Quiz Game – Data Analysis

Fabio Pulvirenti

Quiz Game – Data Analysis

In this first data analysis, I will have fun analyzing the data from a quiz game.

I will have access just to a collection user feedbacks related to the questions.

Dataset Details

Each record of the dataset represents a set of data for a single couple «question-answer»

- Information related to the question
- Information related to the answer and the feedback provided by the user during a game

Specifically, each record is characterized by the following features:

- Question ID : question identifier
- Category ID: category of the question
- Game ID: single game / match identifier
- Question Type: text / image question
- Answer : Correct / Wrong / Other
- Vote : feedback on the questions (thumb up / down)

Dataset Details

Start with some general information about the dataset*:

- Records: Over 1,000,000
  - Rated questions
- Different players: Over 57,000
- Played games: Over 500,000
- Possible different questions: ~ 40,000
- Possibile different question categories: 20

* Collected after the data cleaning process in which all the not provided / null fields were discarded

Analysis

The analysis will cover measures and statistics related to:

1. Questions
2. Games
3. Players

Unfortunately, I do not have any information about the coverage of the provided dataset with respect to the full data collection:

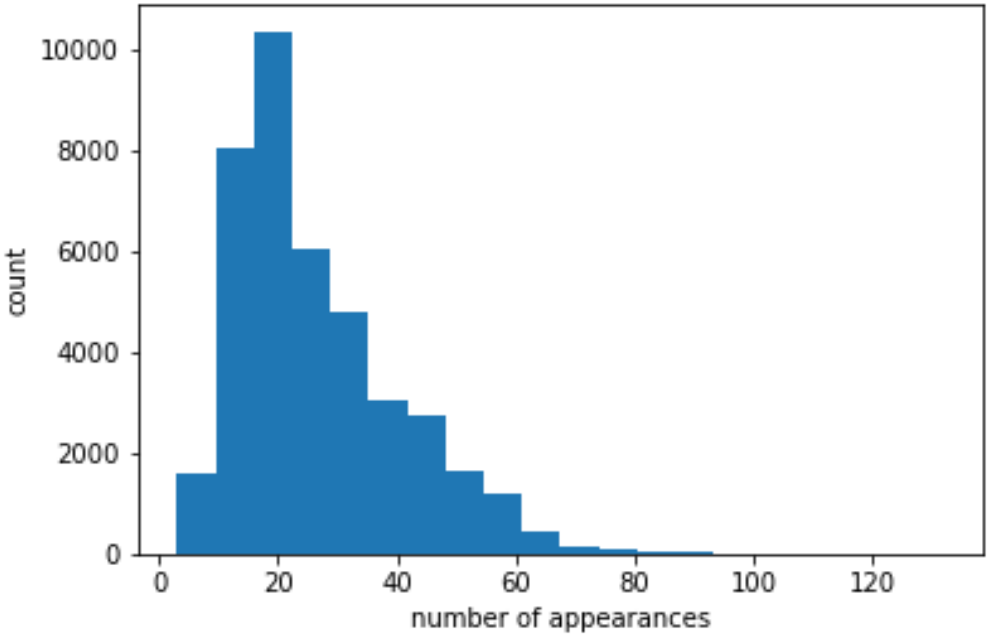
- It is probably just a shard / snapshot
- I do not have any temporal reference (took away a lot of fun!)
- I am assuming that the data distribution of the analyzed dataset respects and is proportional to the distribution of the full data

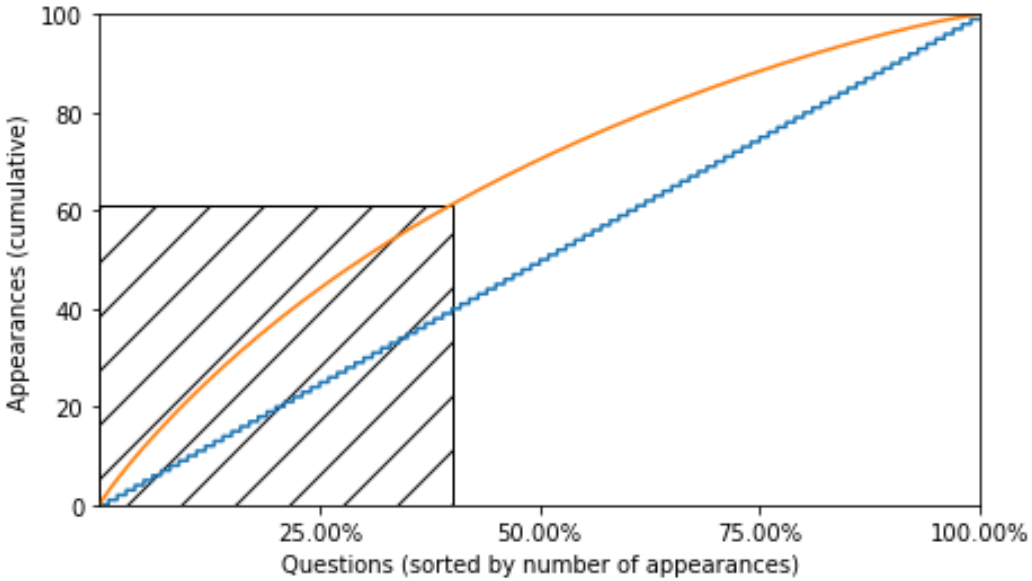
1. Questions

2. Games
3. Players

Questions Data at a Glance

- Possible different questions: ~ 40,000
- Possible different question categories: 20
- Possible question types: 2 (Text or image)
- Average of correct answers: 51%
- Average of positively rated: 68%
- Each question is asked ~ 26 times in the average
- However 40% of the questions achieve the 60% of all the occurrences
  - The orange line represents the aggregated number of played games per user
  - The blue lines in the graph represents a theoretical uniform distribution of games per users

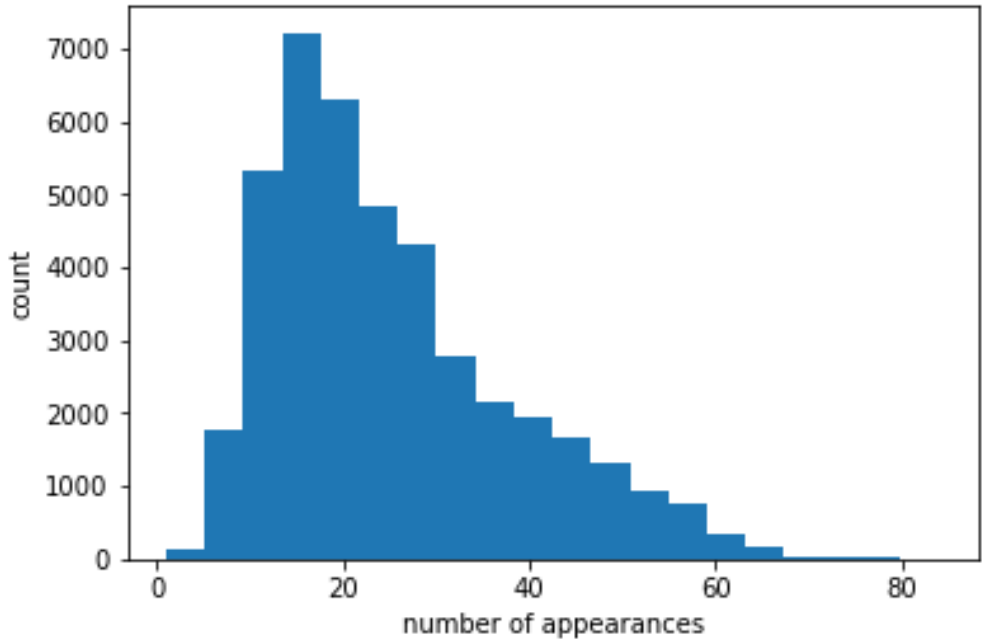


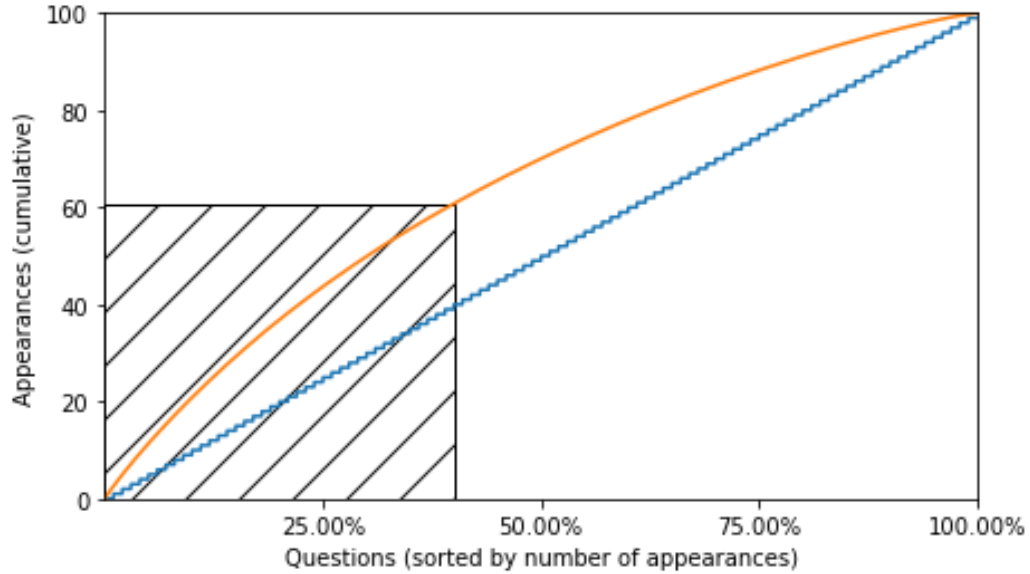


Questions Data at a Glance

What does this mean?

