SC4000/CZ4041/CE4041: Machine Learning

Course Project Description

Kelly KE
College of Computing and Data Science
NTU, Singapore

Detailed Project Description

- This is a group-based course project
- Each group consists of at most 5 members (≤ 5)
- Individual "group" is allowed, but not recommended
- Each group can choose one of the Kaggle competitions listed on the next slide as the course project
- Assessments:
  - Project report (30%)
  - Presentation video (5%)

Course Project Candidates – Kaggle Competitions

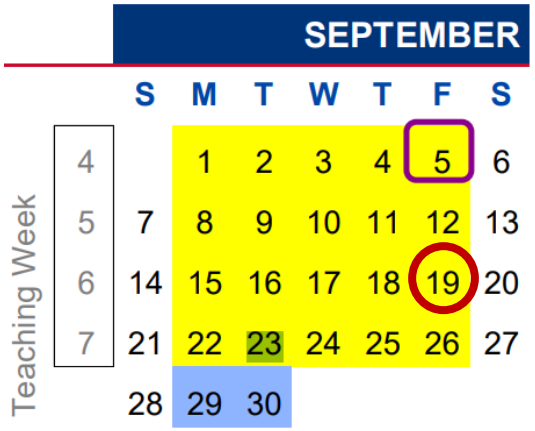
- American Express - Default Prediction
  https://www.kaggle.com/competitions/amex-default-prediction/
- Yale/UNC-CH - Geophysical Waveform Inversion
  https://www.kaggle.com/competitions/waveform-inversion/
- Google Smartphone Decimeter Challenge 2022
  https://www.kaggle.com/competitions/smartphone-decimeter-2022/
- Image Matching Challenge 2025
  https://www.kaggle.com/competitions/image-matching-challenge-2025/
- Linking Writing Processes to Writing Quality
  https://www.kaggle.com/competitions/linking-writing-processes-to-writing-quality/
- HuBMAP – Hacking the Kidney
  https://www.kaggle.com/competitions/hubmap-kidney-segmentation/
- Google - Fast or Slow? Predict AI Model Runtime
  https://www.kaggle.com/competitions/predict-ai-model-runtime/
- LMSYS - Chatbot Arena Human Preference Predictions
  https://www.kaggle.com/competitions/lmsys-chatbot-arena/
- UM – Game-Playing Strength of MCTS Variants
  https://www.kaggle.com/competitions/um-game-playing-strength-of-mcts-variants/
- BirdCLEF+ 2025
  https://www.kaggle.com/competitions/birdclef-2025/

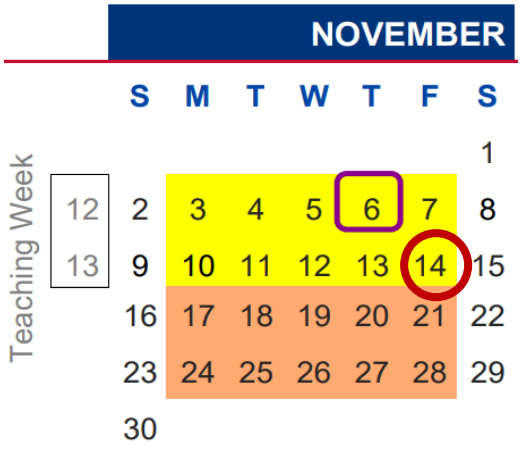
Programming Languages

- Programming Languages:
  - Any programming language can be used, e.g., Python (recommended), C/C++, Java, R, etc.
  - Any open-source ML toolbox can be used
- Note: directly using the source codes released by Kaggle participants are NOT allowed (treated as plagiarism if found)
- Acknowledge in report others' source codes used as references for your own code development
- Do not release your codes publicly!

Key Dates

- Send information on group members via email:
  - by 19th Sept. 2025 (Friday of Week 6)
  - Phase I: find group members by yourself (a forum on "Discussion Board" of NTULearn course site has been created for help)
  - Phase II: will help those who are not able to form a group
- Submit required files via NTULearn:
  - by 11:59pm, 14th Nov. 2025 (Friday of Week 13)





Task 1: Send Group Information

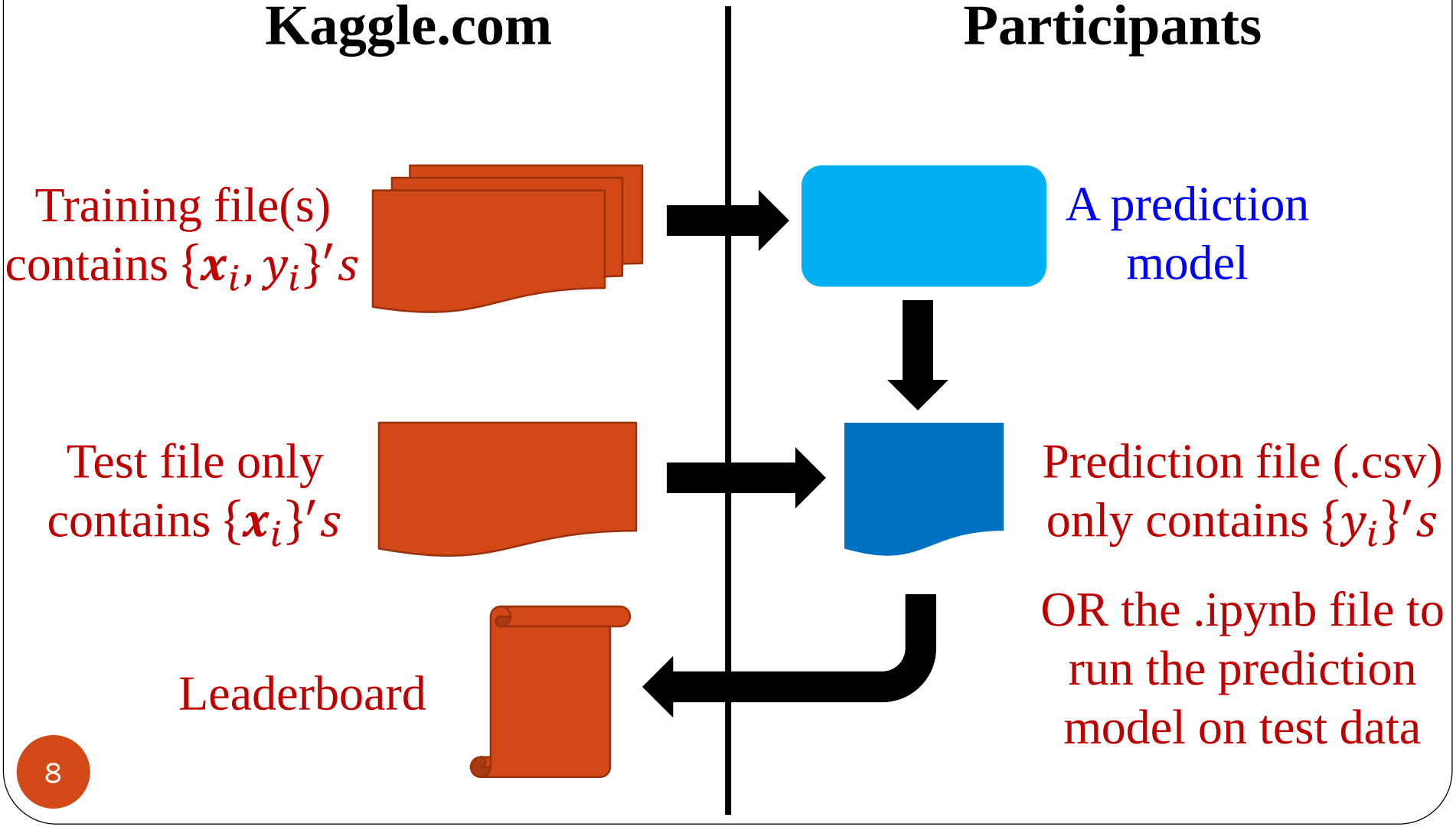
- Send via email to ypke@ntu.edu.sg with all group members on the cc list
- Information needed: send in the following information for all members in a table form:
  - Full name (the name on matric card)
  - Email account (**without** @e.ntu.edu.sg)
  - Example:

