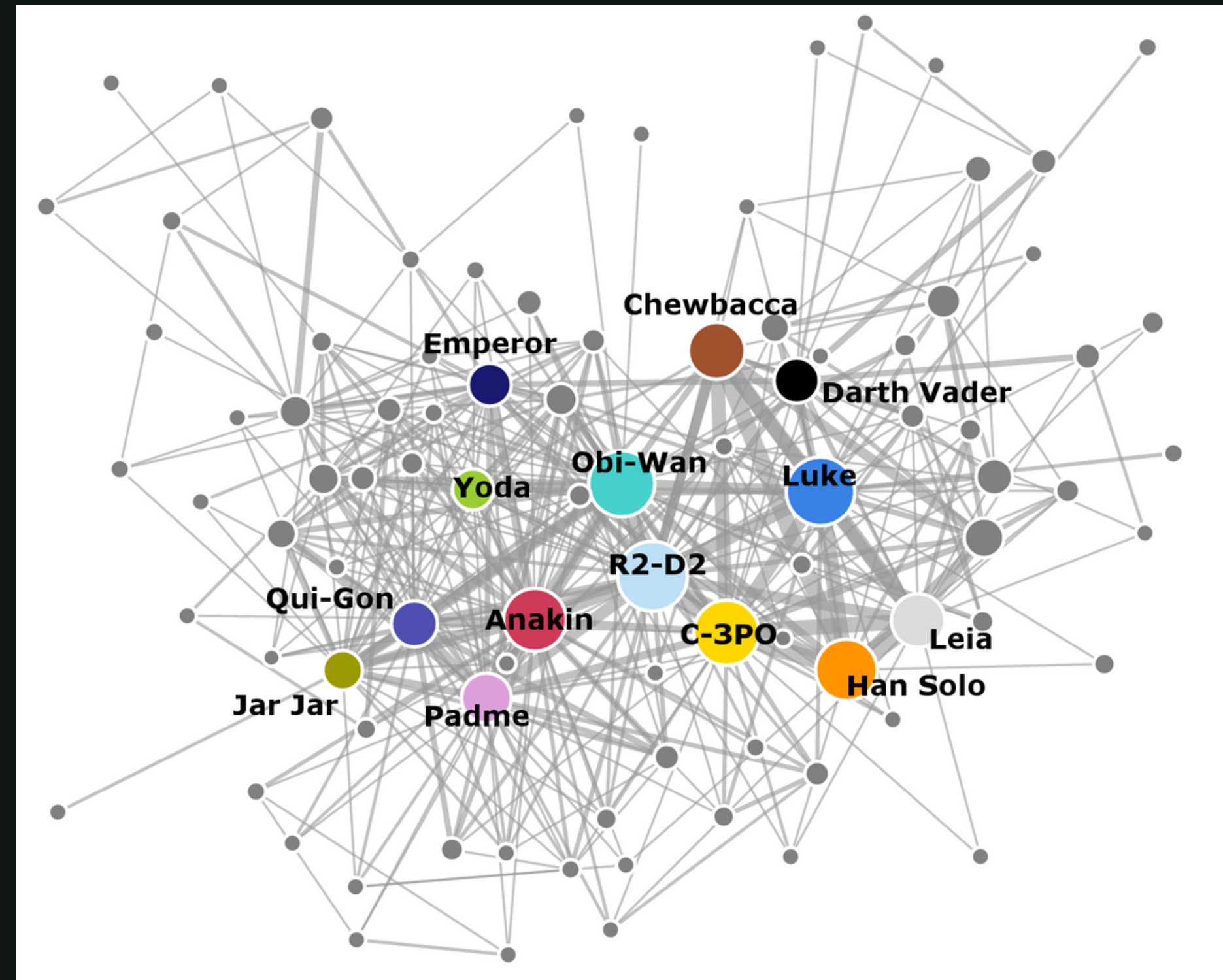


Star Wars Data Analysis

Analysis of data about Star Wars viewers using a data set posted by "*fivethirtyheight*" on github.





About the project :

This is an analysis of data about Star Wars viewers. I used a .csv file posted by "*fivethirtyheight*" on github ([here](#)).

The data set can be divided in two parts : specific questions about Star Wars and questions about the respondent.

I made this project in order to gain experience in data analysis with pandas and numpy libraries.

Step 1

Cleaning
the data



Step 2


Creation of the
algotrithms



Step 3

Analysis of
the outcome





First look at the data:

The sample:

1186 respondents, 935 of which saw the 6 movies and 551 consider themselves as fans.

