“Codeforces 用户数据分析”进展报告

> 撰写时间：2024年5月24日
>
> 撰写人：廖嘉琦

1. 数据获取及预处理

> 主要说明数据的来源，原始数据的基本情况，如数量，字段，含义等， 到目前为止，对数据预处理的情况，如噪声的处理，缺失值的处理等。

1.1 数据来源

原始数据有两个来源：

- Codeforces API：主要来源，获取官方提供的结构化的JSON数据；
- Codeforces爬虫：辅助手段，获取一些API未直接提供，但对分析有用的数据。

1.2 数据说明

数据将涵盖以下几个主要方面：

- **竞赛数据**：利用`contest.list`、`contest.standings`和`contest.status`接口（以及爬虫），获取竞赛的基本信息、排名和提交记录。
- **用户数据**：通过`user.info`、`user.status`和`user.rating`接口，收集选手的基本信息、提交历史和等级变化。
- **社区互动数据**：通过`blogEntry.view`、`blogEntry.comments`和`user.blogEntries`接口，分析社区的讨论热度和互动模式。

例如：

- 竞赛数据：`1853 rows × 8 columns`
  - 来源：API提供

	durationSeconds	name	type	phase	startTime	frozen	id	relativeTimeSeconds
0	10800	Codeforces Round (Div. 1 + Div. 2)	CF	BEFORE	2024-04-06 14:35:00	False	1951	-777732
1	7200	April Fools Day Contest 2024	ICPC	BEFORE	2024-04-01 14:35:00	False	1952	-345732
2	10800	CodeTON Round 8 (Div. 1 + Div. 2, Rated, Prizes!)	CF	BEFORE	2024-03-30 14:35:00	False	1942	-172932
3	8100	Codeforces Round 937 (Div. 4)	ICPC	BEFORE	2024-03-28 14:45:00	False	1950	-734
4	18000	European Championship 2024 - Online Mirror (Un...	ICPC	FINISHED	2024-03-24 10:00:00	False	1949	361968

	durationSeconds	name	type	phase	startTime	frozen	id	relativeTimeSeconds
...	...	...	...	...	...	...	...	...
1848	7200	Codeforces Beta Round 5	ICPC	FINISHED	2010-03-20 16:00:00	False	5	442535569
1849	7200	Codeforces Beta Round 4 (Div. 2 Only)	ICPC	FINISHED	2010-03-12 12:00:00	False	4	443241169
1850	7200	Codeforces Beta Round 3	ICPC	FINISHED	2010-03-07 12:00:00	False	3	443673169
1851	7200	Codeforces Beta Round 2	ICPC	FINISHED	2010-02-25 17:00:00	False	2	444519169
1852	7200	Codeforces Beta Round 1	ICPC	FINISHED	2010-02-19 12:00:00	False	1	445055569

API获取两类ProblemSet

```
2024-03-28 22:32:47,060 - root - INFO - Start fetching contests
2024-03-28 22:33:09,592 - root - INFO - Fetched 1853 contests hosted by Codeforces
2024-03-28 22:33:09,629 - root - INFO - Fetched 1887 contests in gym
```

- 题目数据：`9187 rows × 8 columns`
  - 来源：API+爬虫

	name	type	rating	tags	contestId	points	solvedCount	index
0	Amanda the Amoeba	PROGRAMMING	NaN	['graphs', 'implementation', 'trees', 'two poi...	1949	NaN	193	J
1	Clique Partition	PROGRAMMING	2100.0	['brute force', 'constructive algorithms', 'gr...	1948	NaN	2164	E
2	Array Fix	PROGRAMMING	1100.0	['brute force', 'dp', 'greedy', 'implementation']	1948	NaN	17820	B
3	Birthday Gift	PROGRAMMING	1900.0	['bitmasks', 'brute force', 'constructive algo...	1946	1750.0	3031	D
4	Tree Cutting	PROGRAMMING	1600.0	['binary search', 'dp', 'greedy', 'implementat...	1946	1500.0	7101	C

	name	type	rating	tags	contestId	points	solvedCount	index
...	...	...	...	...	...	...	...	...
9182	Circular RMQ	PROGRAMMING	2200.0	['data structures']	52	1500.0	7983	C
9183	Dancing Lessons	PROGRAMMING	1900.0	['data structures']	45	NaN	842	C
9184	Queue	PROGRAMMING	2300.0	['data structures']	38	NaN	670	G
9185	Points	PROGRAMMING	2800.0	['data structures']	19	NaN	2032	D
9186	Bindian Signalizing	PROGRAMMING	2400.0	['data structures']	5	NaN	1987	E