Name	**Email**
Ng Wei Min	wmng01
Gupta Agrawal	agrawalg
Joel Chew Xin	chewx
Mary Lam	mlam002

Task 2: Project Submission

- Required files to be submitted:
  1. A project report (.pdf or .doc)
  2. A link of presentation video (.txt)
  3. The final .csv file of your prediction results or .ipynb file (with saved models) submitted to the specific Kaggle competition
- Notes:
  - Only the report and video will be assessed
  - The submitted .csv or .ipynb is to double check whether the reported results are correct
  - You may be randomly asked to provide source codes to check whether they have plagiarism issues

General Information of Kaggle



Format and Content of Video

- Presentation video:
  - To summarize your course project in a video of ≤ 10 minutes long
  - You can use any tool to produce the video, e.g., Zoom/MS Teams
  - Upload the video to YouTube (you could set it **private**)
  - Put the link in a .txt file and submit it to NTULearn
  - We will select top-5 video presentations and share their links to the class in our course site
  - Indicate in the .txt file if your group would like to share. If there is no indication, **SHARE** is the default.

Content of Project Report

- Specific roles and contributions of each group member
  - "Lazy" members will be graded differently
  - Report "lazy" members to us as early as possible (preferably before project end date)
  - Keep version control / chat history / email communications (in case)
- Evaluation score, ranking position, and relative ranking of your results for the specific competition in Kaggle
  - Provide screenshots of your evaluation score and ranking position
  - Calculate and report your relative ranking
- Competition description (using your own words)
- Challenges of the problem
- Your proposed solution in detail (preprocessing, feature engineering/representation learning, methodologies, etc.)
- Experimental study to demonstrate why your solution is effective
- Conclusion: what you have learned from the project

Format and Assessment on Project Report

- Report format:
  - 12 point font, single space, ≤ 20 pages
  - Page limit applies to all contents (including title page, references, etc.)

- Leaderboard performance (public leaderboard)
- Convincingness
- Solution novelty
- Writing

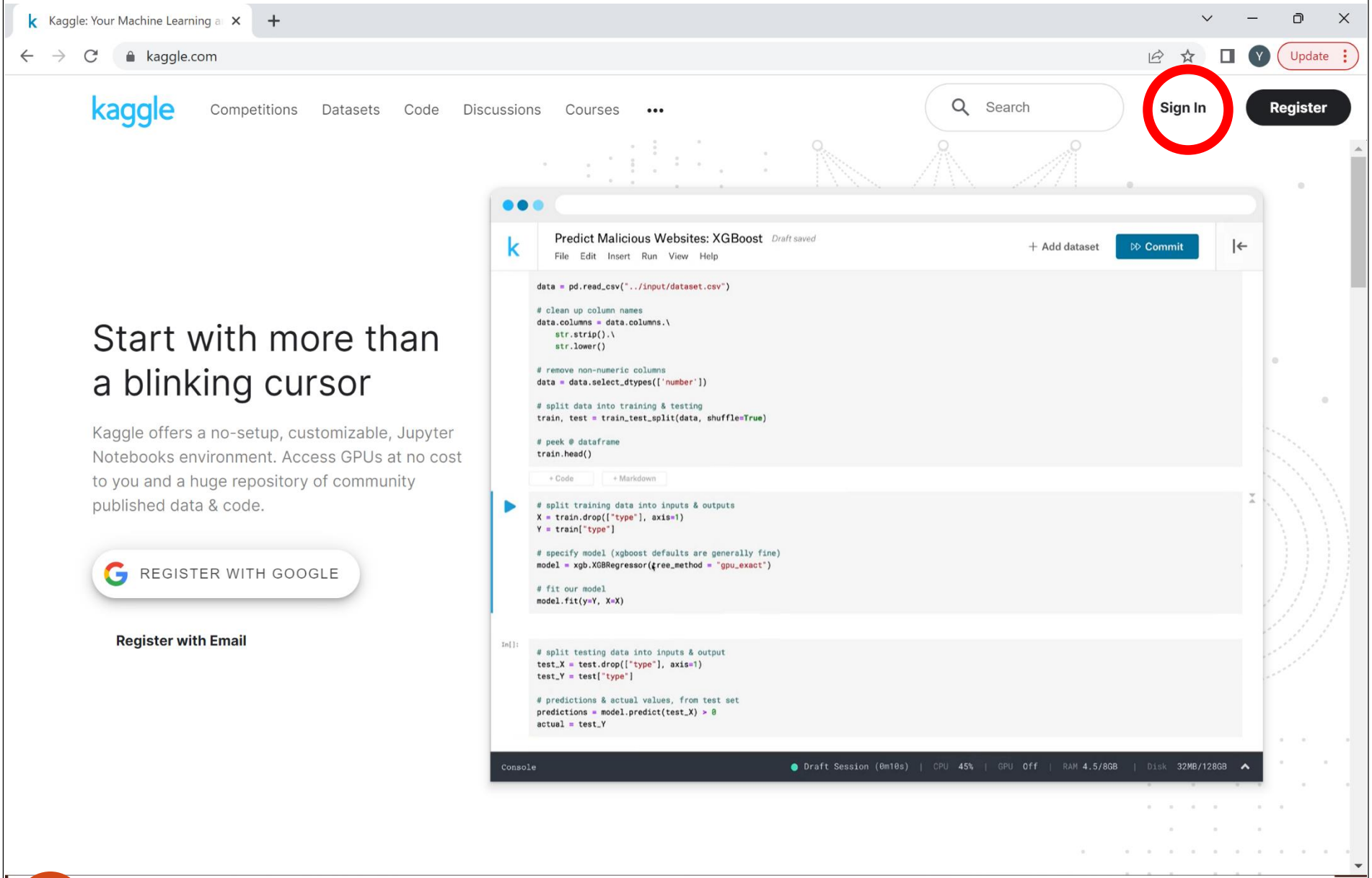
Whether the report is well organized;
Whether the descriptions are logically clear;
Whether the descriptions are detailed enough;
Whether the report contains a lot of typos.

Assessments - Report

- **Leaderboard Performance:** though all the listed Kaggle competitions are completed, you can still submit your results to Kaggle to obtain an evaluation score and find a corresponding ranking position in the **public** leaderboard
- The performance assessment is based on the relative ranking of your results on the specific competition (5 bands: top 10%, top 30%, top 50%, top 70%, and the rest)

Assessments - Report

- Why the **public** leaderboard but not private leaderboard for performance assessment?
- In an on-going Kaggle competition:
  - The public leaderboard is revealed to the participants
  - The private leaderboard is only released when the competition is over and it is used for the final standing
- In our course, the selected competitions are completed:
  - You can tune your model based on both public and private leaderboards
  - The models of prior participants were not tuned based on the private leaderboard but on the public one
  - For fair comparisons, we use public leaderboard, on which both prior participants and you can tune the models


Kaggle: Your Machine Learning a
kaggle.com
Update
kaggle
Competitions
Datasets
Code
Discussions
Courses
Search
Sign In
Register
Start with more than a blinking cursor
Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access GPUs at no cost to you and a huge repository of community published data & code.
REGISTER WITH GOOGLE
Register with Email
Predict Malicious Websites: XGBoost
Draft saved
File Edit Insert Run View Help
+ Add dataset
Commit
data = pd.read_csv("../input/dataset.csv")
# clean up column names
data.columns = data.columns.\
str.strip().\
str.lower()
# remove non-numeric columns
data = data.select_dtypes(['number'])
# split data into training & testing
train, test = train_test_split(data, shuffle=True)
# peek @ dataframe
train.head()
+ Code
+ Markdown
# split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]
# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor(tree_method = "gpu_exact")
# fit our model
model.fit(y=Y, X=X)
In[]:
# split testing data into inputs & output
test_X = test.drop(["type"], axis=1)
test_Y = test["type"]
# predictions & actual values, from test set
predictions = model.predict(test_X) > 0
actual = test_Y
Console
Draft Session (0m10s) | CPU 45% | GPU Off | RAM 4.5/8GB | Disk 32MB/128GB

Kaggle Competitions

kaggle.com/competitions

Update

kaggle

Search

Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Your Work

View Active Events

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the documentation or learn about Community Competitions.