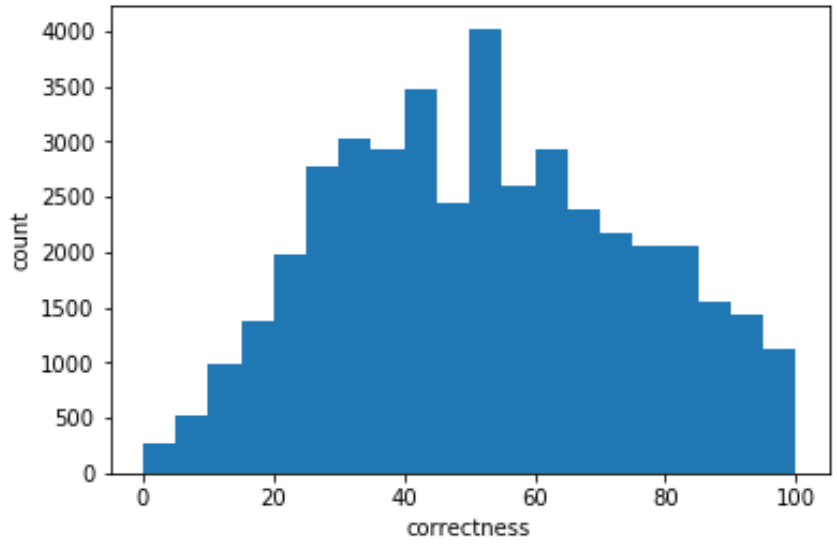
- Not all the questions are proposed the same number of times
- 60% of the questions are proposed just in few games
  - They share only the 40% of the total number of played games



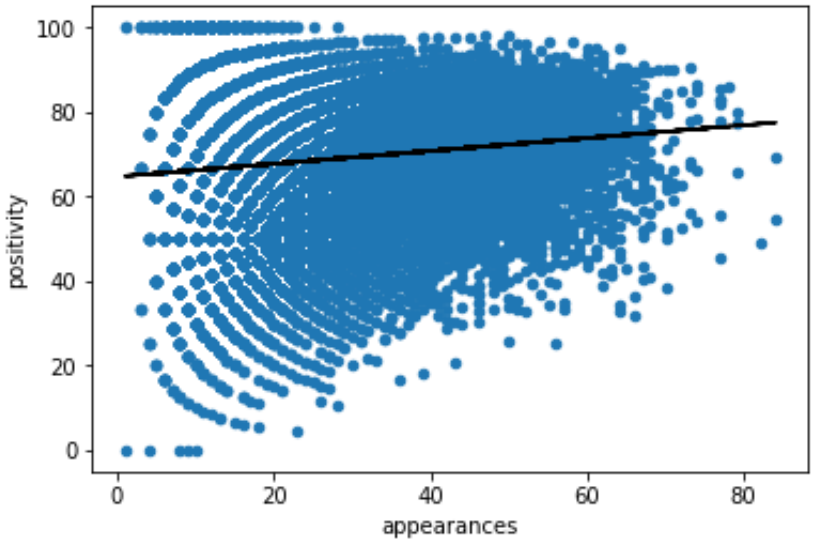


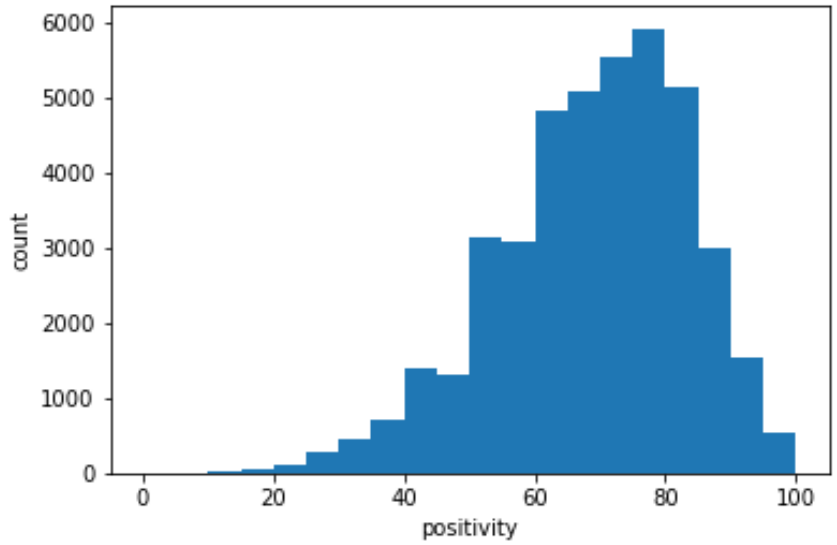
Correct answers and players feedbacks

- No correlation between the number of correct answers and their number of occurrences



- The correlation between the rate of the questions and the number of times that are proposed is low but does exist: from these data it seems that **the most appreciated questions are proposed more often.**



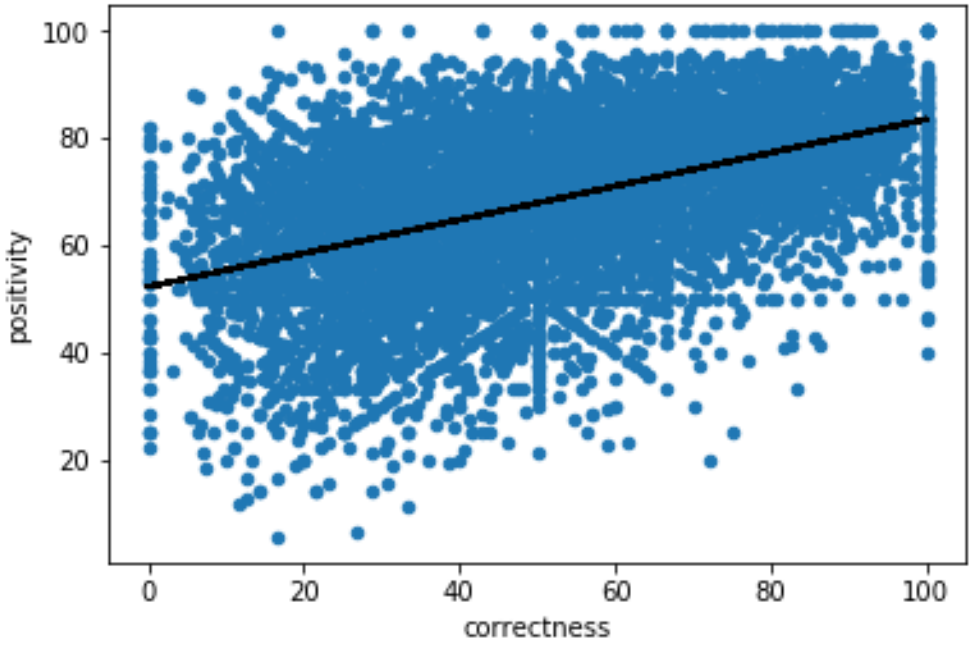


‘’ I like it… if I know it!’’

Quite strong correlation between the number of correct answers and the rate of the question (Pearson Corr. 0.48, the maximum is 1)

People are more likely to appreciate the questions for which they know the answer

- Maybe they reward the clarity of the question?



Text vs Image Questions

Some questions provide an image

We have seen that this type of questions is:

- Easier to answer (62% vs 51%)
- More appreciated (72% vs 68%)

Interesting: some of the questions can be represented in both the ways.