爬取ProblemSet获取所有Tag：

```
2024-03-28 21:44:16,346 - root - INFO - Start fetching tags from page 1 to 94
2024-03-28 21:44:16,622 - root - WARNING - Failed to fetch
https://codeforces.com/problemset/page/1 for 1 times
2024-03-28 21:44:27,754 - root - INFO - Fetched 29 tags from page 1, 29 tags in total
2024-03-28 21:44:39,814 - root - INFO - Fetched 30 tags from page 2, 32 tags in total
2024-03-28 21:44:50,655 - root - INFO - Fetched 31 tags from page 3, 32 tags in total
2024-03-28 21:45:04,617 - root - INFO - Start fetching tags from page 1 to 94
2024-03-28 21:45:07,647 - root - INFO - Fetched 29 tags from page 1, 29 tags in total
2024-03-28 21:45:07,647 - root - INFO - tags: {'ternary search', 'dsu', 'fft',
'combinatorics', 'two pointers', 'greedy', 'bitmasks', 'divide and conquer', 'graph
matchings', 'dfs and similar', 'hashing', 'sortings', 'constructive algorithms', 'shortest
paths', 'implementation', 'data structures', 'math', 'games', 'probabilities', 'number
theory', 'dp', 'geometry', 'graphs', 'interactive', '*special problem', 'trees', 'binary
search', 'strings', 'brute force'}
...
2024-03-28 22:03:08,564 - root - INFO - Fetched 30 tags from page 90, 37 tags in total
2024-03-28 22:03:20,207 - root - INFO - Fetched 28 tags from page 91, 37 tags in total
2024-03-28 22:03:31,645 - root - INFO - Fetched 26 tags from page 92, 37 tags in total
2024-03-28 22:03:42,863 - root - INFO - Fetched 25 tags from page 93, 37 tags in total
2024-03-28 22:03:55,080 - root - INFO - Fetched 26 tags from page 94, 37 tags in total
```