Host a Competition | Your Work

Search competitions

Filters

All Competitions — Everything, past & present

Featured — Premier challenges with prizes

Getting Started — Approachable ML fundamentals

Research — Scientific and scholarly challenges

Community — Created by fellow Kagglers

Get Started

See all

New to Kaggle?

These competitions are perfect for newcomers.

Titanic - Machine Learning from Disaster

Start here! Predict survival on th...

Getting Started

House Prices - Advanced Regression...

Predict sales prices and practice...

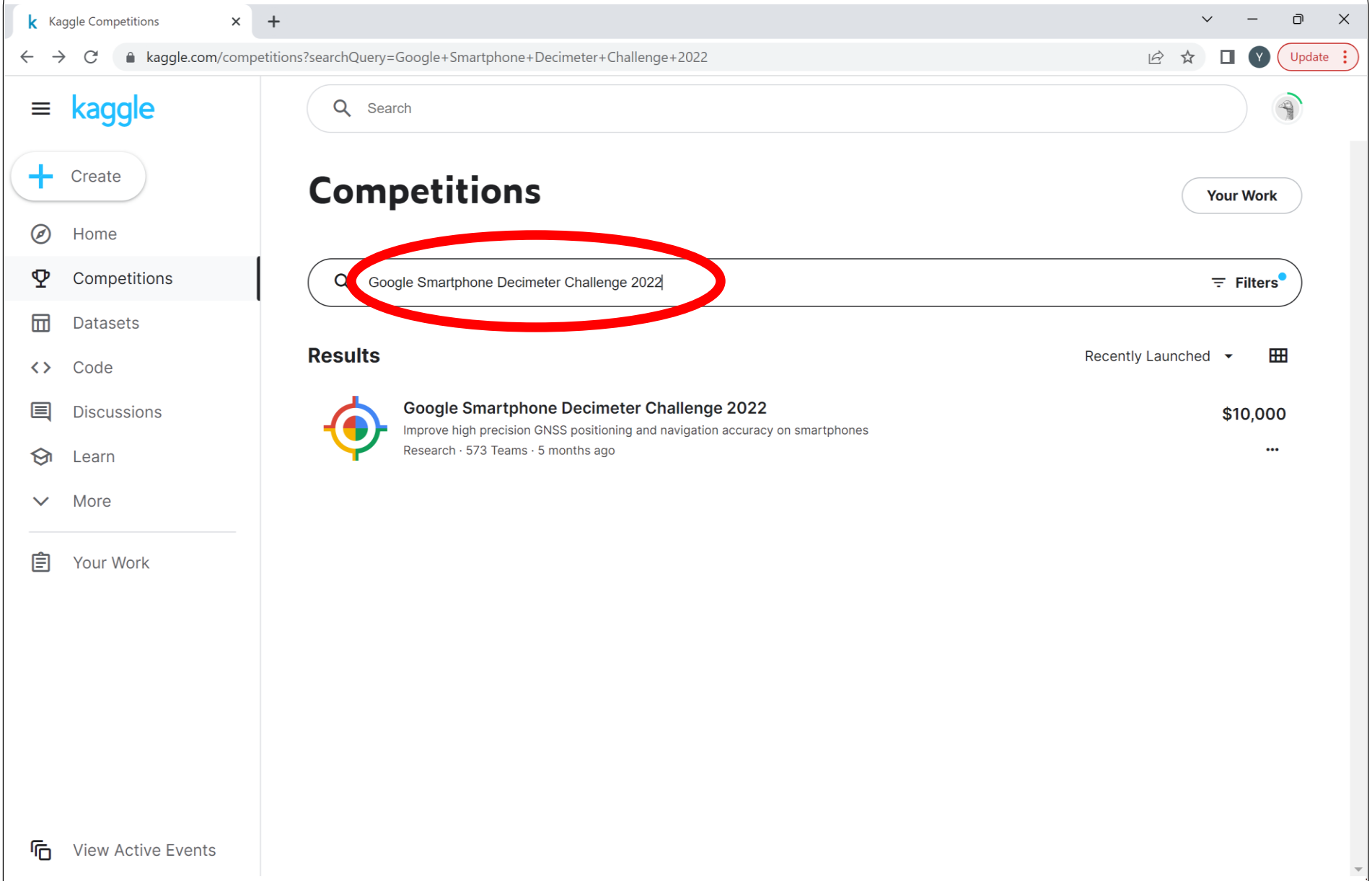
Getting Started

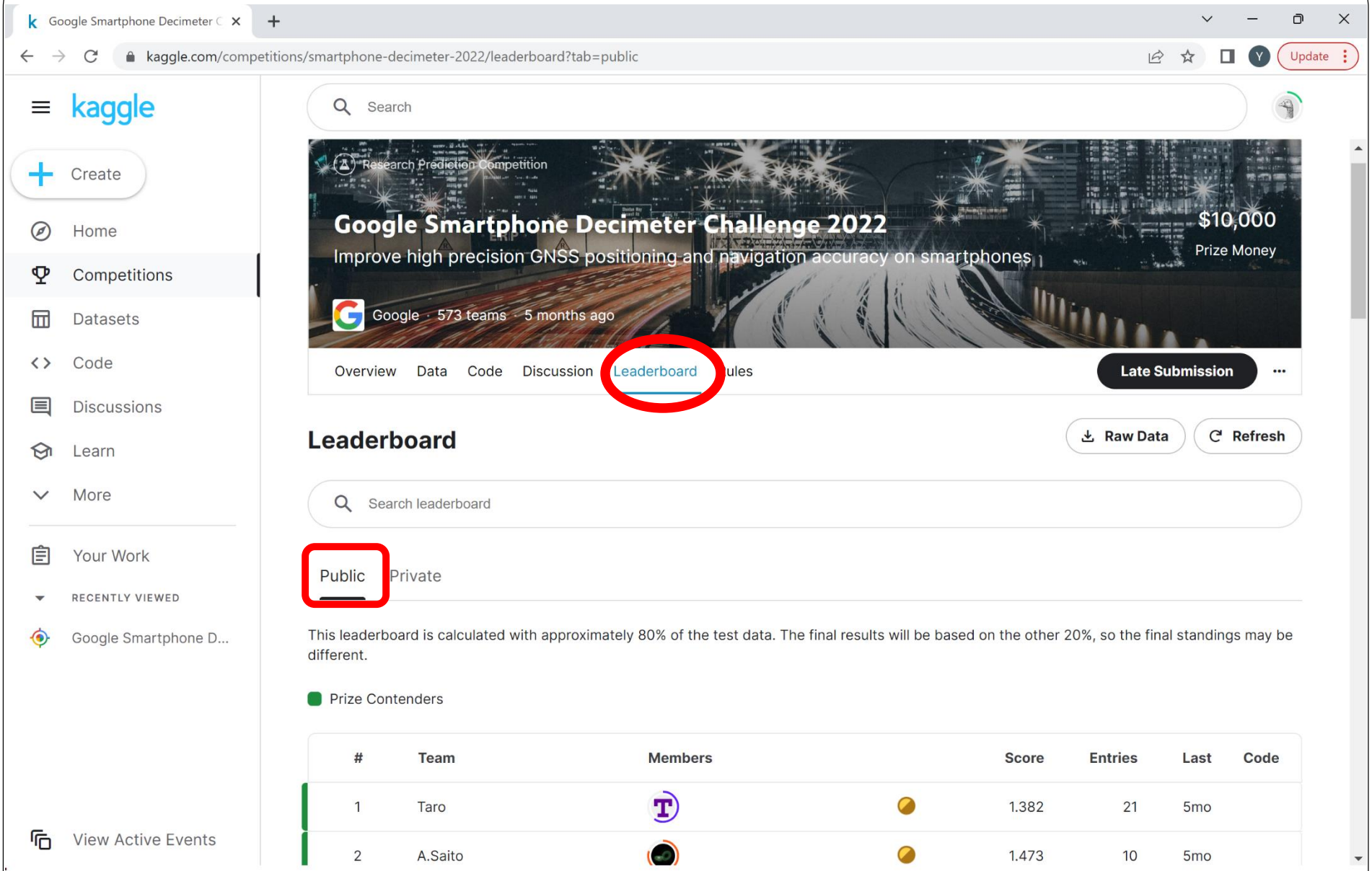
Spaceship Titanic

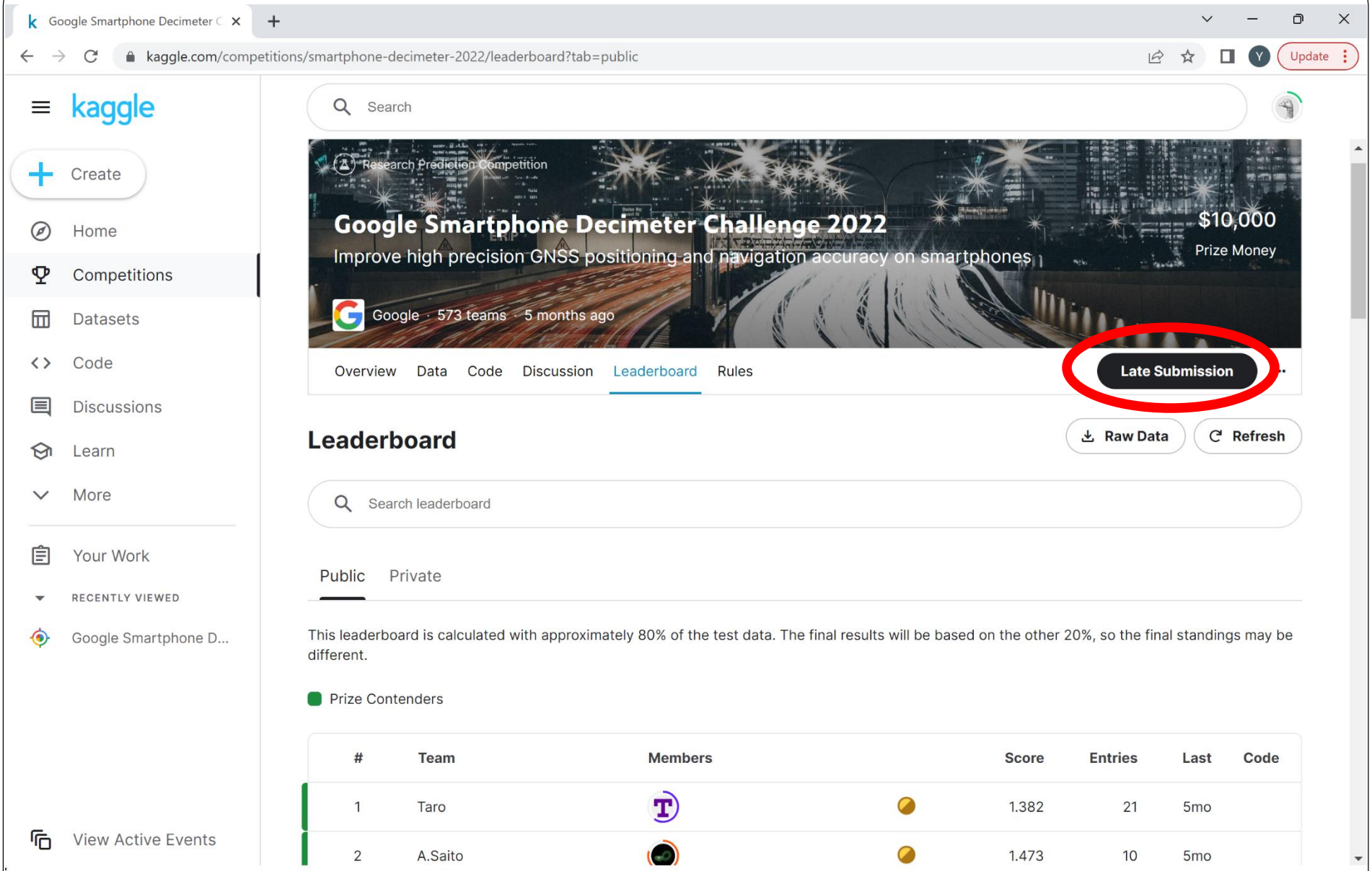
Predict which passengers are tra...

