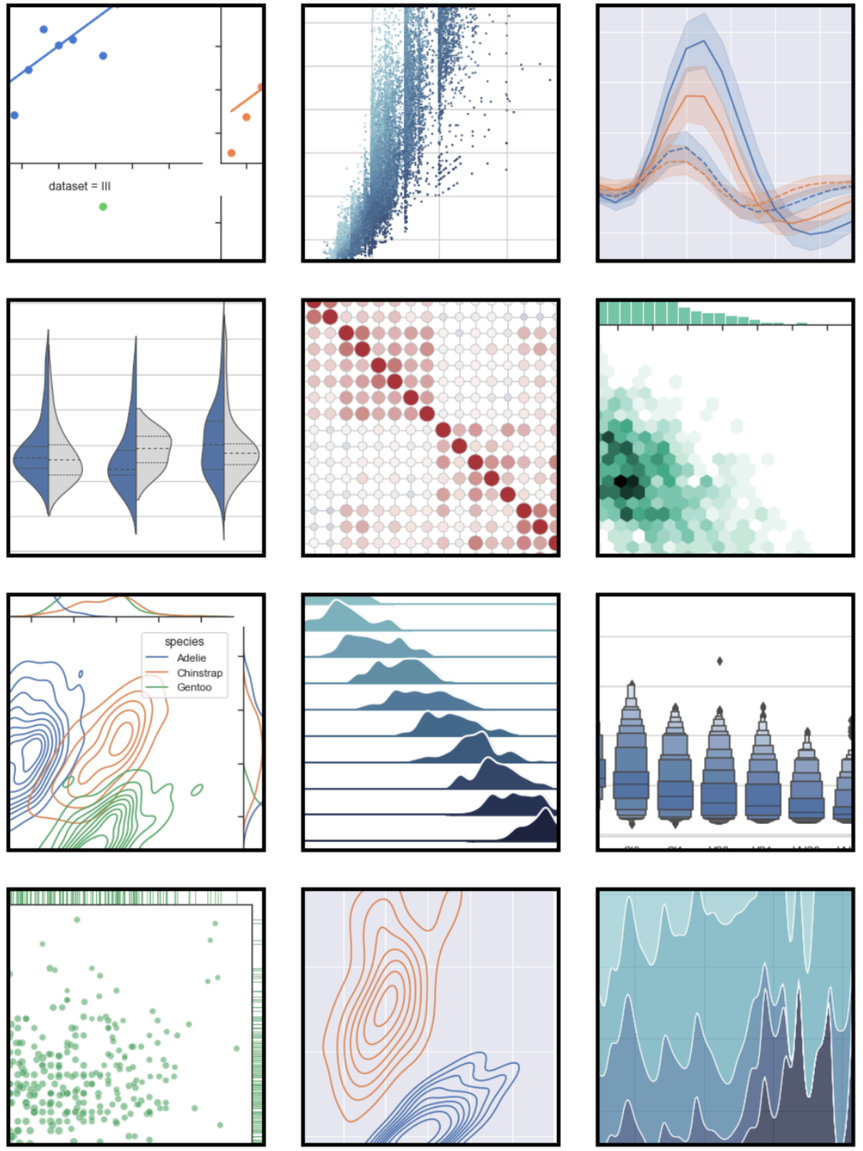
Seaborn基于Matplotlib，包含高阶的可视化方法，适合结合Pandas使用。

✓ Matplotlib使用案例，Seaborn使用案例

✓ 学习难度：⭐⭐

✓ 学习要点：

- ✓ 【基础】能够使用Matplotlib绘制饼图、柱状图、折线图
- ✓ 【基础】能够使用Matplotlib子图，并设置图表标题、坐标等配置
- ✓ 【进阶】能够使用Seaborn绘制箱线图、小提琴图和密度图
- ✓ 【进阶】能够对不同的数据类型，选择Seaborn中合适的画图函数