将Tag与题目匹配：

```
2024-03-28 23:48:55,499 - root - INFO - Start fetching problems by tags
2024-03-28 23:48:55,504 - root - INFO - Load "chinese remainder theorem" locally: total 16
problems
2024-03-28 23:48:57,511 - root - INFO - Load "fft" locally: total 89 problems
2024-03-28 23:48:59,520 - root - INFO - Load "combinatorics" locally: total 631 problems
2024-03-28 23:49:01,531 - root - INFO - Load "two pointers" locally: total 507 problems
2024-03-28 23:49:03,545 - root - INFO - Load "greedy" locally: total 2665 problems
2024-03-28 23:49:05,570 - root - INFO - Load "matrices" locally: total 115 problems
2024-03-28 23:49:07,578 - root - INFO - Load "graph matchings" locally: total 88 problems
2024-03-28 23:49:09,589 - root - INFO - Load "data structures" locally: total 1631 problems
2024-03-28 23:49:11,613 - root - INFO - Load "math" locally: total 2702 problems
2024-03-28 23:49:13,639 - root - INFO - Load "probabilities" locally: total 226 problems
2024-03-28 23:49:15,653 - root - INFO - Load "graphs" locally: total 1020 problems
2024-03-28 23:49:17,671 - root - INFO - Load "binary search" locally: total 992 problems
2024-03-28 23:49:19,687 - root - INFO - Load "strings" locally: total 689 problems
```

```
2024-03-28 23:49:21,700 - root - INFO - Load "brute force" locally: total 1561 problems
2024-03-28 23:49:23,718 - root - INFO - Load "ternary search" locally: total 52 problems
2024-03-28 23:49:25,725 - root - INFO - Load "dsu" locally: total 337 problems
2024-03-28 23:49:27,735 - root - INFO - Load "schedules" locally: total 8 problems
2024-03-28 23:49:29,742 - root - INFO - Load "2-sat" locally: total 30 problems
2024-03-28 23:49:31,750 - root - INFO - Load "bitmasks" locally: total 529 problems
2024-03-28 23:49:50,352 - root - INFO - Load "divide and conquer" remotely: total 269
problems
2024-03-28 23:49:53,466 - root - INFO - Load "string suffix structures" remotely: total 87
problems
2024-03-28 23:49:56,977 - root - INFO - Load "dfs and similar" remotely: total 882 problems
2024-03-28 23:50:00,779 - root - INFO - Load "hashing" remotely: total 193 problems
2024-03-28 23:50:05,140 - root - INFO - Load "constructive algorithms" remotely: total 1642
problems
2024-03-28 23:50:08,359 - root - INFO - Load "shortest paths" remotely: total 258 problems
2024-03-28 23:50:12,573 - root - INFO - Load "implementation" remotely: total 2595 problems
2024-03-28 23:50:16,929 - root - INFO - Load "games" remotely: total 203 problems
2024-03-28 23:50:20,131 - root - INFO - Load "number theory" remotely: total 693 problems
2024-03-28 23:50:23,283 - root - INFO - Load "meet-in-the-middle" remotely: total 47
problems
2024-03-28 23:50:38,893 - root - INFO - Load "dp" remotely: total 1998 problems
2024-03-28 23:50:42,958 - root - INFO - Load "expression parsing" remotely: total 35
problems
2024-03-28 23:50:48,312 - root - INFO - Load "flows" remotely: total 139 problems
2024-03-28 23:50:53,834 - root - INFO - Load "geometry" remotely: total 380 problems
2024-03-28 23:50:57,235 - root - INFO - Load "interactive" remotely: total 212 problems
2024-03-28 23:50:59,747 - root - INFO - Load "*special problem" remotely: total 0 problems
2024-03-28 23:51:07,532 - root - INFO - Load "trees" remotely: total 772 problems
2024-03-28 23:51:10,794 - root - INFO - Load "sortings" remotely: total 1009 problems
2024-03-28 23:51:29,078 - root - INFO - Load "*special" remotely: total 416 problems
2024-03-28 23:51:31,087 - root - INFO - Finished fetching problems by 38 tags with 9187
problems
```

