- Data Mining Project -

kaggle™

©Anna Roytberg & Galina Blokh , ITC 2020

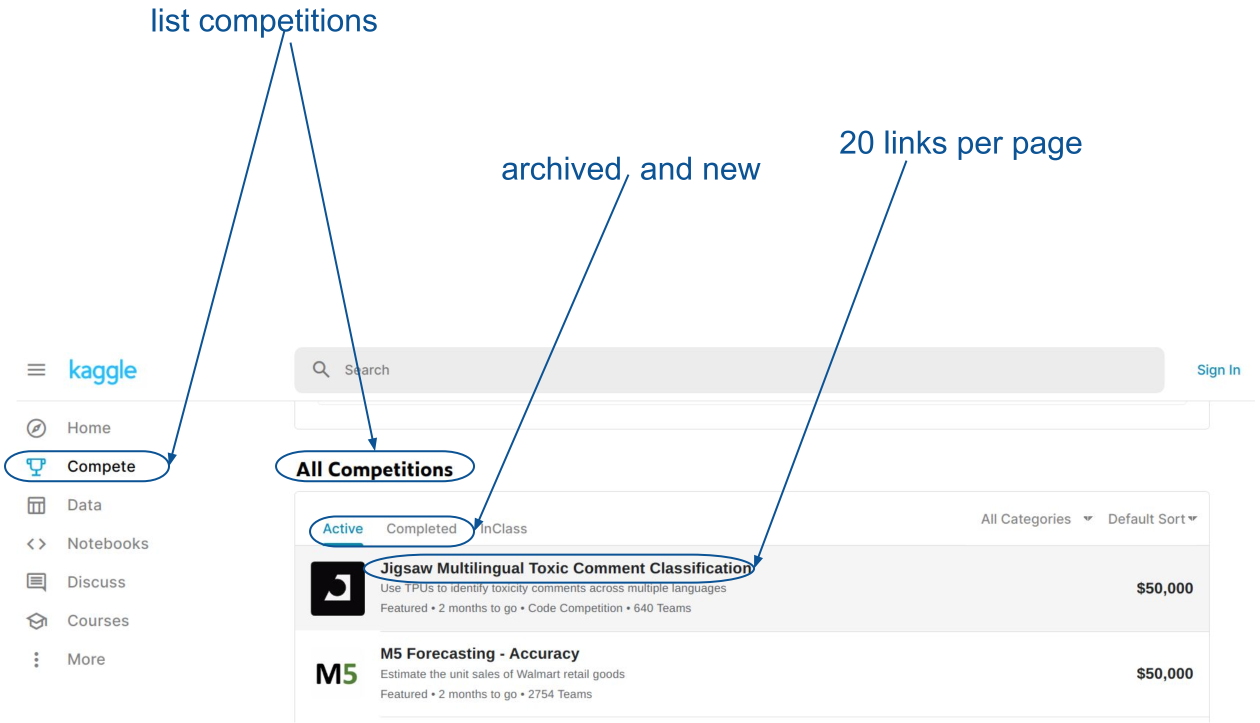
ಠ‿ಠ

The data

First stage

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners.

Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.



list competitions

archived, and new

20 links per page

Jigsaw Multilingual Toxic Comment Classification
Use TPUs to identify toxicity comments across multiple languages
Featured • 2 months to go • Code Competition • 640 Teams
$50,000

M5 Forecasting - Accuracy
Estimate the unit sales of Walmart retail goods
Featured • 2 months to go • 2754 Teams
$50,000