Those represented also by a picture are most frequently correctly answered and appreciated also in their text version

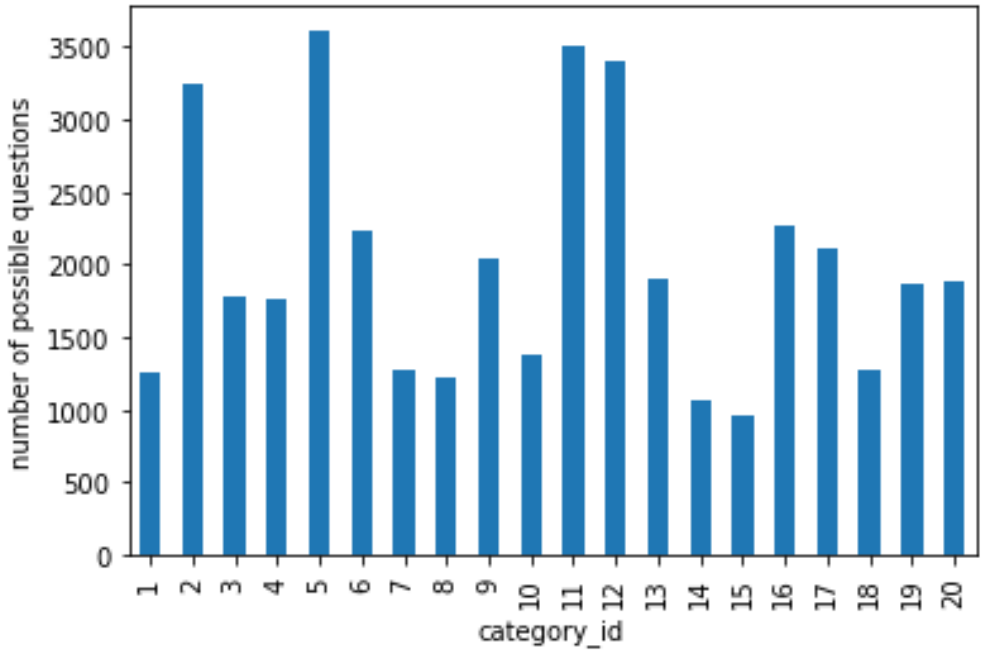
- Could it be related to their «form»?
  - They can be related even to a picture
  - However, the illustrated version is still more effective

	Text	Image	Text Questions (just text)	Text Questions with an illustrated version
Correct Answers	51%	62%	51% ⟷	59%
Positive Rate	68%	72%	68% ⟷	71%

Question Categories

There are 20 categories* of questions

- For each category, on average, there are ~ 2000 questions
  - However, the categories are very unbalanced (standard deviation of more than 800 questions)
  - The reason behind this different amount of possible questions could be related to the «popularity» of the category

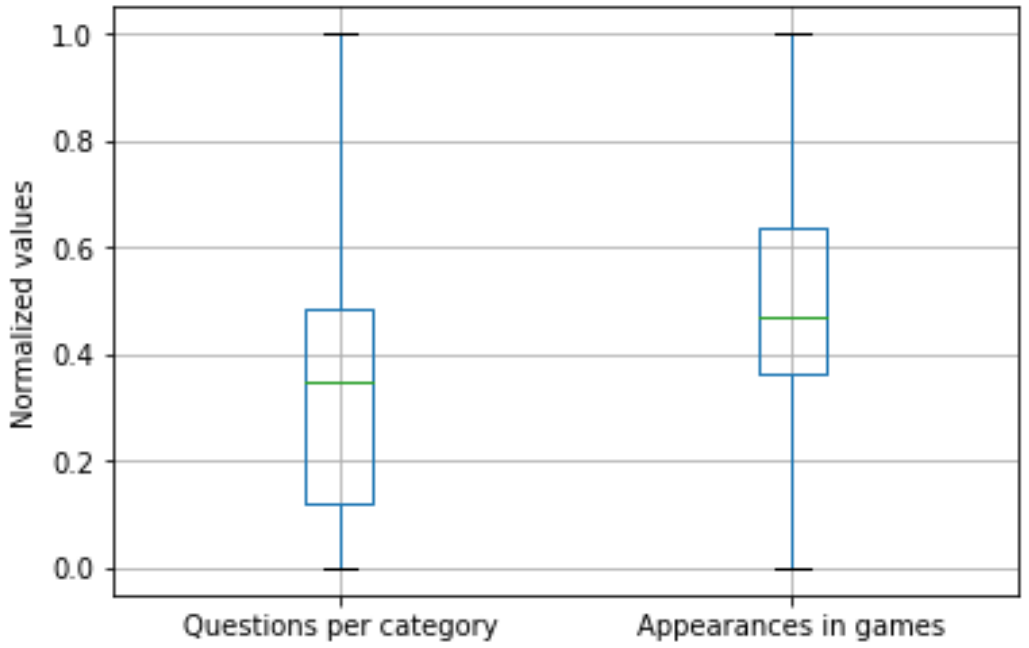


* Plus a «skip» category and a «null» one removed for this analysis

Question Categories

Despite their different number of possible questions, the distribution of the number of times in which each category occurs in the games is relatively more balanced

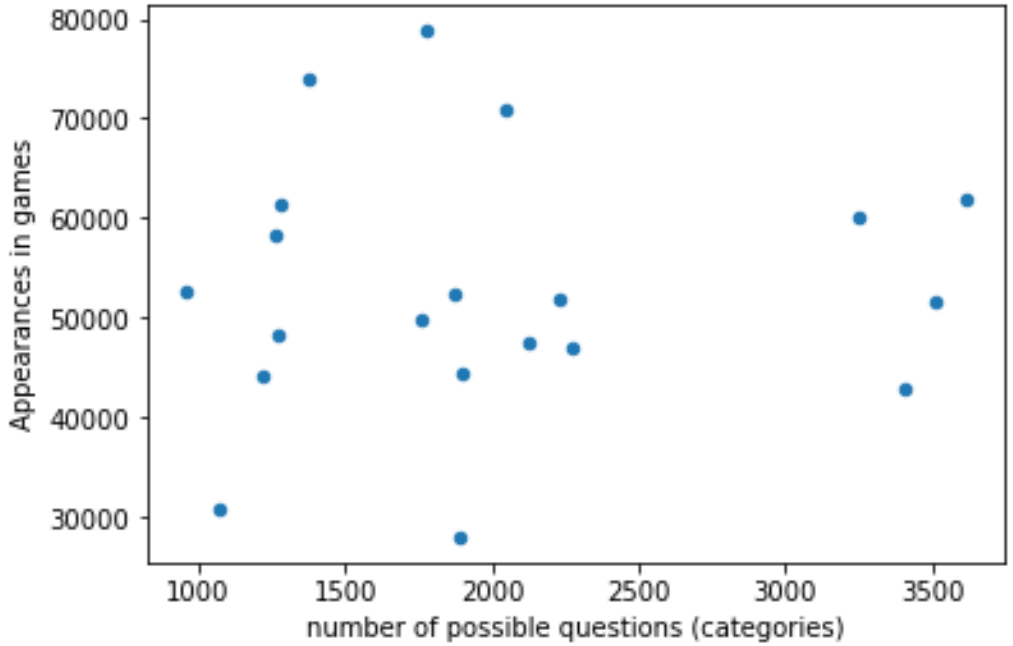
- You first choose the category
- **If you select the least populated categories, you will have less variety of possible questions**



Question Categories and number of questions

Because of the different number of questions per category, **I would have expected a relationship with the respective «popularity»**

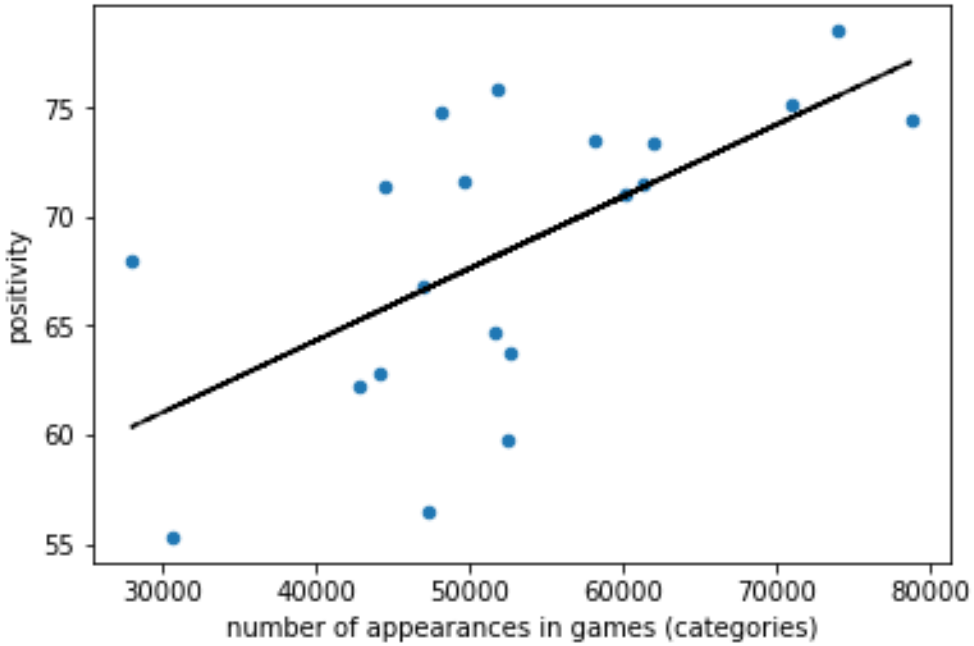
- i.e. more questions are provided by the authors to the most selected categories in order to deliver more variance
- Instead, these features are very poorly correlated (less than 0.1, while I was expecting a negative correlation)