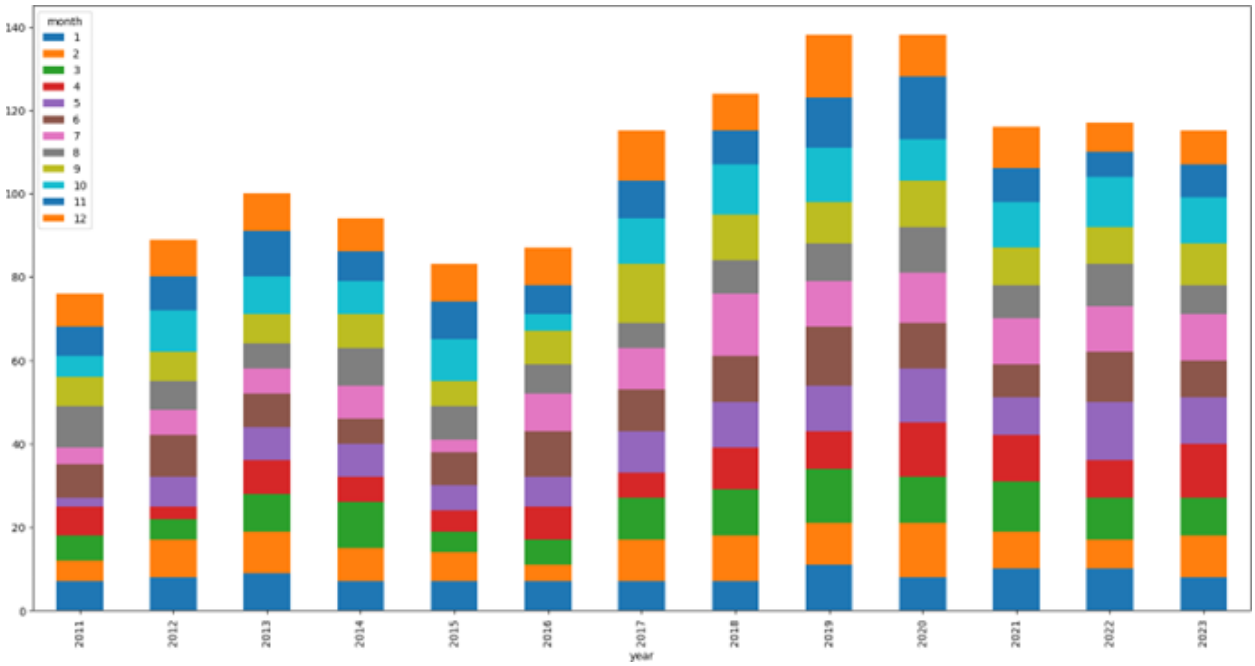
1.3 数据预处理

对数据进行了初步的异常分析，由于是API+爬虫提供，结构化较好，因此不需要太多的预处理。

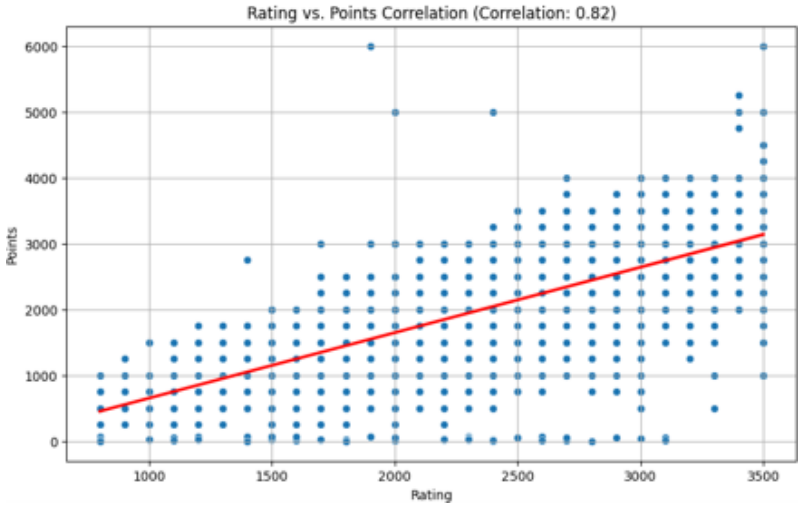
2. 数据分析与可视化

> 数据探索性分析的结果，可以使用统计工具，聚类分析等工具 使用可视化来展示分析结果

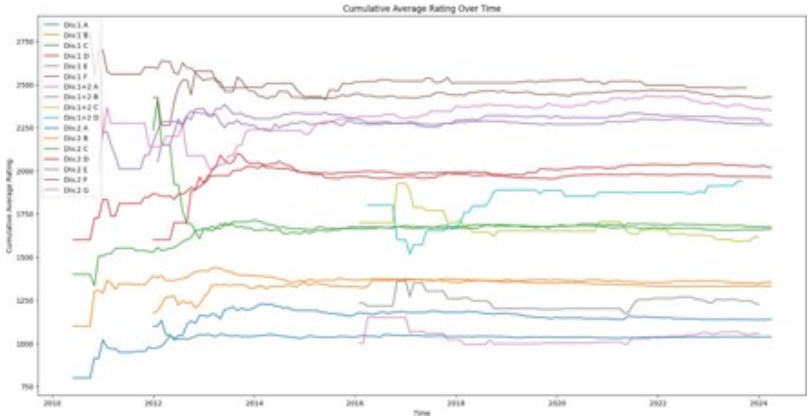
Codeforces官方比赛的年月分布：



比赛前题目难度打分（Points）与赛时难度评分（Rating）的相关性：



题目难度（Div. {1/2/1+2} + {A...G}）的变化：



3. 模型选取

> 围绕选题要解决的问题，考虑使用哪些模型来进行挖掘 说明选择的理由

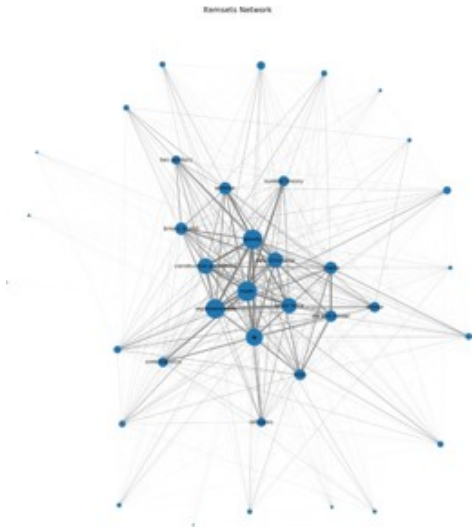