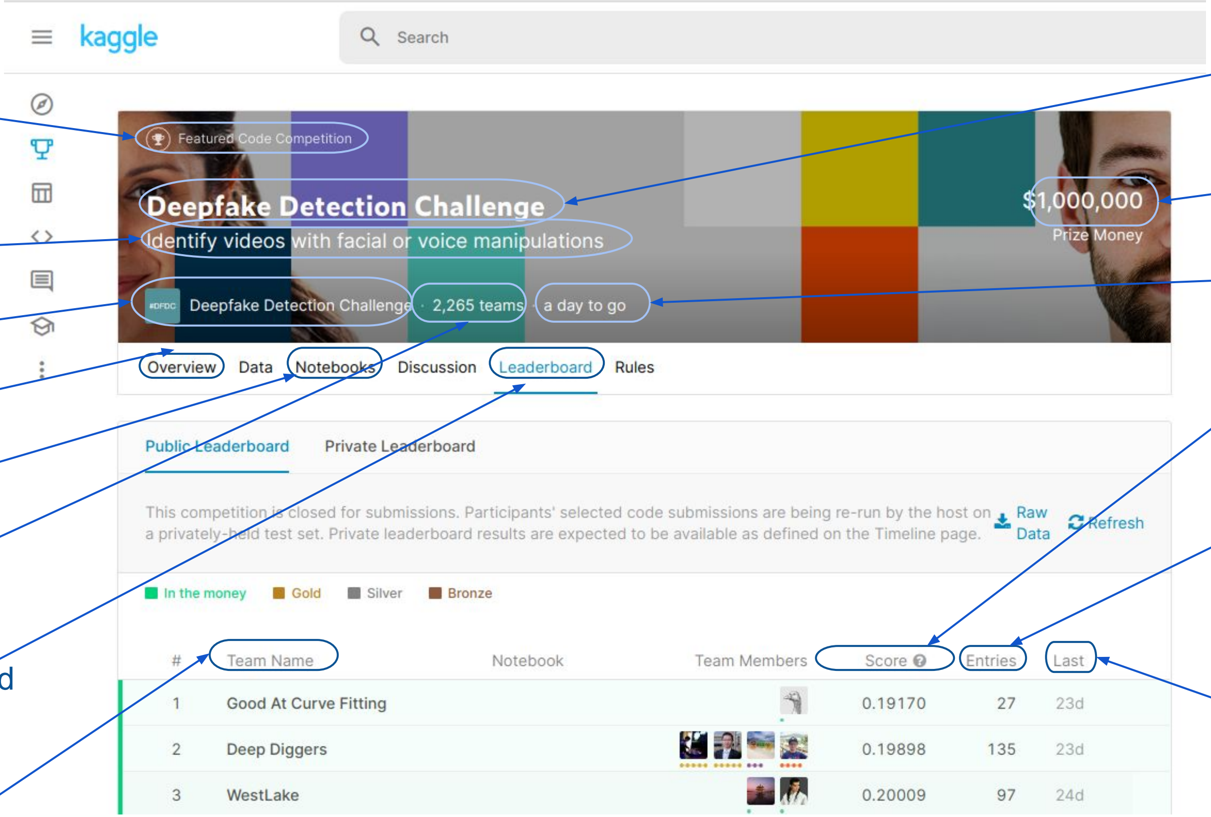
ಠ.ಠ

©https://en.wikipedia.org/wiki/Kaggle

The data

second stage

(~˘▾˘)~



Featured Code Competition

Deepfake Detection Challenge

Identify videos with facial or voice manipulations

Deepfake Detection Challenge · 2,265 teams · a day to go

$1,000,000 Prize Money

Overview Data Notebooks Discussion Leaderboard Rules

Public Leaderboard Private Leaderboard

This competition is closed for submissions. Participants' selected code submissions are being re-run by the host on a privately-held test set. Private leaderboard results are expected to be available as defined on the Timeline page.

Raw Data Refresh

🟩 In the money 🟧 Gold ⬜ Silver 🟫 Bronze

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Good At Curve Fitting			0.19170	27	23d
2	Deep Diggers			0.19898	135	23d
3	WestLake			0.20009	97	24d

topic

title

prize

header

end date

organizator

description

score

notebooks

entries

teams count

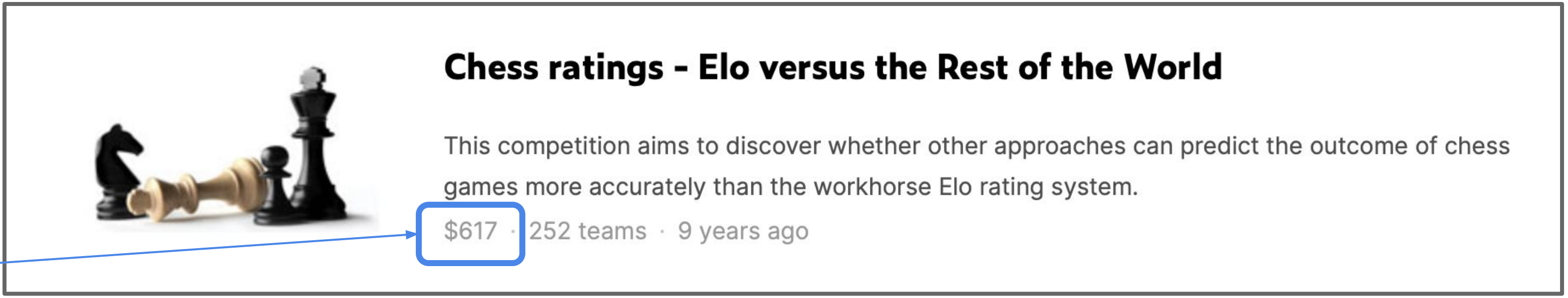
leaderboard

start date

team name

Kaggle changed competition page view:

Before:

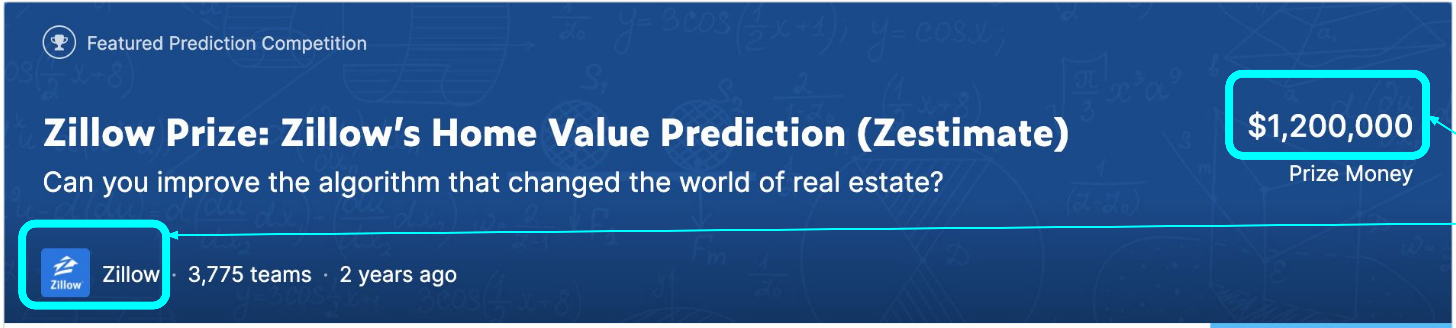


Chess ratings - Elo versus the Rest of the World

This competition aims to discover whether other approaches can predict the outcome of chess games more accurately than the workhorse Elo rating system.

$617 · 252 teams · 9 years ago

⊂◉‿◉つ



Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Zillow · 3,775 teams · 2 years ago

$1,200,000

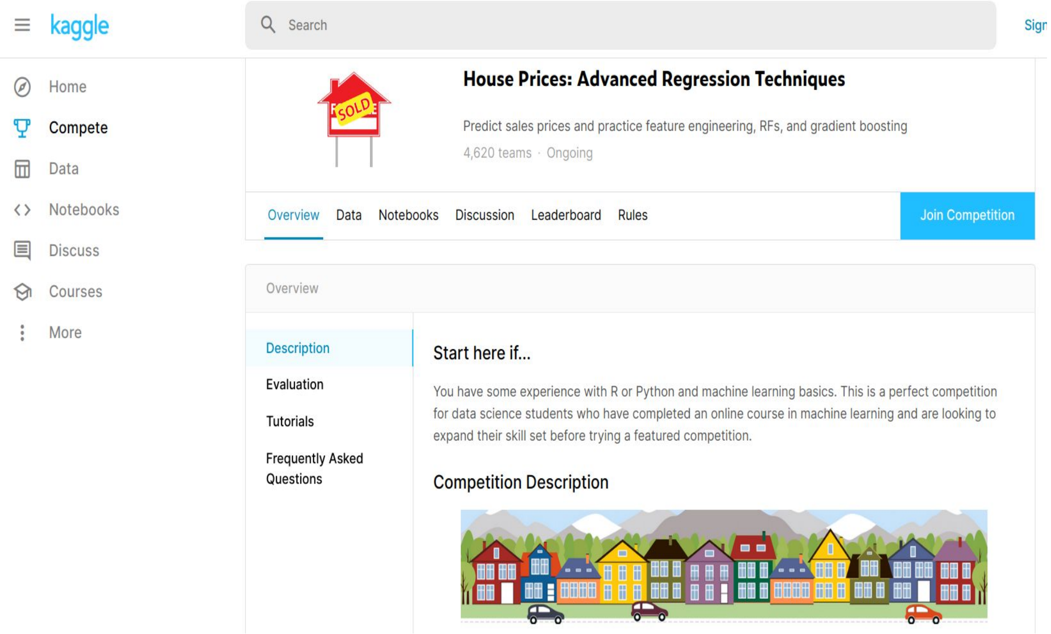
Prize Money

: After

Some numbers of how much data we scraped

- 5,000 competitions
- 15 000 scrapped pages
- 90 000 teams
- Data from 2010 – 2020