Question Categories and positive feedbacks

There is a strong correlation (Pearson 0.68) between the number of games in which the category is proposed and the average rating score of the questions

- **People mostly tend to appreciate the questions of the category they have chosen**
  - This is also related to the fact that they appreciate more the questions they know (already discussed)
- Alternatively, the system tries to propose the categories with the highest positivity



1. Questions

2. Games

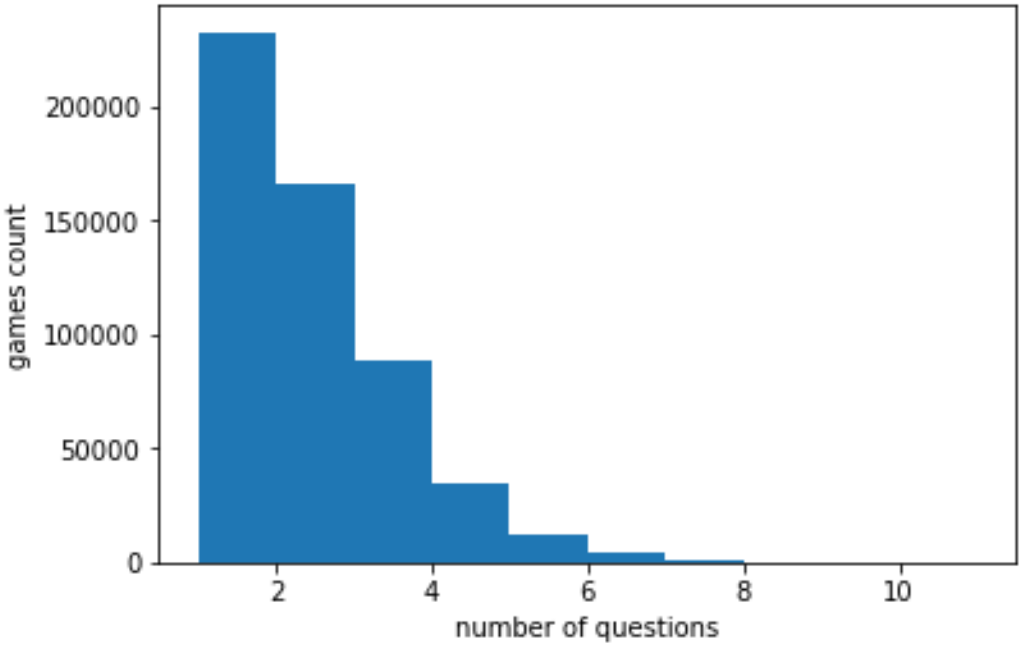
3. Players

Games

The questions in the dataset belong to ~ 550,000 different games (matches)

- Only few of them are related to 2 players (13%): most of the matches are related to just one player
- The number of questions per match is quite low (less than 2 on average)

Probably, not all the questions in a game are rated and the dataset contains just the rated ones.



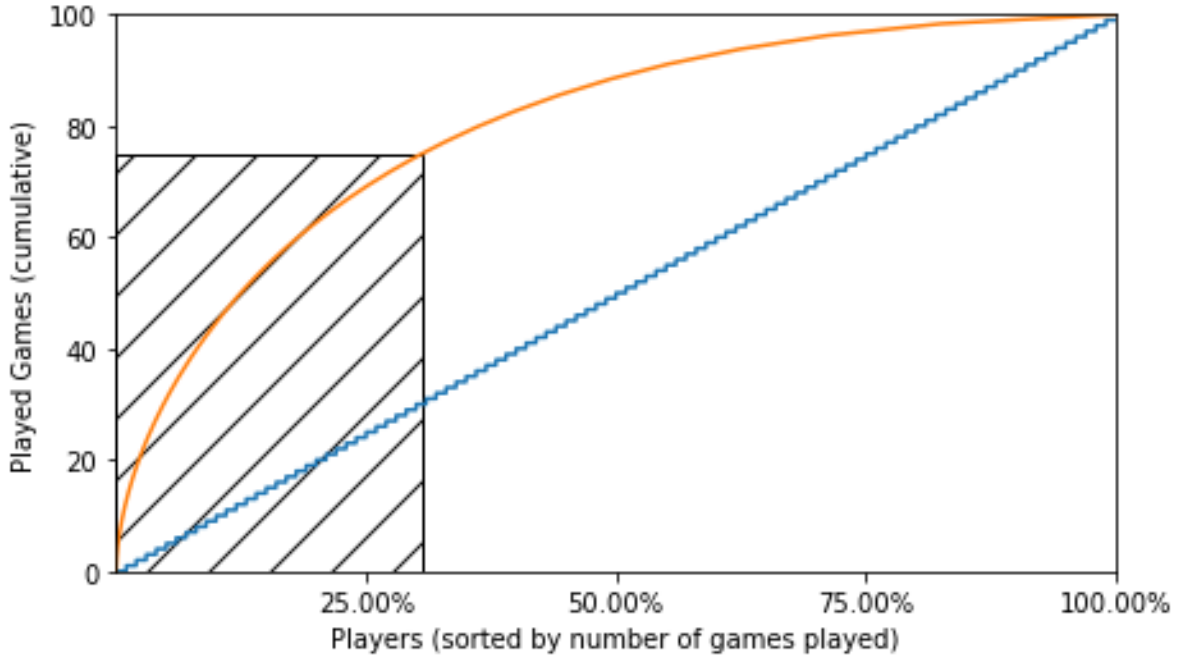
1. Questions
2. Games

3. Players

Players and played games

Let us focus on the players

- A lot of users: ~ 57,000
- However, not with the same behavior
- Average number of played games per user: ~ 10
- 75% of the games played by the top 30% of the users!*
  - The orange line represents the aggregated number of played games per user
  - The blue line in the graph represents a theoretical uniform distribution of games per users



*Assuming that the analyzed data (the rated questions) reflect the nature of the whole scenario

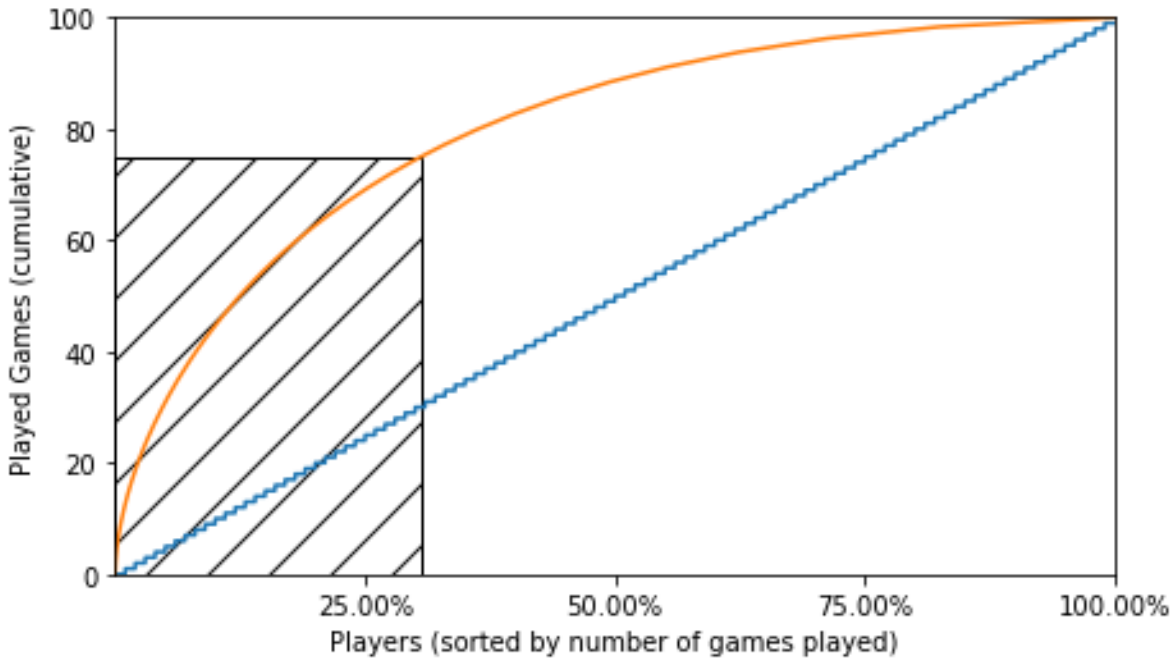
Players and played games

What does this mean?