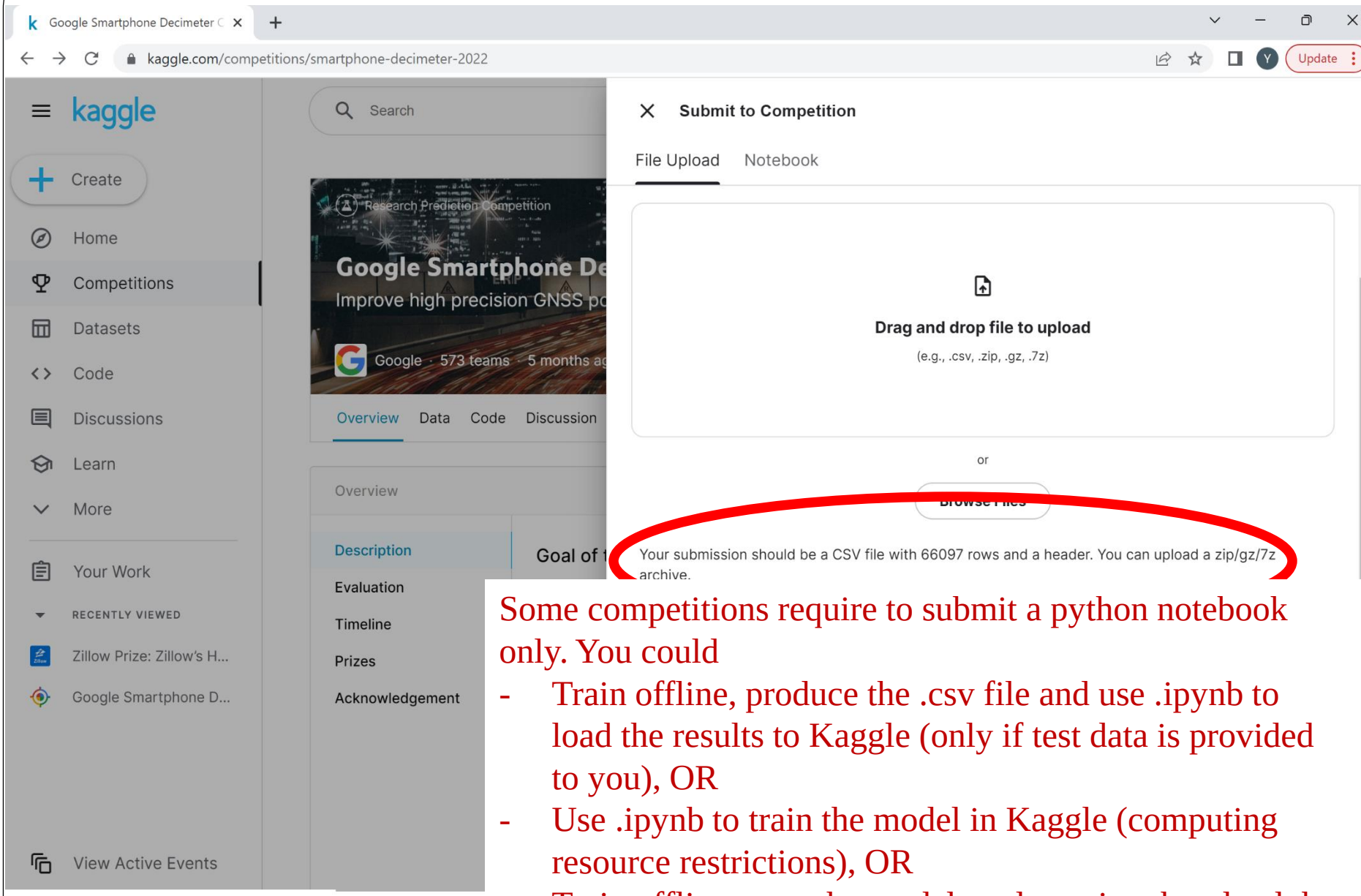
Getting Started

2746 Teams


Kaggle Competitions
kaggle.com/competitions?searchQuery=Google+Smartphone+Decimeter+Challenge+2022
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
View Active Events
Competitions
Your Work
Google Smartphone Decimeter Challenge 2022
Filters
Results
Recently Launched
Google Smartphone Decimeter Challenge 2022
Improve high precision GNSS positioning and navigation accuracy on smartphones
Research · 573 Teams · 5 months ago
$10,000

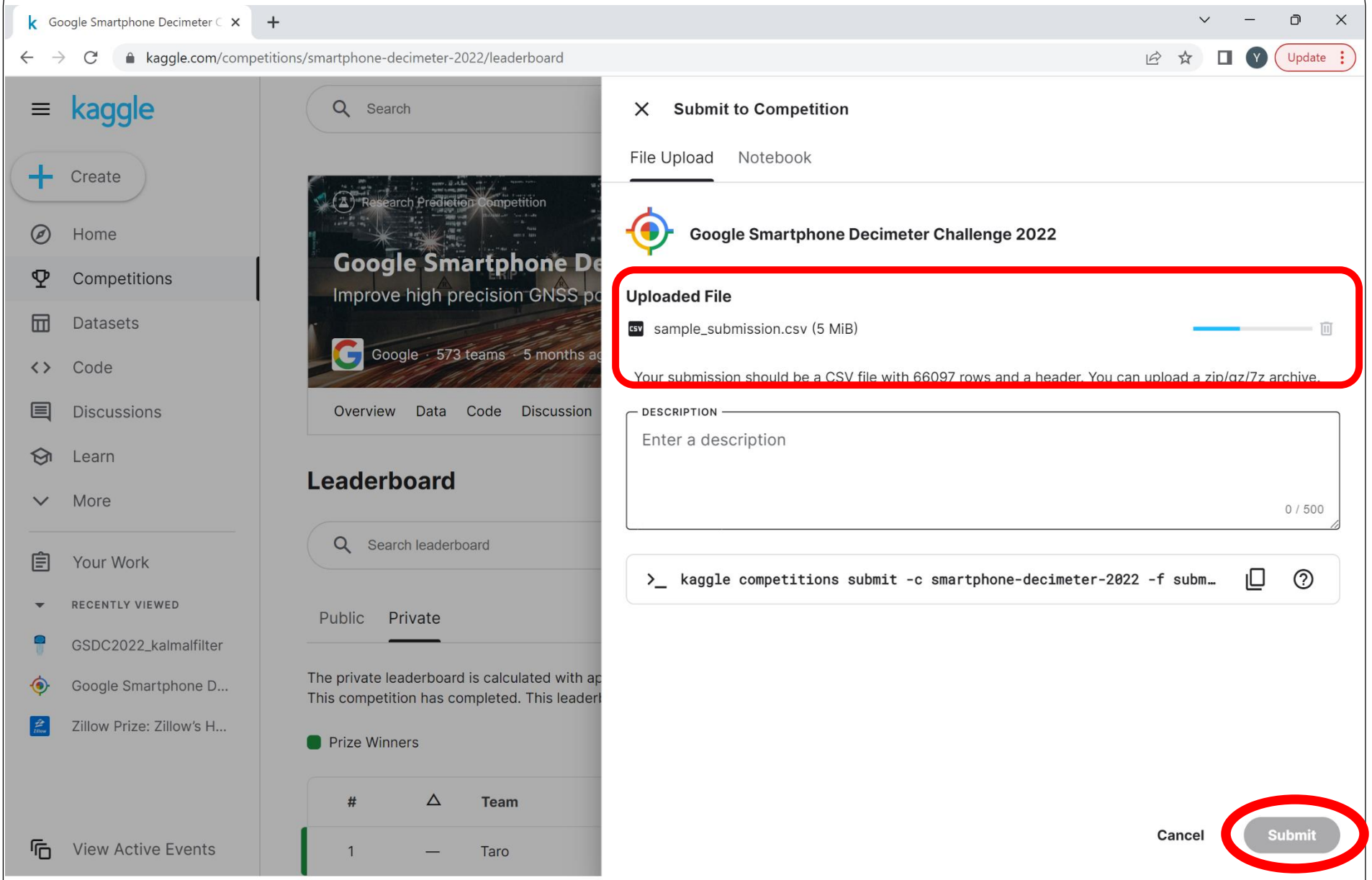

Google Smartphone Decimeter C
kaggle.com/competitions/smartphone-decimeter-2022/leaderboard?tab=public
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
Google Smartphone D...
View Active Events
Research Prediction Competition
Google Smartphone Decimeter Challenge 2022
Improve high precision GNSS positioning and navigation accuracy on smartphones
$10,000
Prize Money
Google · 573 teams · 5 months ago
Overview
Data
Code
Discussion
Leaderboard
ules
Late Submission
Leaderboard
Raw Data
Refresh
Search leaderboard
Public
Private
This leaderboard is calculated with approximately 80% of the test data. The final results will be based on the other 20%, so the final standings may be different.
Prize Contenders
Team Members Score Entries Last Code
1 Taro 1.382 21 5mo
2 A.Saito 1.473 10 5mo