- **关联规则学习**：发现不同编程问题之间的关系，例如哪些类型的题目经常一起出现，或者哪些技能在解决某类问题时特别重要。这对于理解竞赛题目的结构和参赛者的解题模式非常有帮助。

- **关联规则学习**：发现选手表现与学习资源、讨论话题之间的关联，例如通过Apriori算法寻找常见的题目组合或讨论主题。
- **聚类算法（Clustering Algorithms）**：如K-means或层次聚类（Hierarchical clustering），这可以用来发现具有相似特征的参赛者群体或题目类型，从而帮助理解数据中的模式和关系。

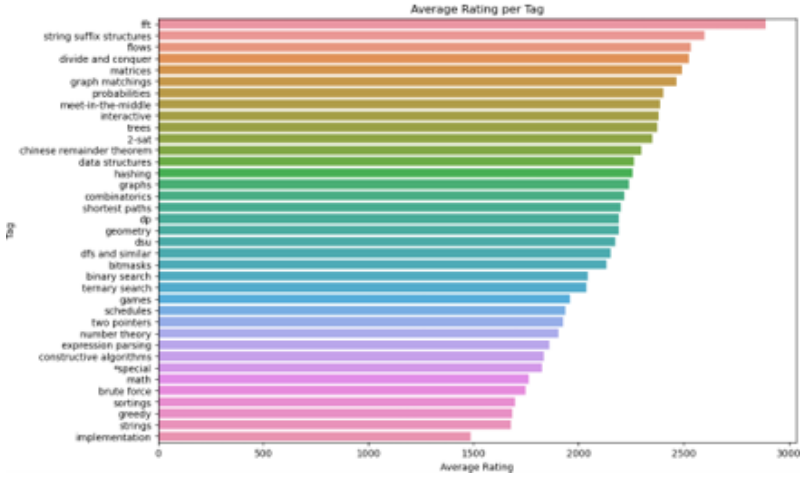
4. 挖掘实验的结果

4.1 频繁项集挖掘

- **关联规则学习**：对编程问题标签进行频繁项集挖掘，研究哪些类型的解题思路经常一起出现，并进行可视化展示，例如，动态规划与组合问题，图论与深度优先搜索问题等常一同出现在题目中。