- Just a small percentage of all the users is playing a lot

Pareto Principle effect (almost!):

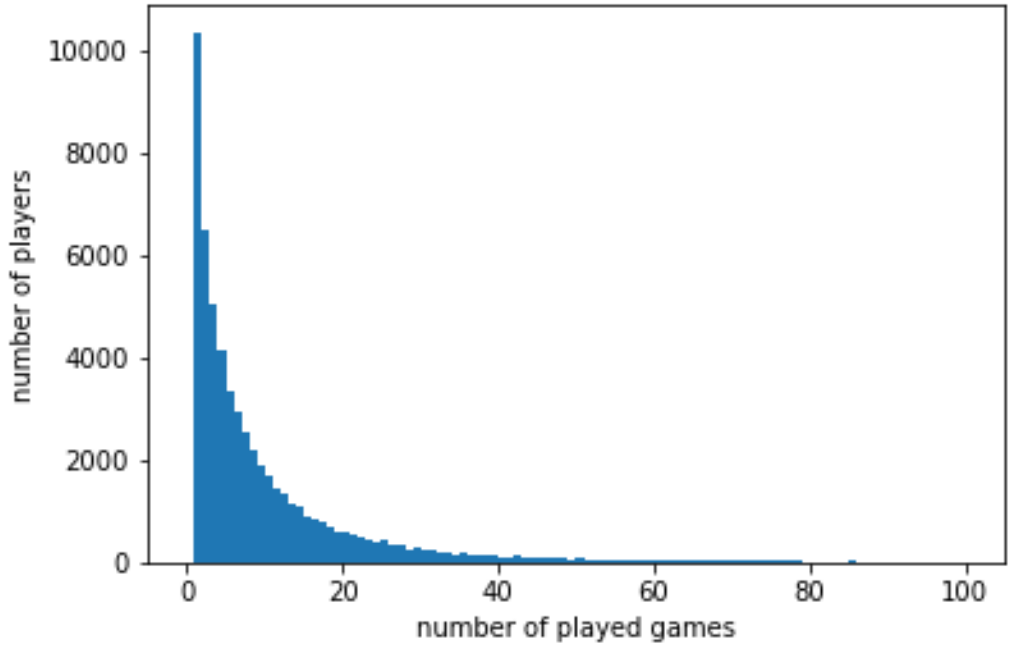
- 30% of the users played 75% of the games
- 70% of the users played just a few games
  - They share only the 25% of the total number of played games



Players and played games

There is a huge number of players who have played just a few games

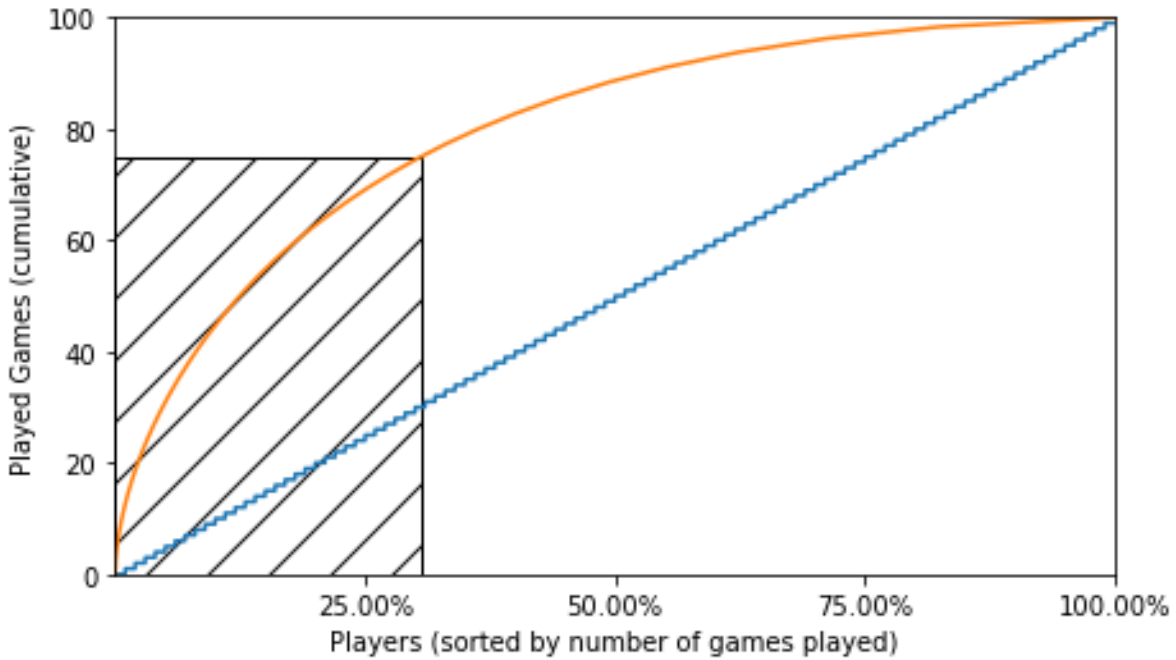
- A lot of them (18%) have played just 1 game: casual gamers or newbies?
- A lot of them (11%) have played 2 games
- Almost 40% of the users have played less than 4 games



Players and played games

Two possible strategies:

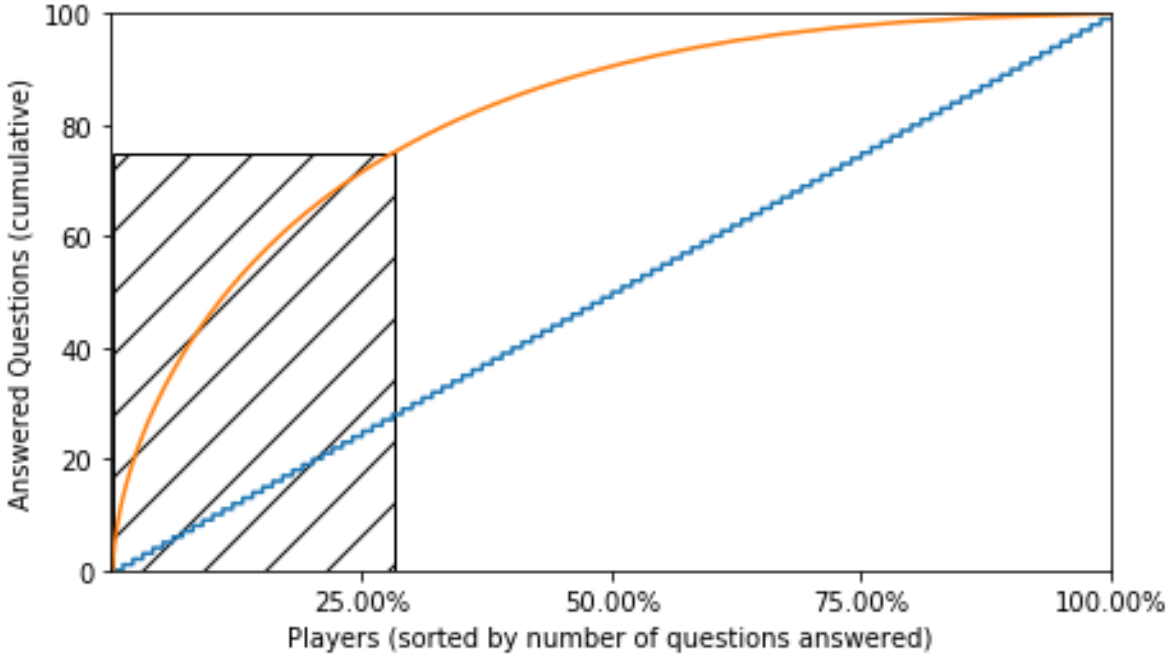
- Focusing on heavy hitters (most active players)
- Trying to involve more the less active players
- Best one? Probably depending on the revenue-model
  - Revenue proportional to players or games / questions?
  - Probably to questions (ad after the question?)
  - What about the players of the premium version?