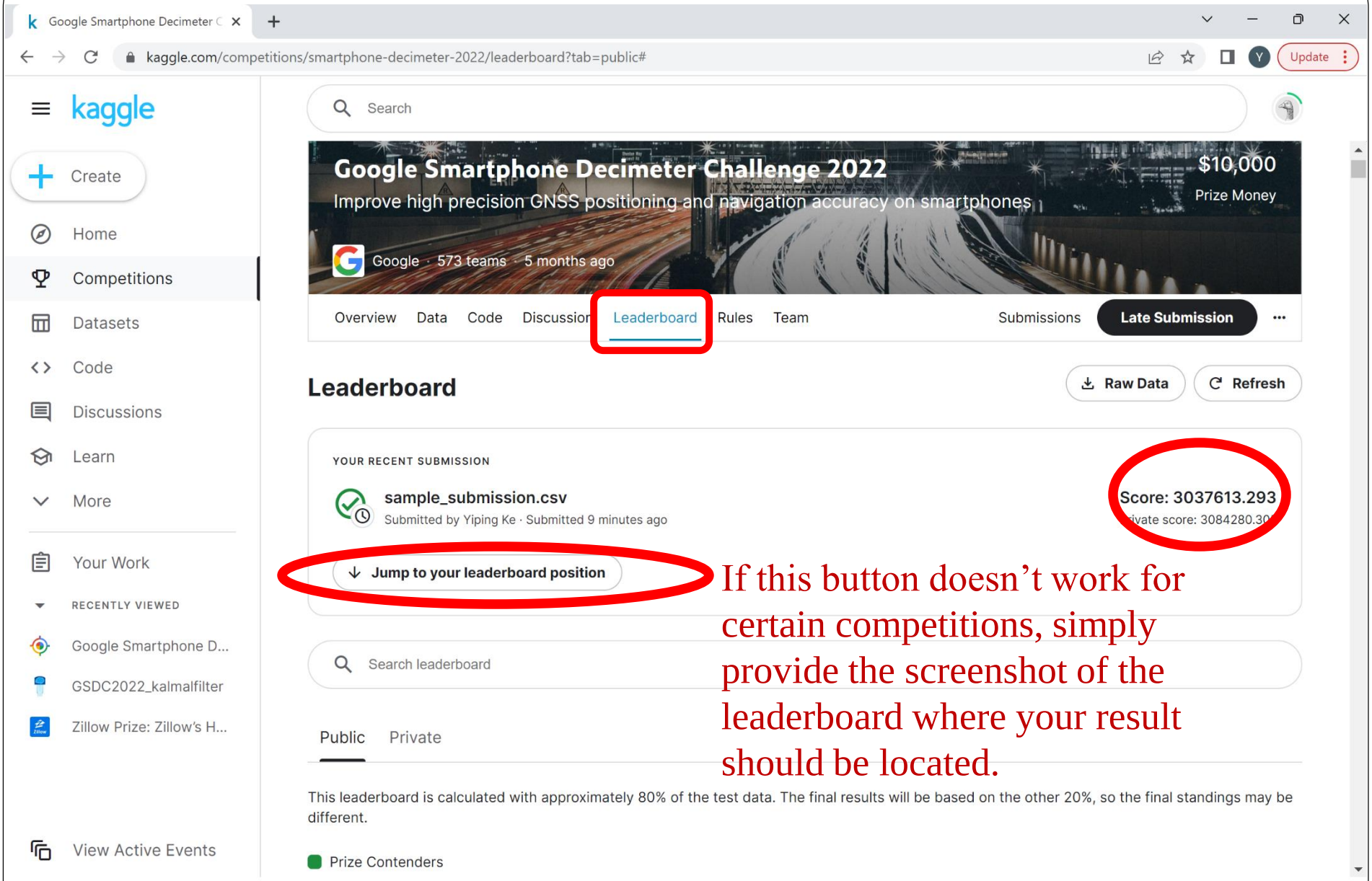

Google Smartphone Decimeter
kaggle.com/competitions/smartphone-decimeter-2022/leaderboard?tab=public
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
Google Smartphone D...
View Active Events
Research Prediction Competition
Google Smartphone Decimeter Challenge 2022
Improve high precision GNSS positioning and navigation accuracy on smartphones
$10,000
Prize Money
Google · 573 teams · 5 months ago
Overview
Data
Code
Discussion
Leaderboard
Rules
Late Submission
Leaderboard
Raw Data
Refresh
Search leaderboard
Public
Private
This leaderboard is calculated with approximately 80% of the test data. The final results will be based on the other 20%, so the final standings may be different.
Prize Contenders
Team Members Score Entries Last Code
1 Taro 1.382 21 5mo
2 A.Saito 1.473 10 5mo



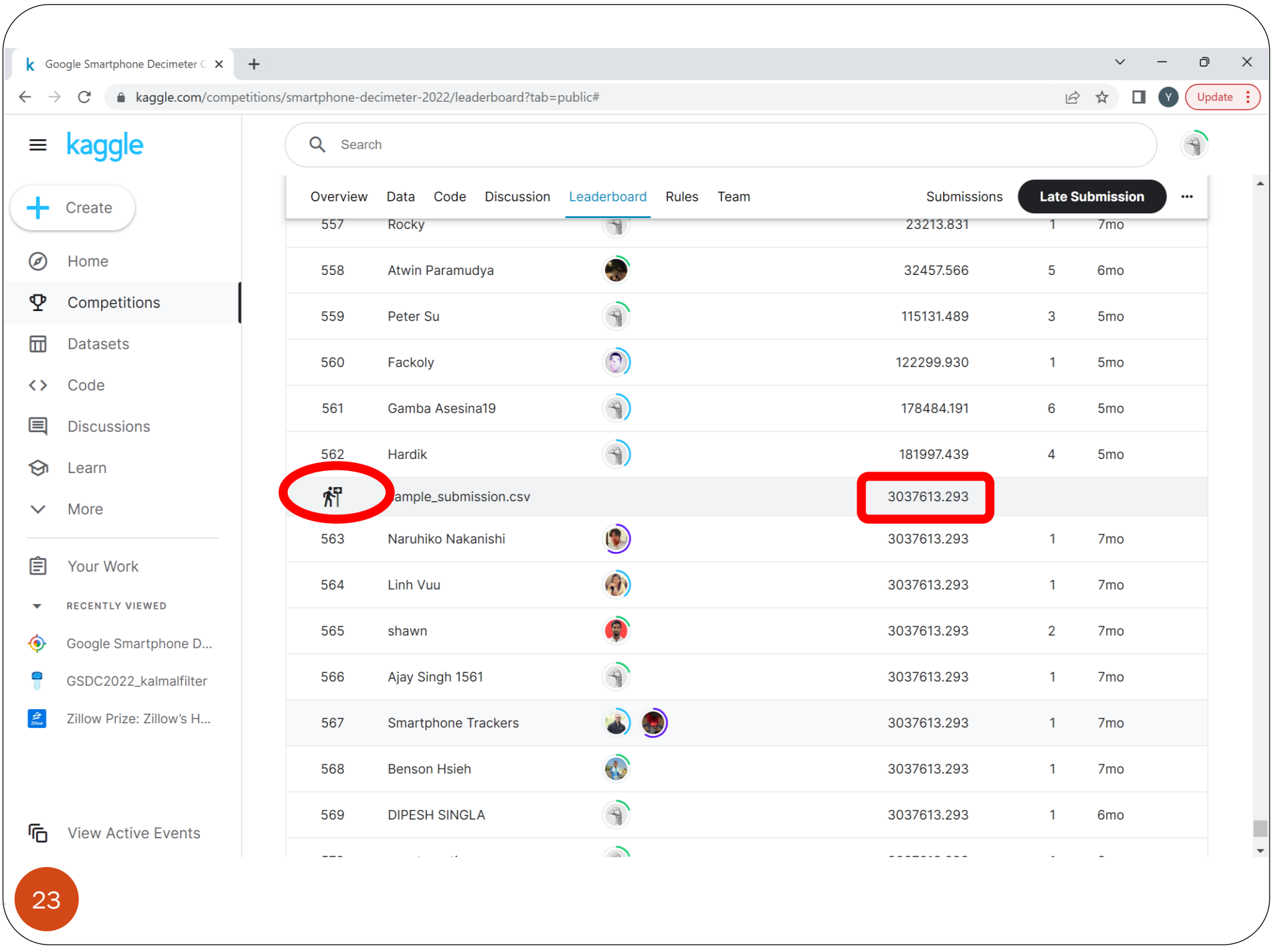
Some competitions require to submit a python notebook only. You could