⊙_⊙



House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,620 teams · Ongoing

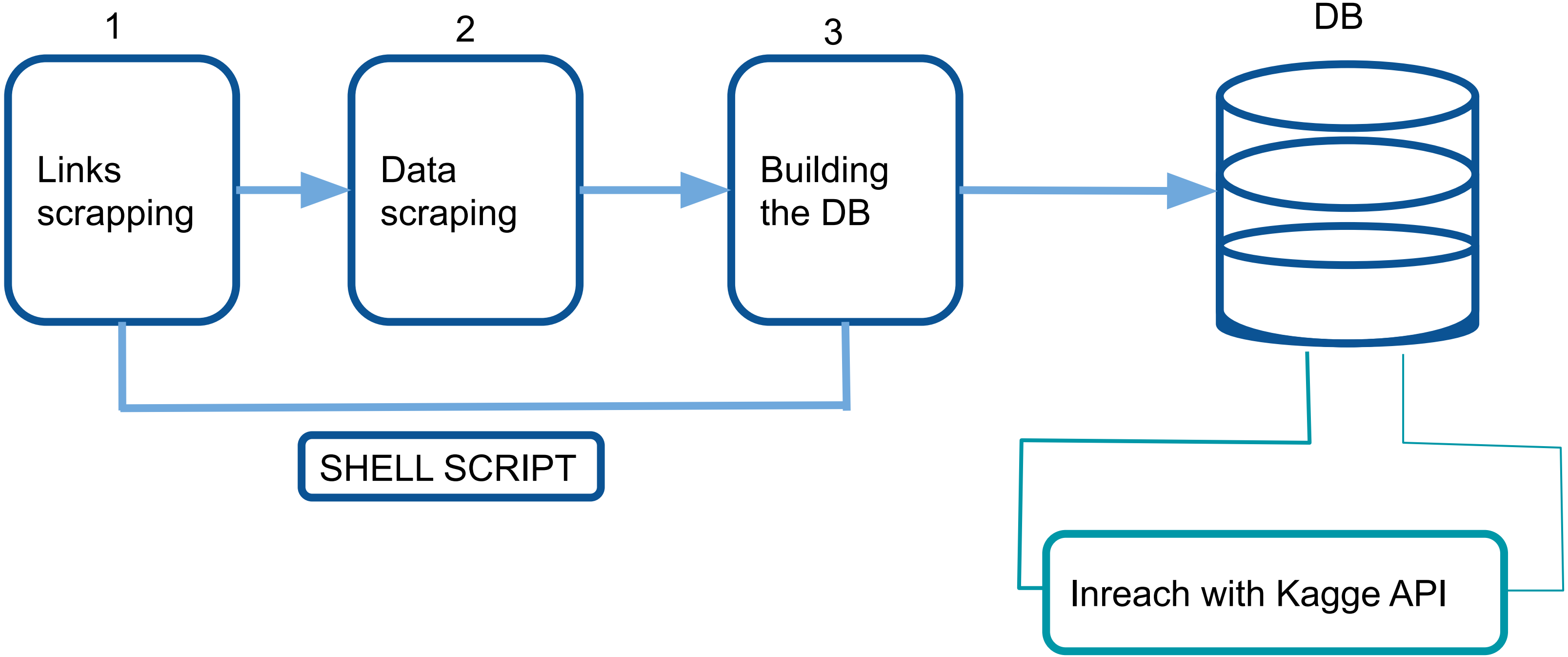
Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