- **编程问题标签与难度关联**：对编程问题标签和题目难度关系进行研究，并进行可视化展示，例如，快速傅里叶变换相关题目具有较高难度评级。



- **频繁项集支持度**：对编程问题关联进行挖掘并使用多种指标进行评测。

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(brute force)	(implementation)	0.169914	0.282464	0.059867	0.352338	1.247372	0.011873	1.107886	0.238909
1	(implementation)	(brute force)	0.282464	0.169914	0.059867	0.211946	1.247372	0.011873	1.053337	0.276383
2	(brute force)	(math)	0.169914	0.294111	0.054860	0.322870	1.097782	0.004887	1.042471	0.107305
3	(math)	(brute force)	0.294111	0.169914	0.054860	0.186528	1.097782	0.004887	1.020424	0.126584
4	(constructive algorithms)	(greedy)	0.178731	0.290084	0.069664	0.389769	1.343641	0.017817	1.163356	0.311413
5	(greedy)	(constructive algorithms)	0.290084	0.178731	0.069664	0.240150	1.343641	0.017817	1.080831	0.360259
6	(constructive algorithms)	(math)	0.178731	0.294111	0.058234	0.325822	1.107819	0.005668	1.047036	0.118507
7	(math)	(constructive algorithms)	0.294111	0.178731	0.058234	0.198001	1.107819	0.005668	1.024028	0.137877
8	(data structures)	(greedy)	0.177533	0.290084	0.051486	0.290006	0.999732	-0.000014	0.999891	-0.000326
9	(greedy)	(data structures)	0.290084	0.177533	0.051486	0.177486	0.999732	-0.000014	0.999942	-0.000377
10	(graphs)	(dfs and similar)	0.111026	0.096005	0.052248	0.470588	4.901694	0.041589	1.707546	0.895402
11	(dfs and similar)	(graphs)	0.096005	0.111026	0.052248	0.544218	4.901694	0.041589	1.950435	0.880524
12	(greedy)	(dp)	0.290084	0.217481	0.058888	0.203002	0.933423	-0.004200	0.981833	-0.091298
13	(dp)	(greedy)	0.217481	0.290084	0.058888	0.270771	0.933423	-0.004200	0.973516	-0.083535
14	(math)	(dp)	0.294111	0.217481	0.054751	0.186158	0.855975	-0.009212	0.961512	-0.192483
15	(dp)	(math)	0.217481	0.294111	0.054751	0.251752	0.855975	-0.009212	0.943388	-0.176970
16	(implementation)	(greedy)	0.282464	0.290084	0.072603	0.257033	0.886064	-0.009336	0.955515	-0.151972
17	(greedy)	(implementation)	0.290084	0.282464	0.072603	0.250281	0.886064	-0.009336	0.957073	-0.153353
18	(math)	(greedy)	0.294111	0.290084	0.077392	0.263138	0.907112	-0.007925	0.963432	-0.126687
19	(greedy)	(math)	0.290084	0.294111	0.077392	0.266792	0.907112	-0.007925	0.962740	-0.126059
20	(greedy)	(sortings)	0.290084	0.109829	0.058670	0.202251	1.841510	0.026810	1.115854	0.643692
21	(sortings)	(greedy)	0.109829	0.290084	0.058670	0.534192	1.841510	0.026810	1.524054	0.513348
22	(math)	(implementation)	0.294111	0.282464	0.064548	0.219467	0.776973	-0.018528	0.919289	-0.289089
23	(implementation)	(math)	0.282464	0.294111	0.064548	0.228516	0.776973	-0.018528	0.914976	-0.285737
24	(math)	(number theory)	0.294111	0.075433	0.055513	0.188749	2.502219	0.033328	1.139681	0.850499
25	(number theory)	(math)	0.075433	0.294111	0.055513	0.735931	2.502219	0.033328	2.673120	0.649336

4.2 神经网络拟合CF比赛Rating变化的计算机制

- **挖掘思路**：爬取近 200 场比赛，筛选用户的相关数据，包括`rank, old rating, new rating`，通过深度学习拟合`delta rating`与其他信息的关系。
- **神经网络结构**：通过分析数据体量和拟合任务目标，设计了一个相对小型的网络。

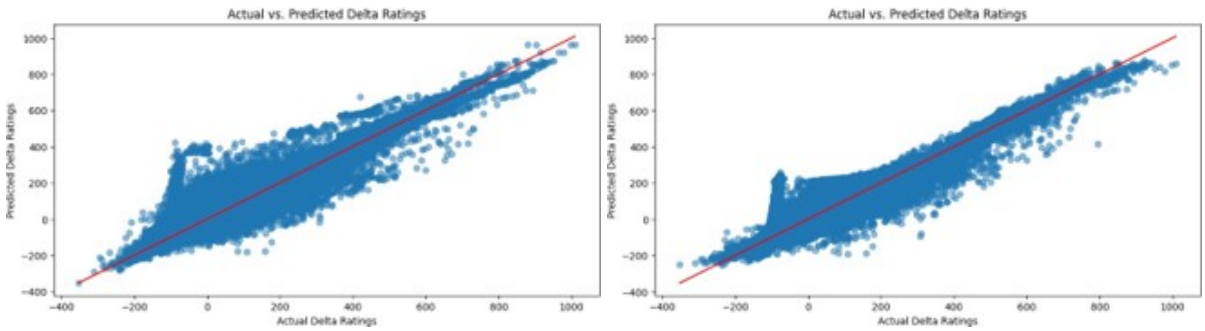
```
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 64)                256

dense_1 (Dense)              (None, 64)                4160

dense_2 (Dense)              (None, 1)                 65

=================================================================
Total params: 4,481
Trainable params: 4,481
Non-trainable params: 0
_________________________________________________________________
```