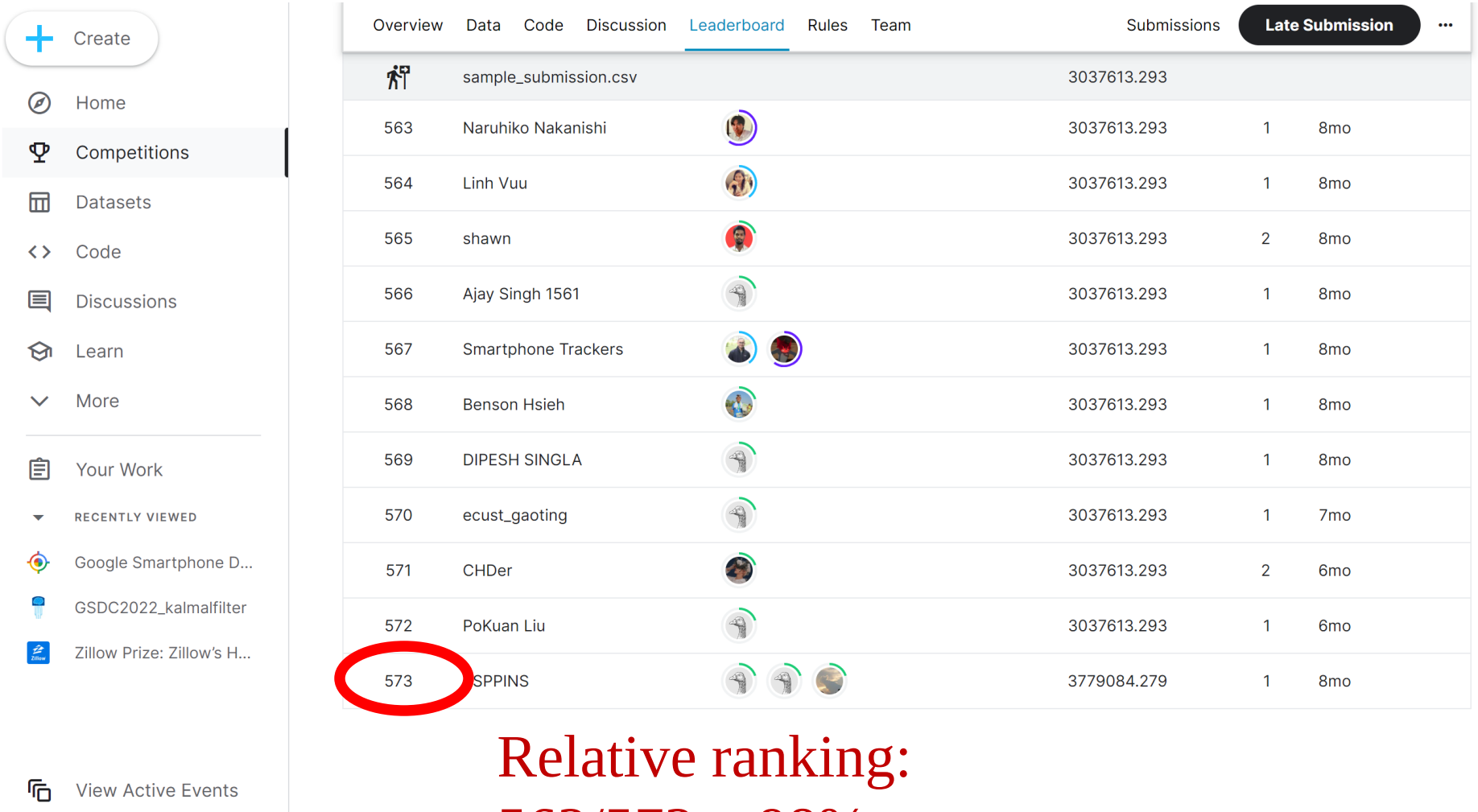
- Train offline, produce the .csv file and use .ipynb to load the results to Kaggle (only if test data is provided to you), OR
- Use .ipynb to train the model in Kaggle (computing resource restrictions), OR
- Train offline, save the model, and use .ipynb to load the model for test


Google Smartphone Decimeter C
kaggle.com/competitions/smartphone-decimeter-2022/leaderboard
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
GSDC2022_kalmalfilter
Google Smartphone D...
Zillow Prize: Zillow's H...
View Active Events
Research Prediction Competition
Google Smartphone De
Improve high precision GNSS po
Google · 573 teams · 5 months ag
Overview
Data
Code
Discussion
Leaderboard
Search leaderboard
Public
Private
The private leaderboard is calculated with ap
This competition has completed. This leaderb
Prize Winners
#
Team
1
Taro
Submit to Competition
File Upload
Notebook
Google Smartphone Decimeter Challenge 2022
Uploaded File
sample_submission.csv (5 MiB)
Your submission should be a CSV file with 66097 rows and a header. You can upload a zip/gz/7z archive.
DESCRIPTION
Enter a description
0 / 500
>_ kaggle competitions submit -c smartphone-decimeter-2022 -f subm…
Cancel
Submit



If this button doesn't work for certain competitions, simply provide the screenshot of the leaderboard where your result should be located.

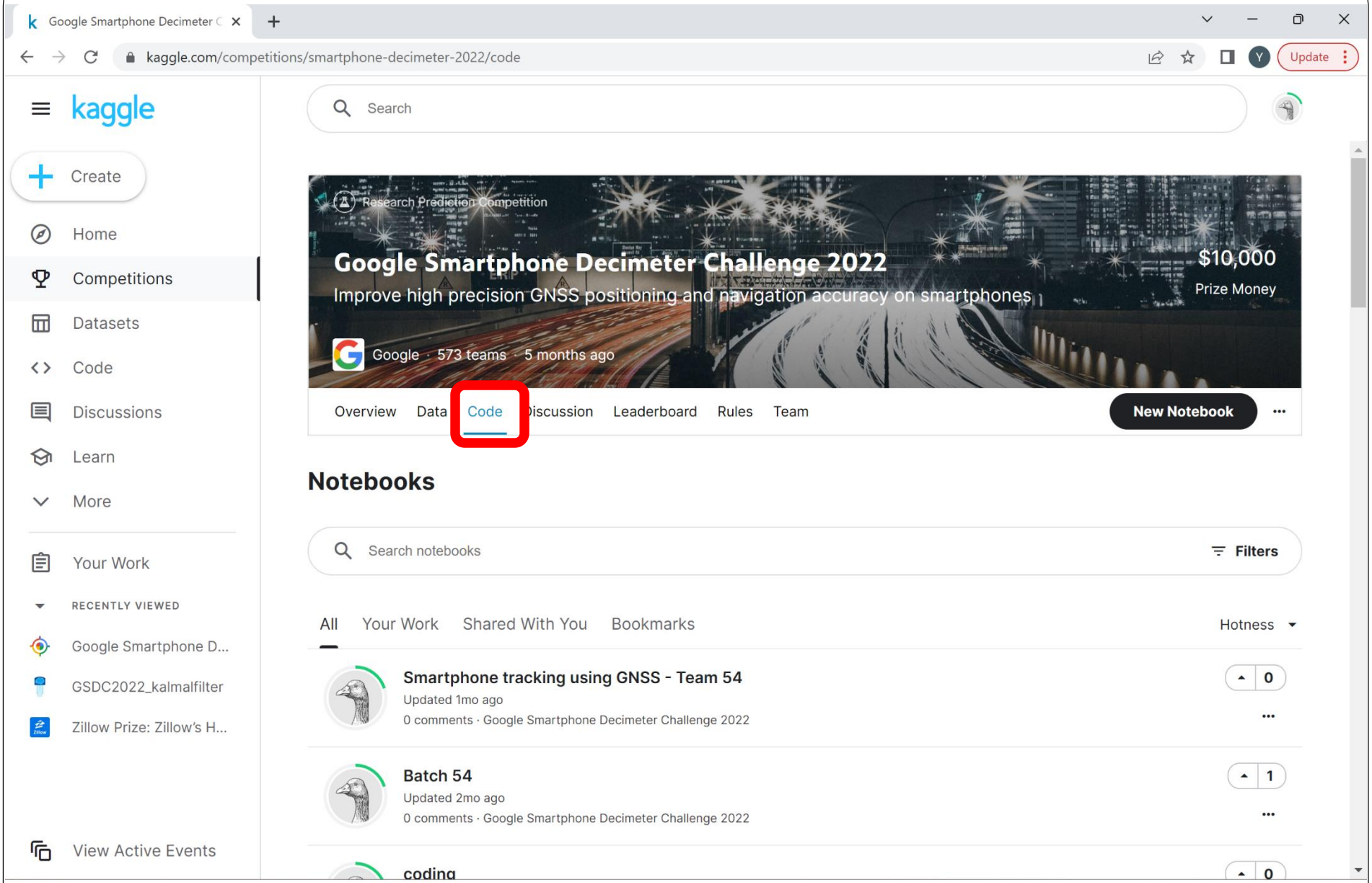

Google Smartphone Decimeter C
kaggle.com/competitions/smartphone-decimeter-2022/leaderboard?tab=public#
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
Google Smartphone D...
GSDC2022_kalmalfilter
Zillow Prize: Zillow's H...
View Active Events
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
Submissions
Late Submission
557 Rocky 23213.831 1 7mo
558 Atwin Paramudya 32457.566 5 6mo
559 Peter Su 115131.489 3 5mo
560 Fackoly 122299.930 1 5mo
561 Gamba Asesina19 178484.191 6 5mo
562 Hardik 181997.439 4 5mo
ample_submission.csv 3037613.293
563 Naruhiko Nakanishi 3037613.293 1 7mo
564 Linh Vuu 3037613.293 1 7mo
565 shawn 3037613.293 2 7mo
566 Ajay Singh 1561 3037613.293 1 7mo
567 Smartphone Trackers 3037613.293 1 7mo
568 Benson Hsieh 3037613.293 1 7mo
569 DIPESH SINGLA 3037613.293 1 6mo