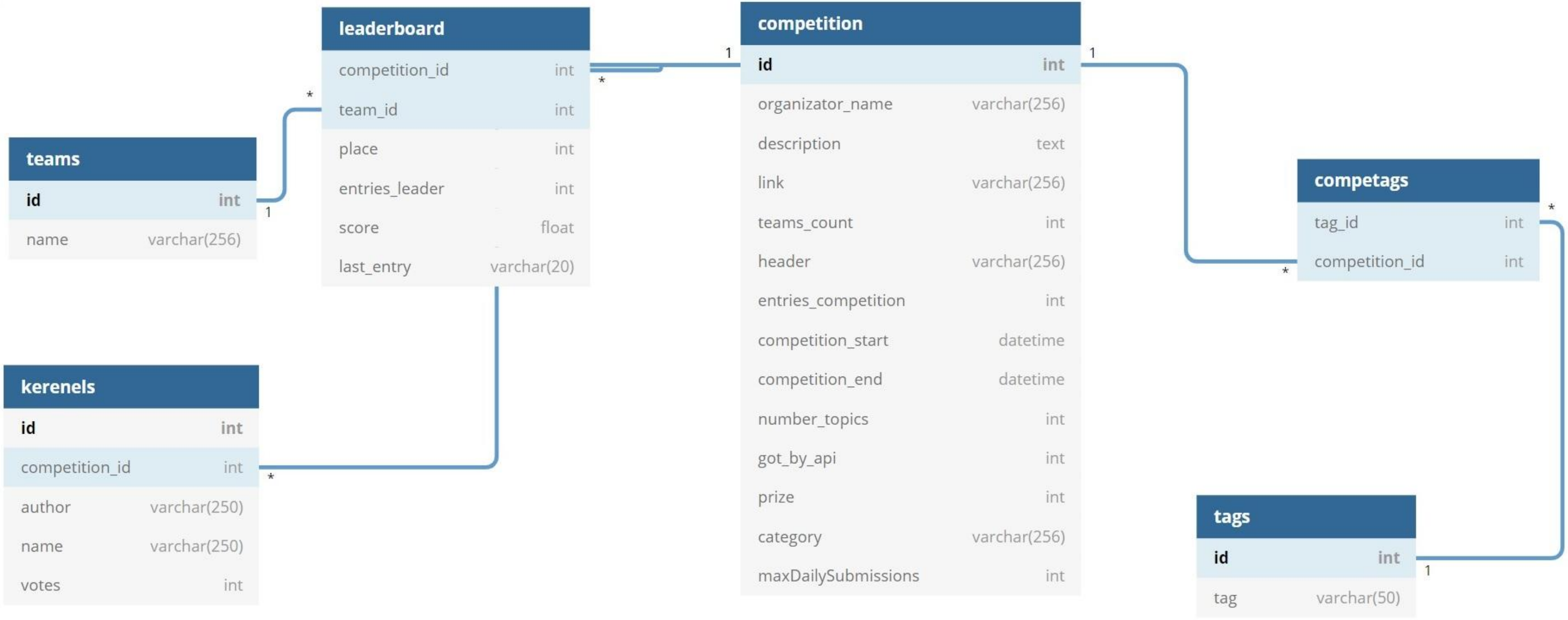
Competition Description

Project structure

(⌐■_■)



Database structure



leaderboard	
competition_id	int
team_id	int
place	int
entries_leader	int
score	float
last_entry	varchar(20)

competition	
id	**int**
organizator_name	varchar(256)
description	text
link	varchar(256)
teams_count	int
header	varchar(256)
entries_competition	int
competition_start	datetime
competition_end	datetime
number_topics	int
got_by_api	int
prize	int
category	varchar(256)
maxDailySubmissions	int

teams	
id	**int**
name	varchar(256)

competags	
tag_id	int
competition_id	int

kerenels	
id	**int**
competition_id	int
author	varchar(250)
name	varchar(250)
votes	int

tags	
id	**int**
tag	varchar(50)

*why we did our structure in this way?

(•_•)>⌐■-■

©https://dbdiagram.io/d/5e8392fb4495b02c3b890e62

Kaggle api

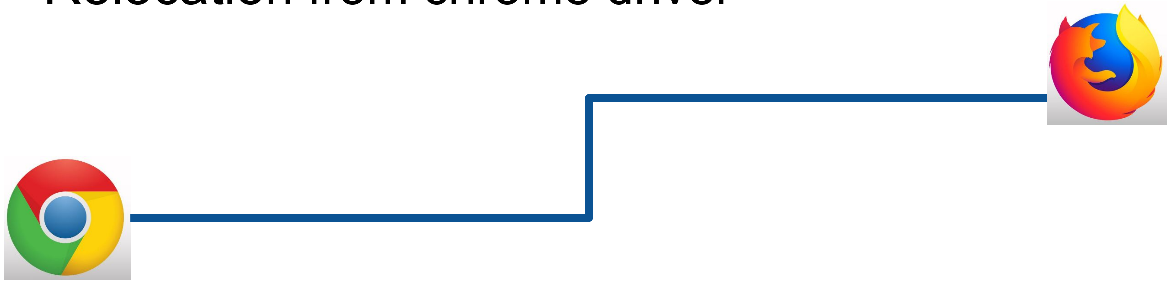
pro	contra
+ quick work	- Scrap information only for 400 competitions of 6500
+ fast connection	- need to create account on kaggle to get a private key

(•̀ᴗ•́)و ̑̑

Challenges we experienced and how we faced them

- At first the feeling that “I didn’t understand at all what was wanted from me” ಠ_ಠ
- Kaggle is pretty old and changed the page structure several times
- A lot of data: downloading takes a lot of time

- Relocation from chrome driver

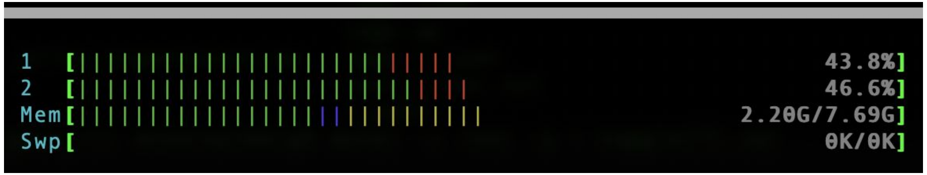


Challenges we experienced and how we faced them

- Troubles on Amazon server with connections ("!!!!!a server has to get rest. And what about me?!")
- Memory issue: ಠ_ಥ