- **拟合效果**：验证猜想，CF比赛Rating变化与用户的比赛排名和原 rating 的绝对值相关；与用户的新/旧排名的百分比相关。



- **其他结论**：通果获取高水平选手的用户`ID`，分析高频标签，探究高水平选手的做题倾向相对集中，标签集中在博弈论与后缀结构等。

5. 系统交互设计

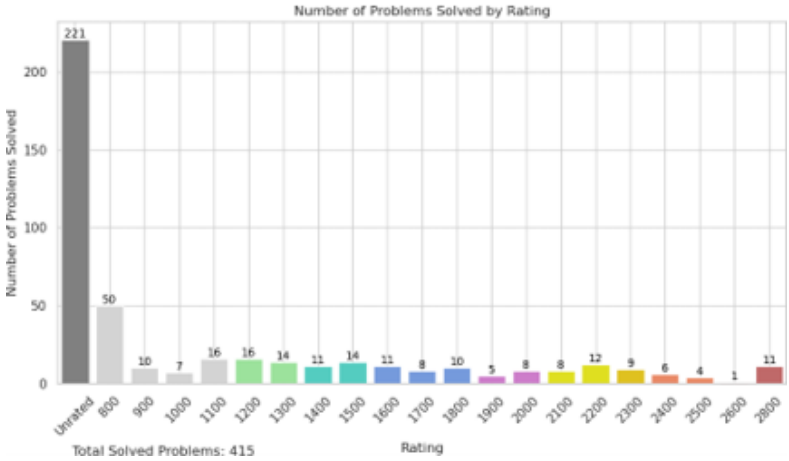
> 主要描述提供的功能及使用方法，输入、输出等。

1. **功能概览**

- **数据浏览**：允许用户浏览和搜索Codeforces的历史比赛和题目数据。
- **数据分析**：提供各种预设的数据分析选项，如趋势分析、参赛者表现评估等。
- **报告生成**：用户可以生成和下载分析报告，包括图表和统计摘要。
- **实时数据追踪**：跟踪实时比赛数据和用户表现。

2. 用户交互设计

- **主页**：展示最新比赛、热门题目和最近活动的概览。
- **数据查询界面**：提供一个表单或搜索栏，用户可以输入特定条件（如日期范围、题目类型等）进行数据查询。
- **分析工具页**：用户可以选择不同的数据分析模型，并设置相应参数来分析数据。
- **结果展示界面**：以表格、图表或文本形式展示分析结果，例如趋势线图、柱状图等。



3. 输入输出设计

- 输入：
  - **查询条件**：用户通过表单输入搜索条件（如比赛编号、日期、题目标签等）。
  - **分析参数**：在分析工具页，用户可以设定具体的分析参数，如时间窗口、聚类数等。
- 输出：
  - **数据视图**：直观显示查询结果，如比赛列表、题目详情等。
  - **分析报告**：输出包括图表和关键指标的详细分析报告，用户可以下载PDF或Excel格式的报告。

4. 交互元素设计

- **导航栏**：方便用户切换不同的界面和功能模块。
- **数据筛选工具**：提供动态筛选工具，让用户根据多种标准（如难度、标签、日期等）筛选数据。
- **交互式图表**：图表支持用户交互，如点击图表元素查看详细数据点。

5. 任务分配与完成情况

- 廖嘉琦： 数据获取、算法实现、文档撰写
- 张亦晴：算法实现、可视化、文档撰写
- 曹健：系统设计、可视化、文档撰写