Part6 Kaggle学习路径

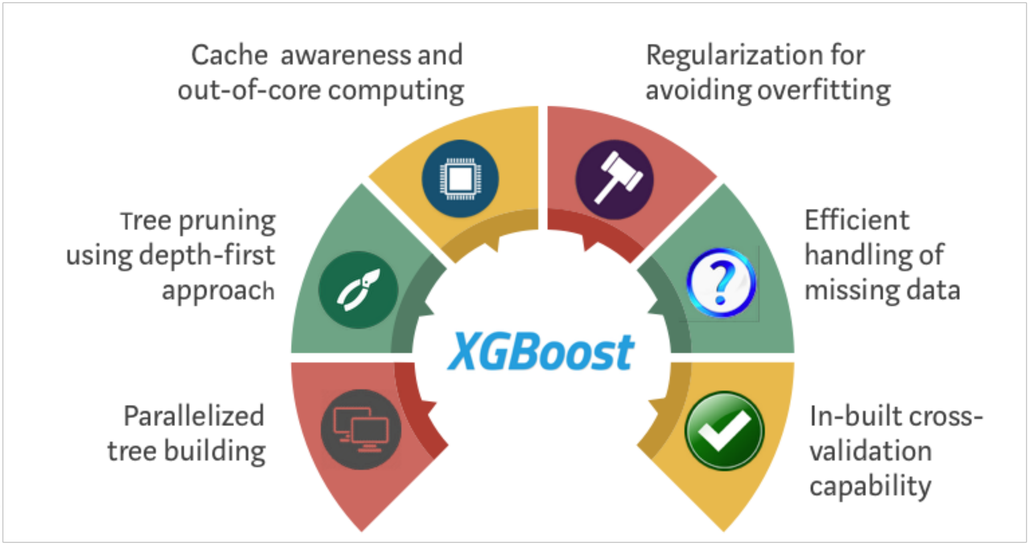
Coggle数据科学

【学习资料】：XGBoost / LightGBM / CatBoost

XGBoost / LightGBM / CatBoost是高阶的树模型，是数据挖掘竞赛必备库。

✓ XGBoost使用案例，LightGBM使用案例，CatBoost使用案例

✓ 学习难度：⭐⭐⭐⭐⭐



✓ 学习要点：

- ✓ 【基础】能完成训练和与预测
- ✓ 【进阶】能进行交叉验证进行验证和预测，能使用Early Stop
- ✓ 【进阶】能使用GPU进行训练和预测，并对类别进行编码
- ✓ 【进阶】能对模型进行可视化，计算特征重要性，并进行特征筛选
- ✓ 【深入】理解模型超参数含义，会对模型进行调参
- ✓ 【深入】能自定义损失函数与评价函数