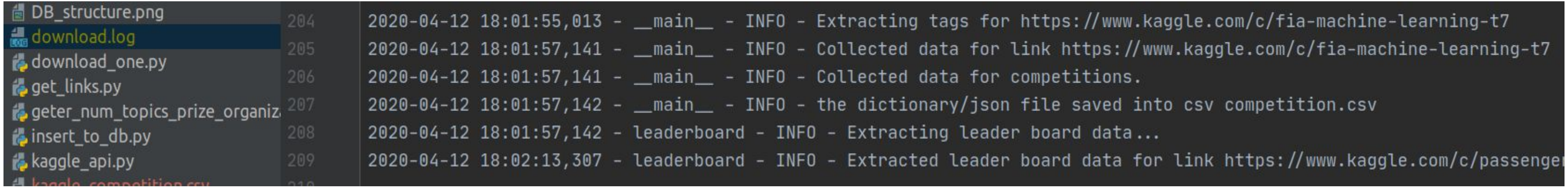
  the driver get all the memory and AWS freezes us.

  Solution: recreate the driver instance every 25 links



```
1   [|||||||||||||||||||||||||||                     43.8%]
2   [||||||||||||||||||||||||||||                    46.6%]
Mem[|||||||||||||||||||||||||||||               2.20G/7.69G]
Swp[                                                 0K/0K]
```

Never forget about logging!!!



```
DB_structure.png
download.log
download_one.py
get_links.py
geter_num_topics_prize_organiz
insert_to_db.py
kaggle_api.py
kaggle_competition.csv
```

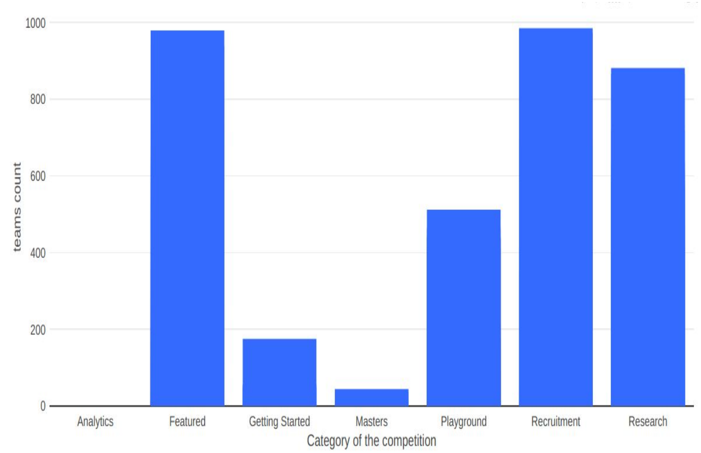
```
2020-04-12 18:01:55,013 - __main__ - INFO - Extracting tags for https://www.kaggle.com/c/fia-machine-learning-t7
2020-04-12 18:01:57,141 - __main__ - INFO - Collected data for link https://www.kaggle.com/c/fia-machine-learning-t7
2020-04-12 18:01:57,141 - __main__ - INFO - Collected data for competitions.
2020-04-12 18:01:57,142 - __main__ - INFO - the dictionary/json file saved into csv competition.csv
2020-04-12 18:01:57,142 - leaderboard - INFO - Extracting leader board data...
2020-04-12 18:02:13,307 - leaderboard - INFO - Extracted leader board data for link https://www.kaggle.com/c/passenge
```

A short demo from redash.io

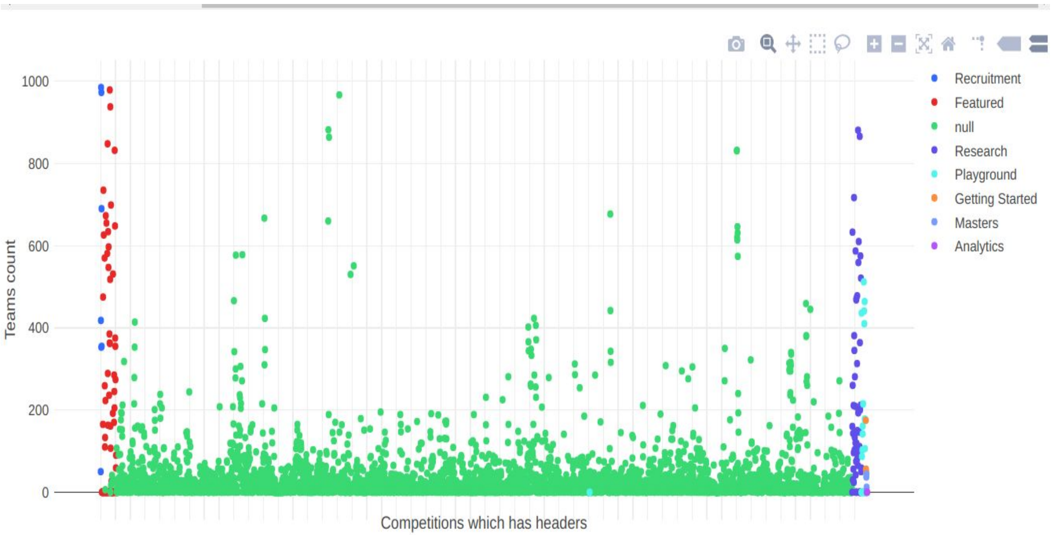
°˖✧◝(⁰▿⁰)◜✧˖°

Not a good platform for making visualisations!!!!