Players and questions

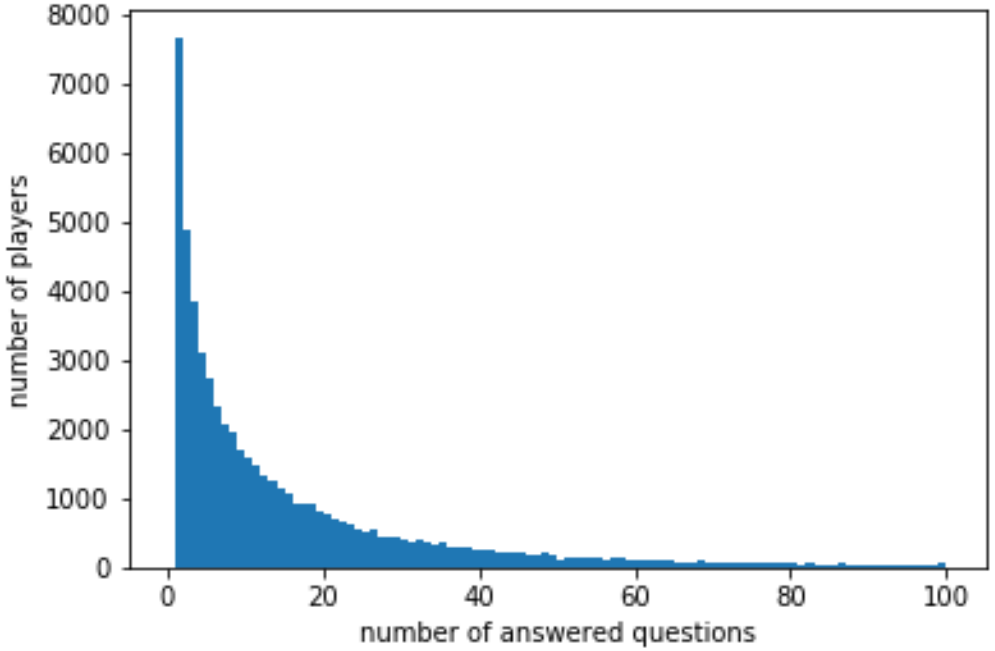
Very similar behavior in the relationship player- answered questions

- 18.5 questions answered per player on average
- More than 75% of the questions answered by the top 30% of the users!
- Predictable:
  - The number of questions and the number of played games are strongly correlated (Pearson Correlation 0.98, the max value is 1)



Players and questions

- 13 % of the users have answered to just 1 question
  - Almost 40% of the users quit after answering less than 5 questions
- **Question: are they discouraged after failing?**
  - In this case an increasing difficulty rate of question could be a solution to push these players to hold on
- **Answer: No, let's see why…**



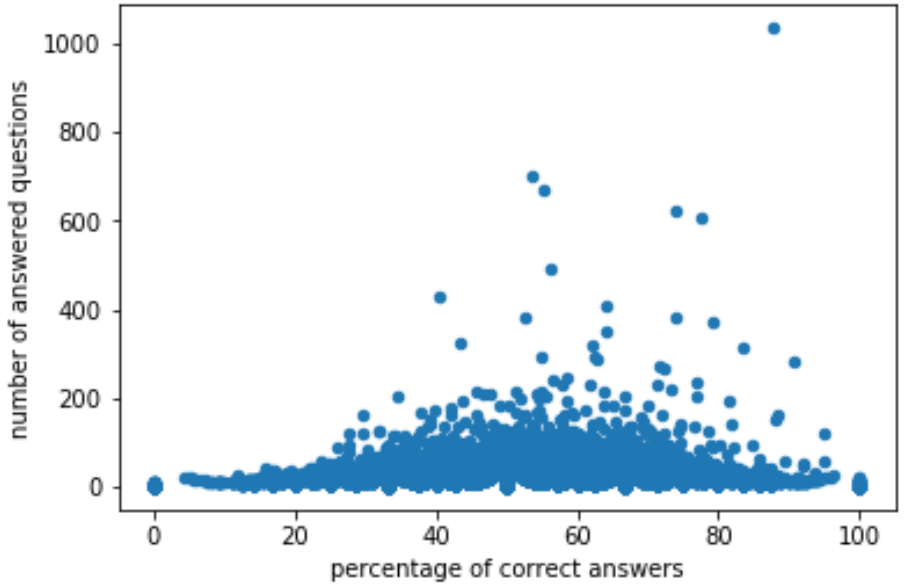
Players and questions

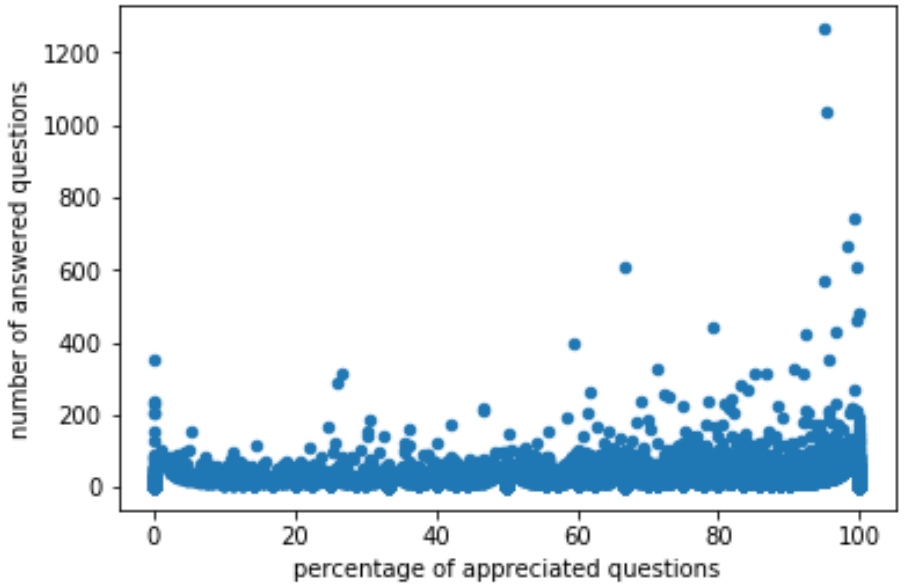
Very low or no correlation between:

- **Number of played games and percentage of right answers (Pearson: 0.006)**
  - **A high value could have meant the presence of a kind of «discouraging» factor**
- Number of played games and percentage of positively rated answers (Pearson: 0.10)

Note that the correlation with the «positivity» is slightly higher than the one with the «correctness» of the answer

- **It is more likely that they quit because they do not like the questions rather than because of the difficulty of the game.**

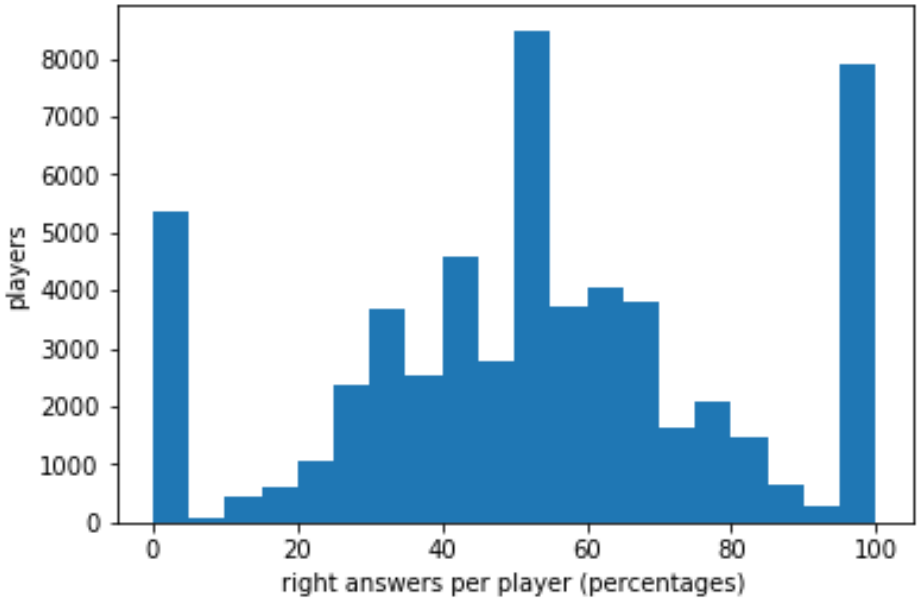


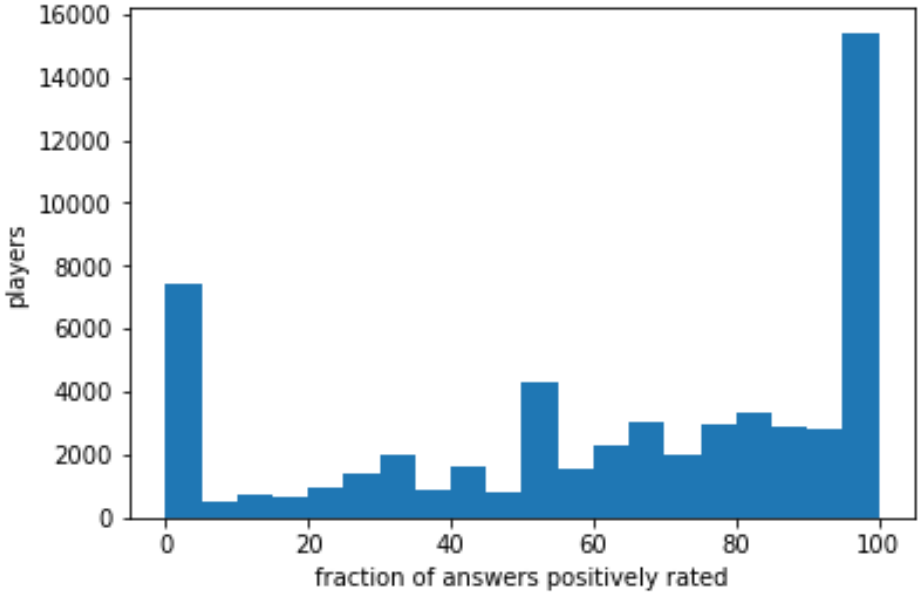


Other per-user statistics

- Correct answers per user: average = 53%, Very unbalanced towards the extreme (0% - 100%)
- Positively rated questions: average = 63%

Question: which is the impact of the not expert players? How do they behave?