Part6 Kaggle学习路径

Coggle数据科学

【学习资料】：Pytorch

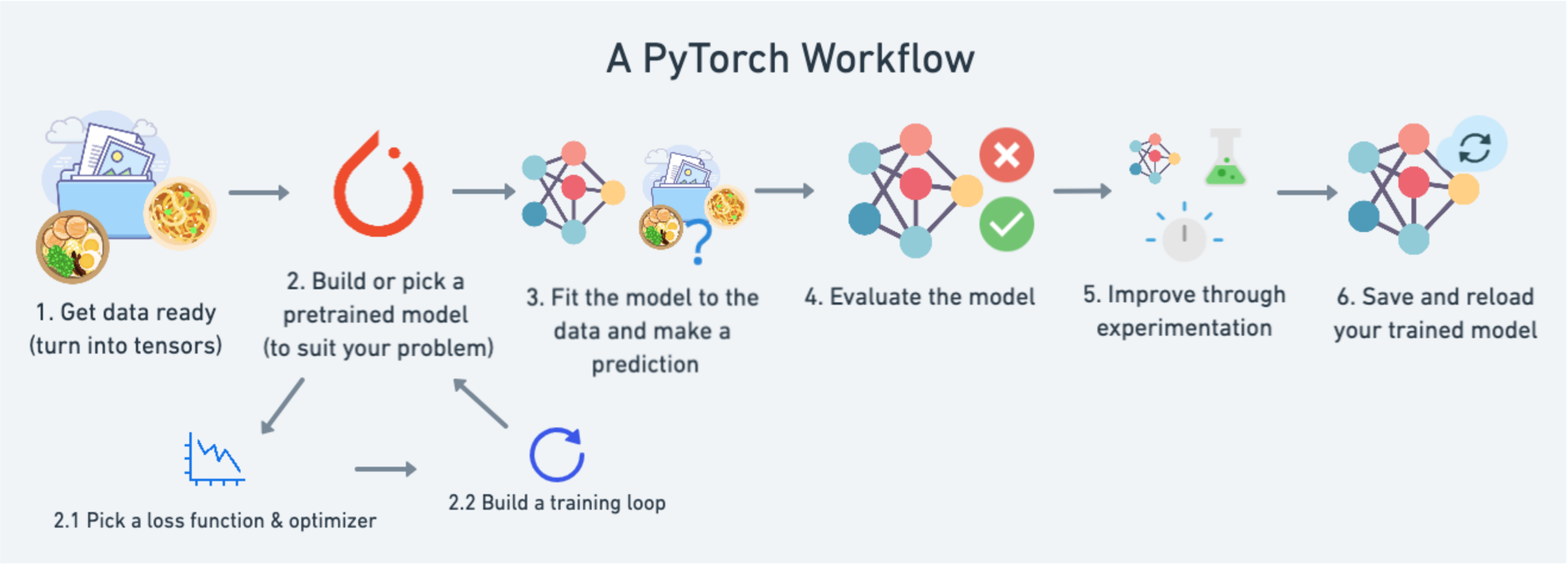
Pytorch是常见的深度学习库，支持深度学习模型搭建和训练

✓ Pytorch官网文档，Pytorch使用案例

✓ 学习难度：⭐⭐⭐⭐

✓ 学习要点：

- ✓ 【基础】能读取自定义数据集、搭建模型
- ✓ 【入门】能完成模型训练、验证、预测
- ✓ 【基础】能加载预训练模型，进行finetune



Part6 Kaggle学习路径

Coggle数据科学

【学习资料】：BERT & transformers

BERT是NLP比赛必备模型，transformers加载和训练BERT的库。

✓ transformers官方文档，transformers使用案例

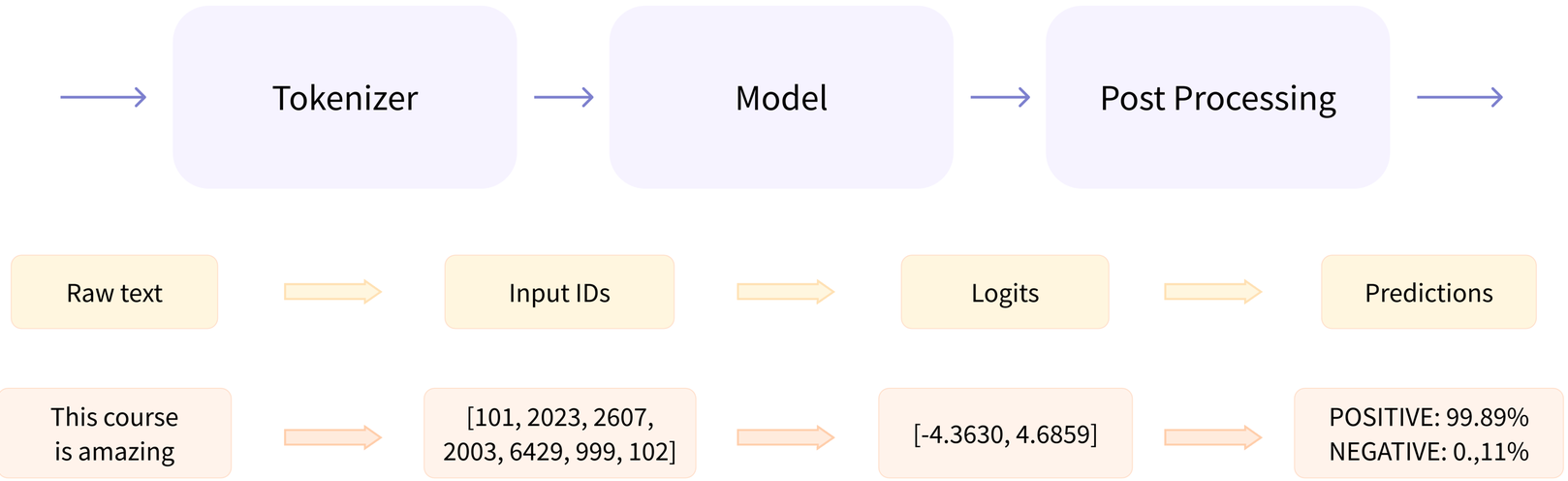
✓ 学习难度：⭐⭐⭐⭐

✓ 学习要点：

✓ 【入门】能加载BERT模型，对文本进行编码

✓ 【进阶】能使用BERT完成分类、匹配、NER

✓ 【深入】能在自定义数据集上进行预训练



Part6 Kaggle学习路径

Coggle数据科学

全网最全 & 不定期更新
比赛干货 & Kaggle学习路线
https://coggle.club/blog/kaggle-roadmap

Coggle数据科学

感谢观看

Thanks for watching

2022

Coggle数据科学