Teams count grouped by category



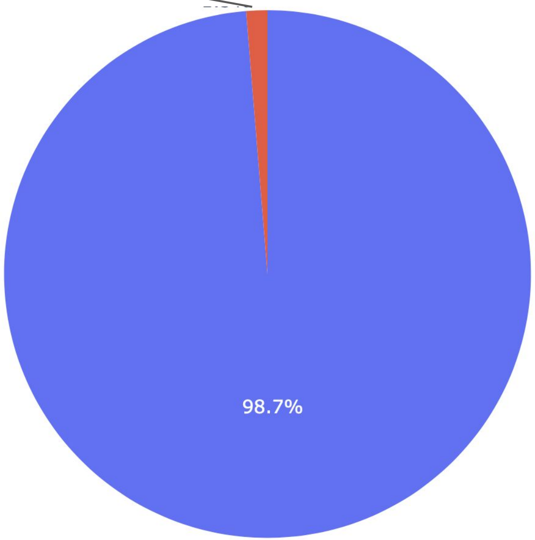
Competitions by teams count grouped by category



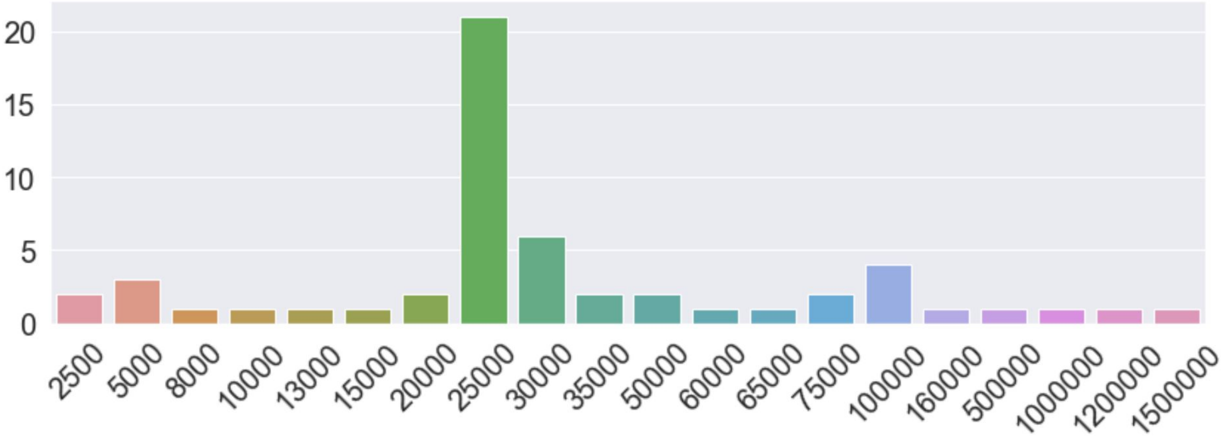
©http://ec2-54-183-241-235.us-west-1.compute.amazonaws.com/queries/3#6