Non-Expert players or Newbies (Players with less than 5 answers)

Interesting:

- They have a higher score than all the users in terms of right answers (55% vs 53%) *
- They are less positive towards the question rating (58% vs 63%)

Again: It is more likely that they quit because they do not like the questions rather than because they find the game too difficult.

*(Difference very clear from data distribution but nevertheless tested and verified through statistical hypothesis testing or A/B test)

Non-Expert players or Newbies (Players with less than 5 answers)

However, keep in mind that we have analyzed a dataset which could be just an horizontal slice of the full collection (e.g. 1 year of logs)

Unfortunately, we are not able to infer any further interesting information about the new players

- Understanding their first-game day would have helped us classifying them as quitters or new players
- Even leveraging the user_id order to better understand the players' experience, we did not notice any distribution skew between all the users and the least answering ones
  - A concentration of low-answering users with high values of user_id could have lead us to think that some of the low answering users had just joined the game

Without newbies

Correct answers per user:

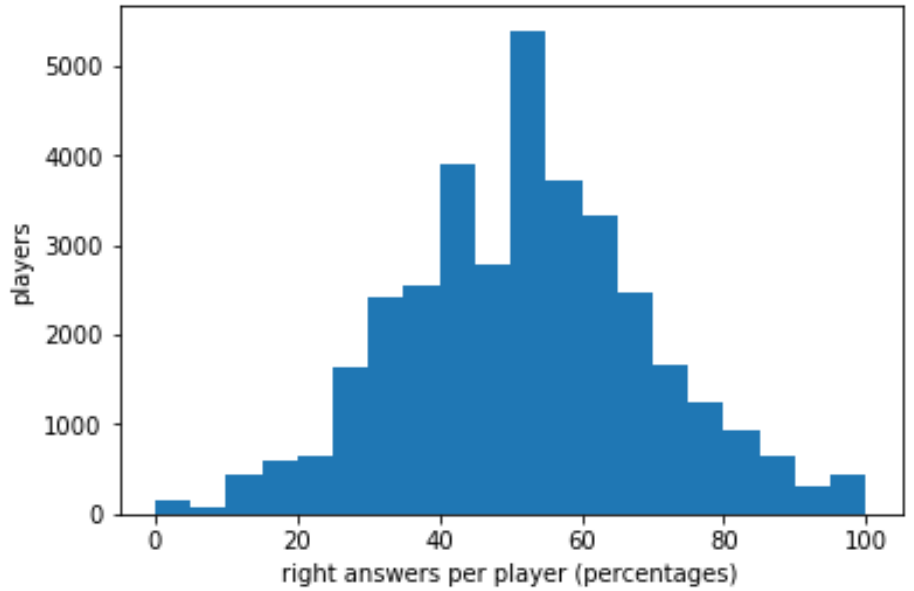
- Normal distribution around the average (51%), consistently less people with 100% of correctness

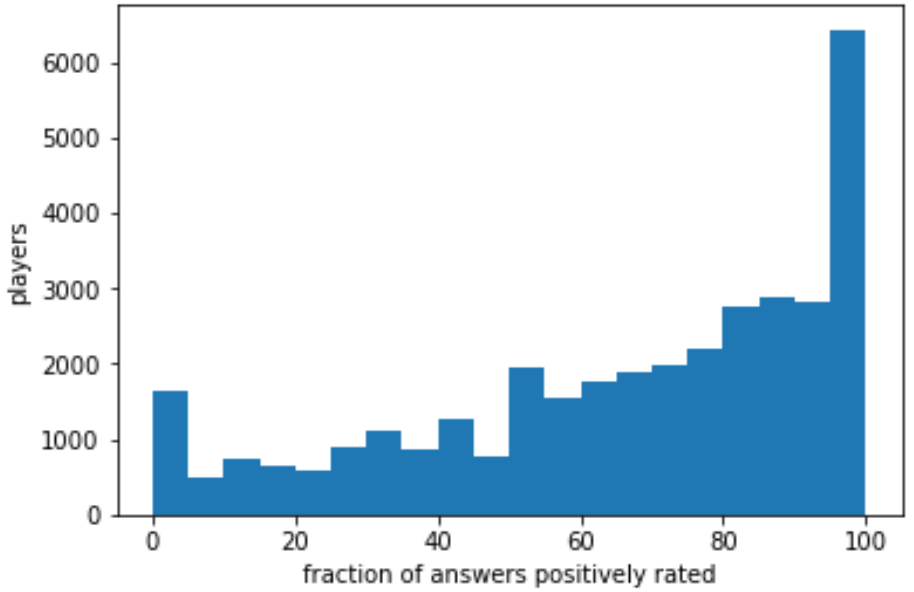
Positively rated questions per user:

- Still a lot of full positive o negative feedbacks but way more balanced.

In general, people do like the questions but there is a bunch of players who do not like any of them

- **Or, probably, a lot of people rate all the questions in the same way.**
- **Can we think about excluding the 100% or 0% positive guys in our statistics? Would it give us a more realistic environment?**





Without newbies

Correct answers per user:

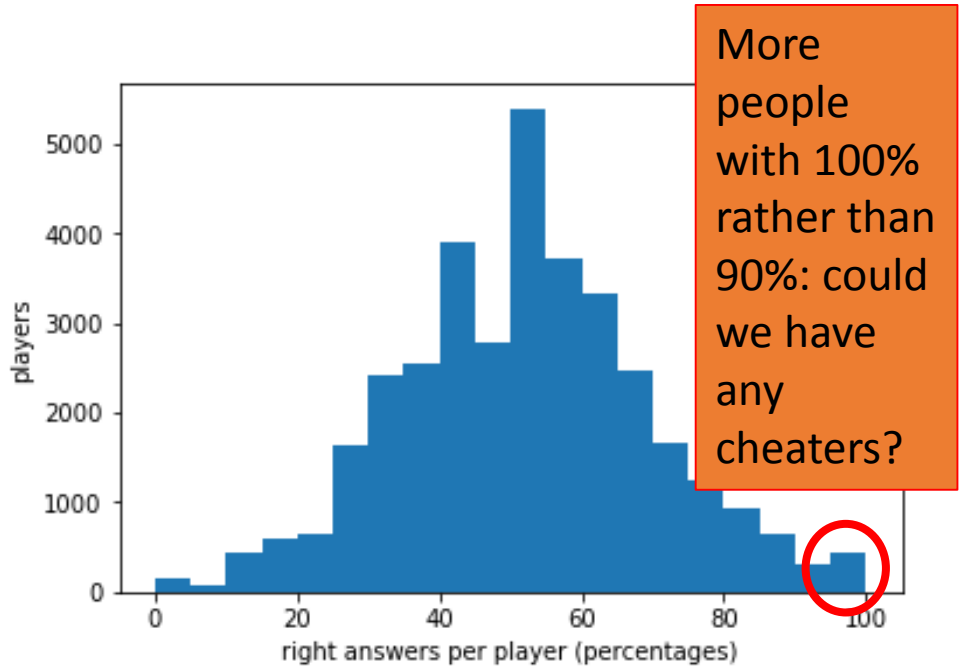
- Normal distribution around the average (51%), consistently less people with 100% of correctness

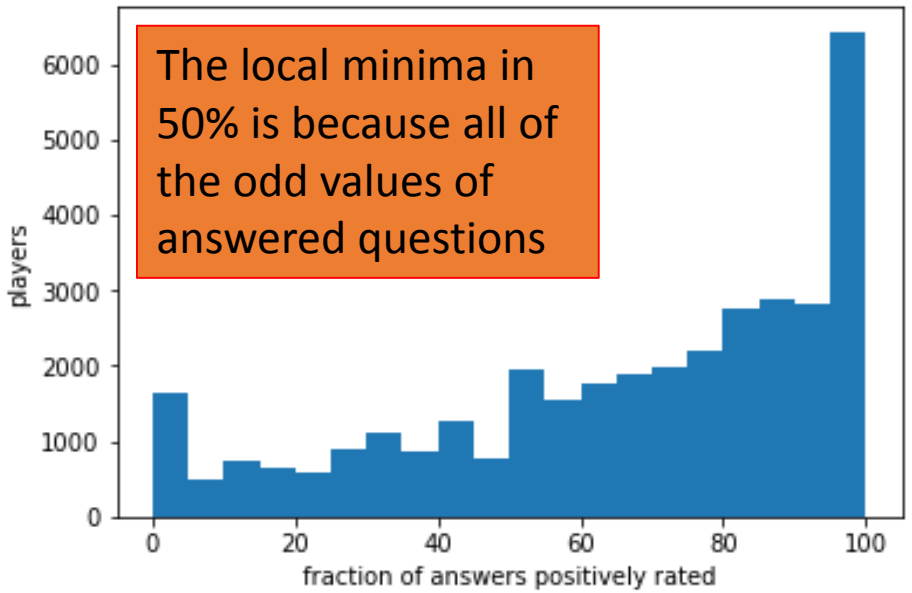
Positively rated questions per user:

- Still a lot of full positive o negative feedbacks but way more balanced.