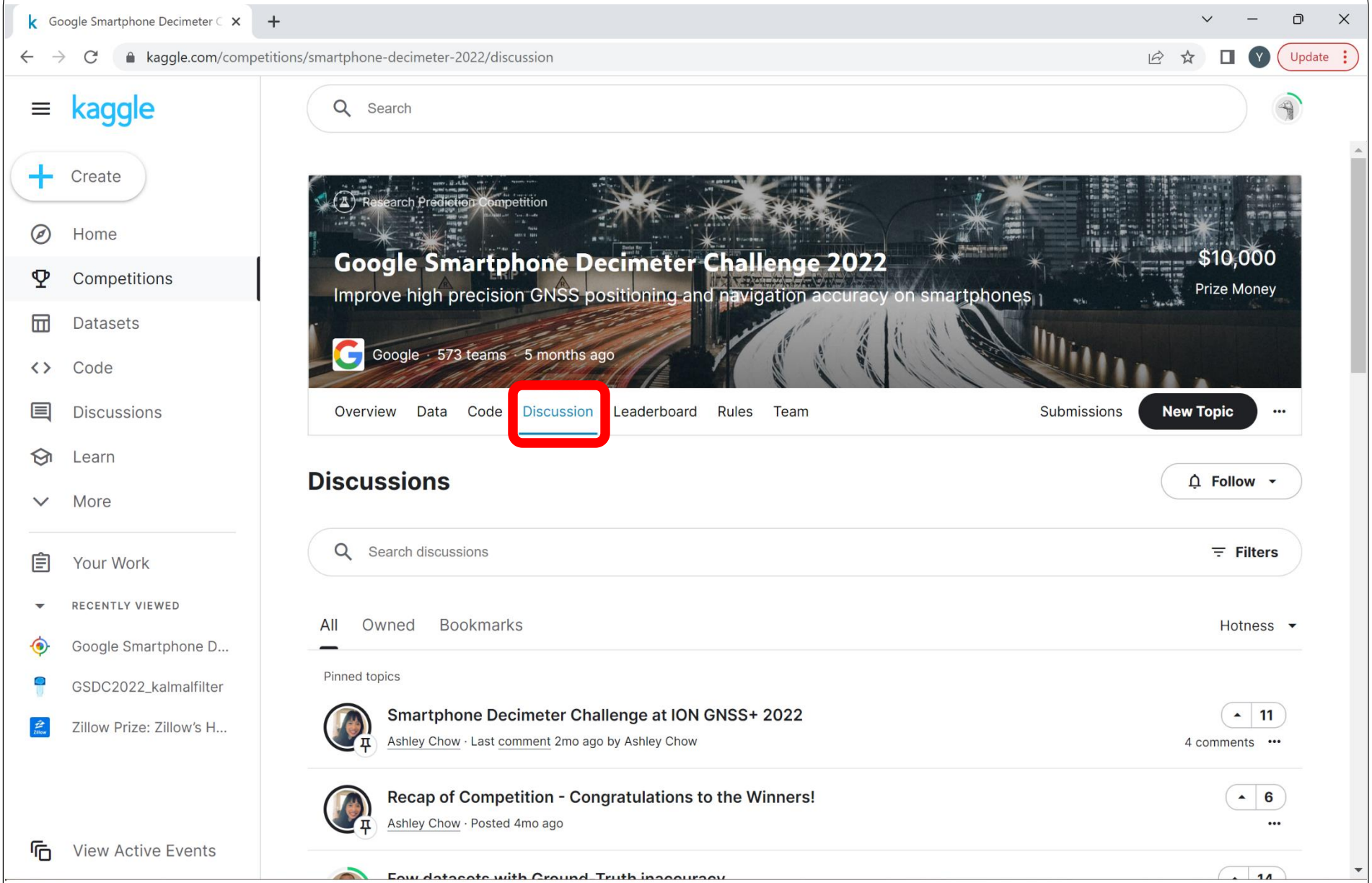


Relative ranking:
563/573 = 98%

Assessments – Report (cont.)

- **Solution Novelty:** as on Kaggle.com, most participants or winners may discuss or even release their solutions (with codes) on the forum of each specific competition
  - If you propose a new and effective solution, you can get credits. You are encouraged to propose your own solutions based on your own understandings on the competitions. In the report and presentation video, highlight your new ideas.
  - Directly reuse released source codes are not allowed!
  - Acknowledge any reference of existing codes!


Google Smartphone Decimeter C
kaggle.com/competitions/smartphone-decimeter-2022/code
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
Google Smartphone D...
GSDC2022_kalmalfilter
Zillow Prize: Zillow's H...
View Active Events
Research Prediction Competition
Google Smartphone Decimeter Challenge 2022
Improve high precision GNSS positioning and navigation accuracy on smartphones
$10,000
Prize Money
Google · 573 teams · 5 months ago
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
New Notebook
Notebooks
Search notebooks
Filters
All
Your Work
Shared With You
Bookmarks
Hotness
Smartphone tracking using GNSS - Team 54
Updated 1mo ago
0 comments · Google Smartphone Decimeter Challenge 2022
0
Batch 54
Updated 2mo ago
0 comments · Google Smartphone Decimeter Challenge 2022
1
coding
0


Google Smartphone Decimeter C
kaggle.com/competitions/smartphone-decimeter-2022/discussion
Update
kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Learn
More
Your Work
RECENTLY VIEWED
Google Smartphone D...
GSDC2022_kalmalfilter
Zillow Prize: Zillow's H...
View Active Events
Research Prediction Competition
Google Smartphone Decimeter Challenge 2022
Improve high precision GNSS positioning and navigation accuracy on smartphones
$10,000
Prize Money
Google · 573 teams · 5 months ago
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
Submissions
New Topic
Discussions
Follow
Search discussions
Filters
All
Owned
Bookmarks
Hotness
Pinned topics
Smartphone Decimeter Challenge at ION GNSS+ 2022
Ashley Chow · Last comment 2mo ago by Ashley Chow
11
4 comments
Recap of Competition - Congratulations to the Winners!
Ashley Chow · Posted 4mo ago
6

Assessments - Report (cont.)

- **Convincingness:** the goal of the project report is to convince readers that your proposed solution is proper to solve the specific machine learning task. To do so, in the report,
  - You need to provide detailed motivations and explanations of the techniques you used in the solution, e.g., what is the motivation of a new feature you proposed, why you proposed a specific pre-processing step, why you use the proposed classifier but not others
  - You also need to conduct experiments to further verify your proposed ideas

Assessments – Report (cont.)

- Weight priority:

  Convincingness = Leaderboard Performance > Solution Novelty = Writing

Frequent Q&A

- Can we choose another Kaggle competition which is not on the candidate list?
  - No, you can only choose one from the candidate list.
- Are there requirements on the format of the report?
  - 12 point font, single space, $\leq$ 20 pages.
- What if our report or presentation video exceed the length limit?
  - We will only grade till the length limit for fairness.
- Can we use other ML techniques beyond what are taught in this module, e.g., deep learning models, for the course project? Can we use other data sources than those provided by the specific competition?
  - Yes, you are free to do so as long as you find them useful.
- Can the report and video submission deadline be extended?
  - No, it is a hard deadline.

Thank you!