Prizes on Kaggle could be huge, but they are extremely rare

Only 1.3% have prize



The biggest prize - $1500000

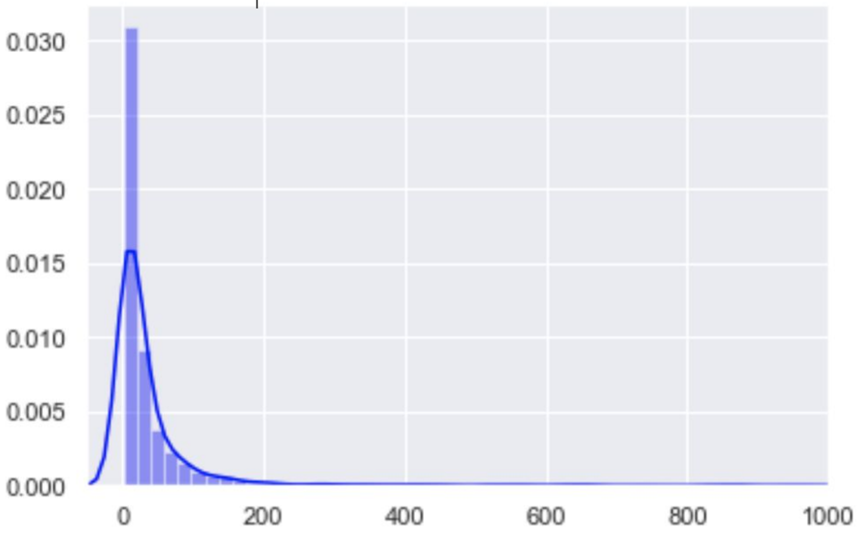


☆ ˋ(◍‿̈◍)ˊ ☆

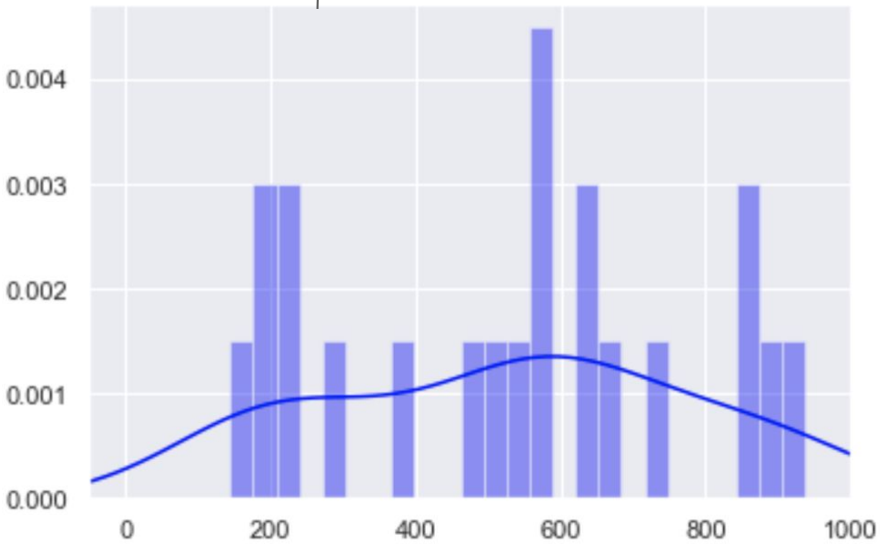
More teams if there is a prize

(╭ರ_•́) Number of teams

Without prize

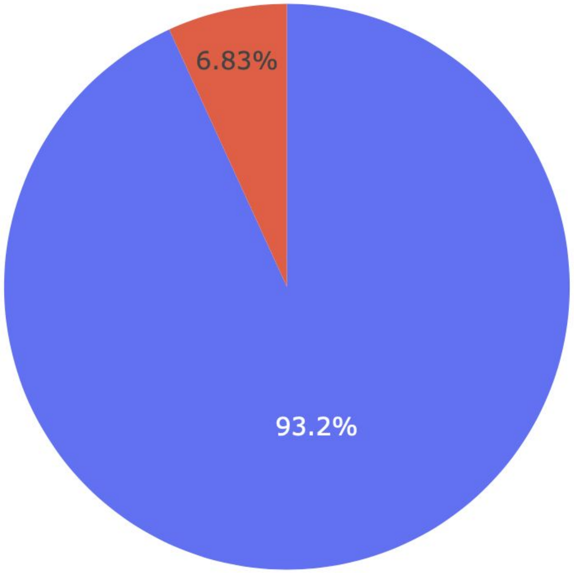


With prize

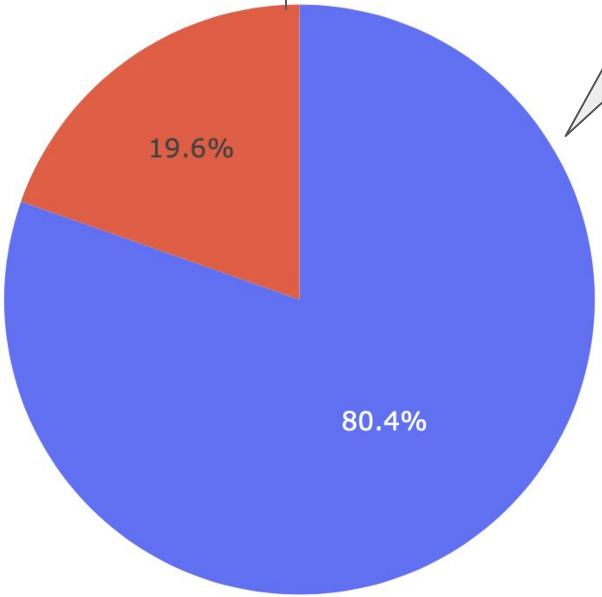


Number of topics in discussion

Without prize



With prize



People discuss more the competitions with prizes

☆(✿‿✿)☆

Data Insights

- Very little of Kaggle competitions have prizes
- Competitions popularity doesn’t depend on size of a prize
- Most of the competitions are some university assignments
- There are no teams that won more than 1 big prize.

Good luck on Kaggle!

f(ಠ‿↼)z