In general, people do like the questions but there is a bunch of players who do not like any of them

- **Or, probably, a lot of people rate all the questions in the same way.**
- **Can we think about excluding the 100% or 0% positive guys in our statistics? Would it give us a more realistic environment?**





Take-aways (1)

- Trend to positively rate the questions for which the answer is known
- Image questions are much easier and more appreciated
- Text questions which assume an illustrated version are easier and more liked
- Categories do not have the same number of possible questions
- The most “selected” categories do not have more possible questions:
  - Higher probability of question repetition
- Most appreciated categories occur more
  - People have selected it because they like it more
  - Do they positively rate the question just because they like the topic?
- Most appreciated questions appear more frequently
  - The system proposed them or, simply, people appreciate more the questions of the categories they like (and select)

Take-aways (2)

- Probably, not all the questions of each games are rated
- 75% of the games and 80% of the questions played just by 30% of the users
- People do not stop playing because the game is too hard
  - They more likely quit because they don’t like the questions
- Early quitters (Less than 5 answers) are better players with respect to the others (+2% correctness)
  - At the same time, they are more “picky” with the question (-5% of positive rates)
- Somebody is probably rating all the questions in the same way (0% or 100% positivity)
- Is anybody cheating?

Potentially useful data

The analyzed dataset provided no detailed information that could deliver further interesting insights about users behavior

Users	
Geographical	Country or region-based trends
Temporal	Useful to identify quitters or newbies Rushing hours for the games
Personal	Sex and Age-based profiling
Games	
Temporal	Time to complete, time of the day, correlation between time to complete the question and the score
Personal	Similarity between users

What can we do with what we have?

We have seen that the question appreciation is a factor pushing players to keep playing

What do we have?

- History of the rates by each user

What we can do?

- Given the users «taste», we can propose questions that the users will probably appreciate:

Recommendation system given the user previous rates

User-rate recommendation system

User based (Collaborative Filtering)

Ratio: “We have similar interests, look at these questions I liked”

Content based

Ratio: These questions are often positively rated together. If you liked question A, maybe you would like question B

User-rate recommendation system

On the users:

Collaborative Filtering

Basic (very naïve explanation)

- Person A likes item 1, 2, 3
- Person B like 2, 3, 4

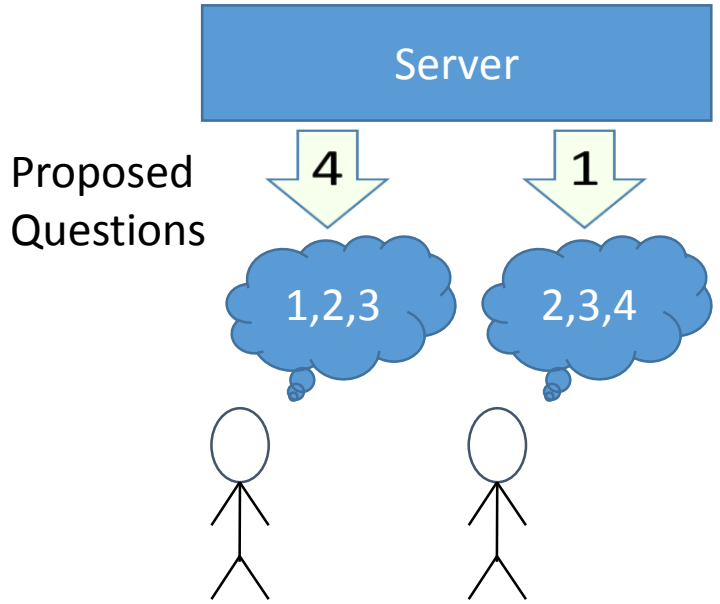
User-rate recommendation system

On the users:

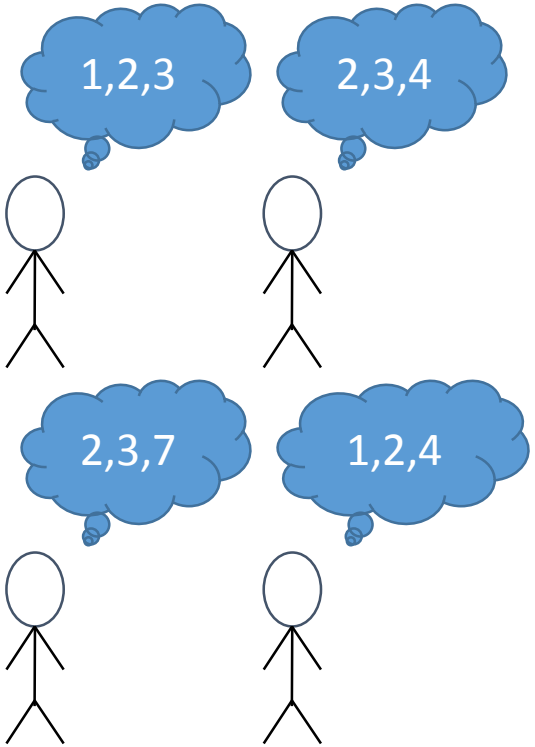
Collaborative Filtering

Basic (very naïve explanation)

- Person A likes item 1, 2, 3
- Person B like 2, 3, 4
- They have similar interests!
  - A should like item 4 and B should like item 1.



User-rate recommendation system



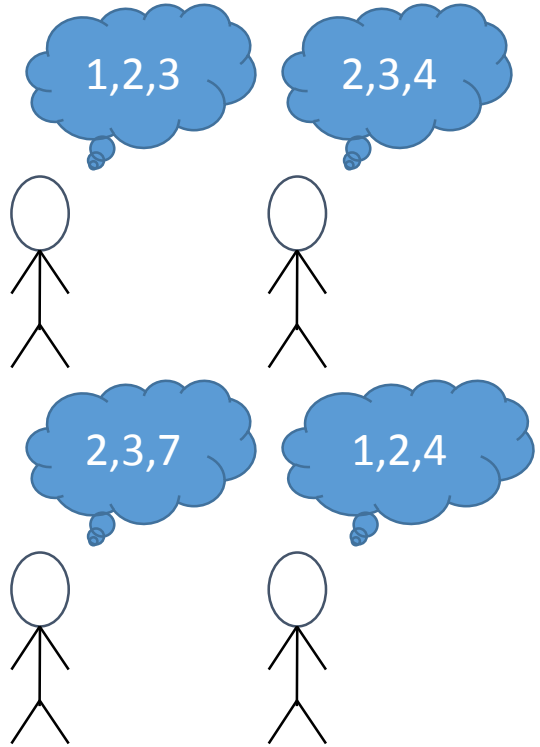
On the content:

Association Rules

Basic (very naïve explanation):

- Question 2 and Question 3 are often liked by the same players (Lot of users appreciate both 2 and 3)
- They must be “similar”!

User-rate recommendation system



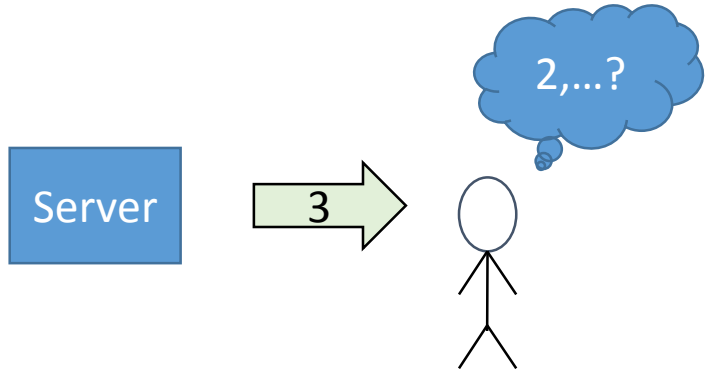
On the content:

Association Rules*

Basic (very naïve explanation):

- Question 2 and Question 3 are often liked by the same players (Lot of users appreciate both 2 and 3)
- They must be “similar”!
- Person A likes question 2:
  - He would probably appreciate question 3 as well!

*Very basic implementation in the sources



Thank you