The angle:

As the first Star Wars movie has been released in 1977, the age of the respondent is significant for the analysis. I also analysed the effect of their education.

The respondents:

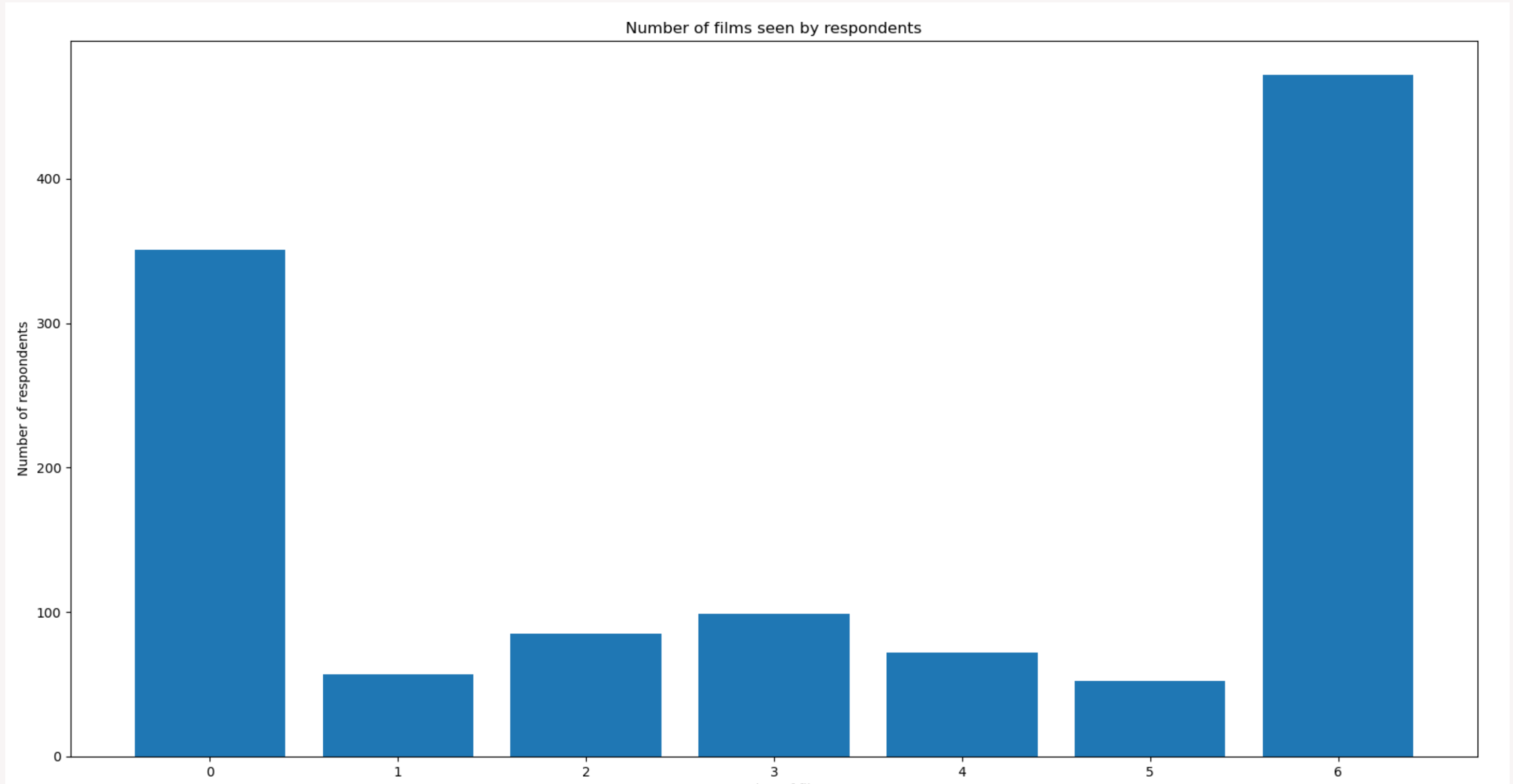
The profiles of the respondents are highly diversified. The data includes their age (from 18 to more than 64), their education, their income...

The focus:

One of the greatest question in the Star Wars universe is "*who shot first?*". It refers to a scene in the 4th movie that as been changed over the years (VHS, DVD...) where you can see either Han Solo or Greedo shooting first (see more [here](#)).

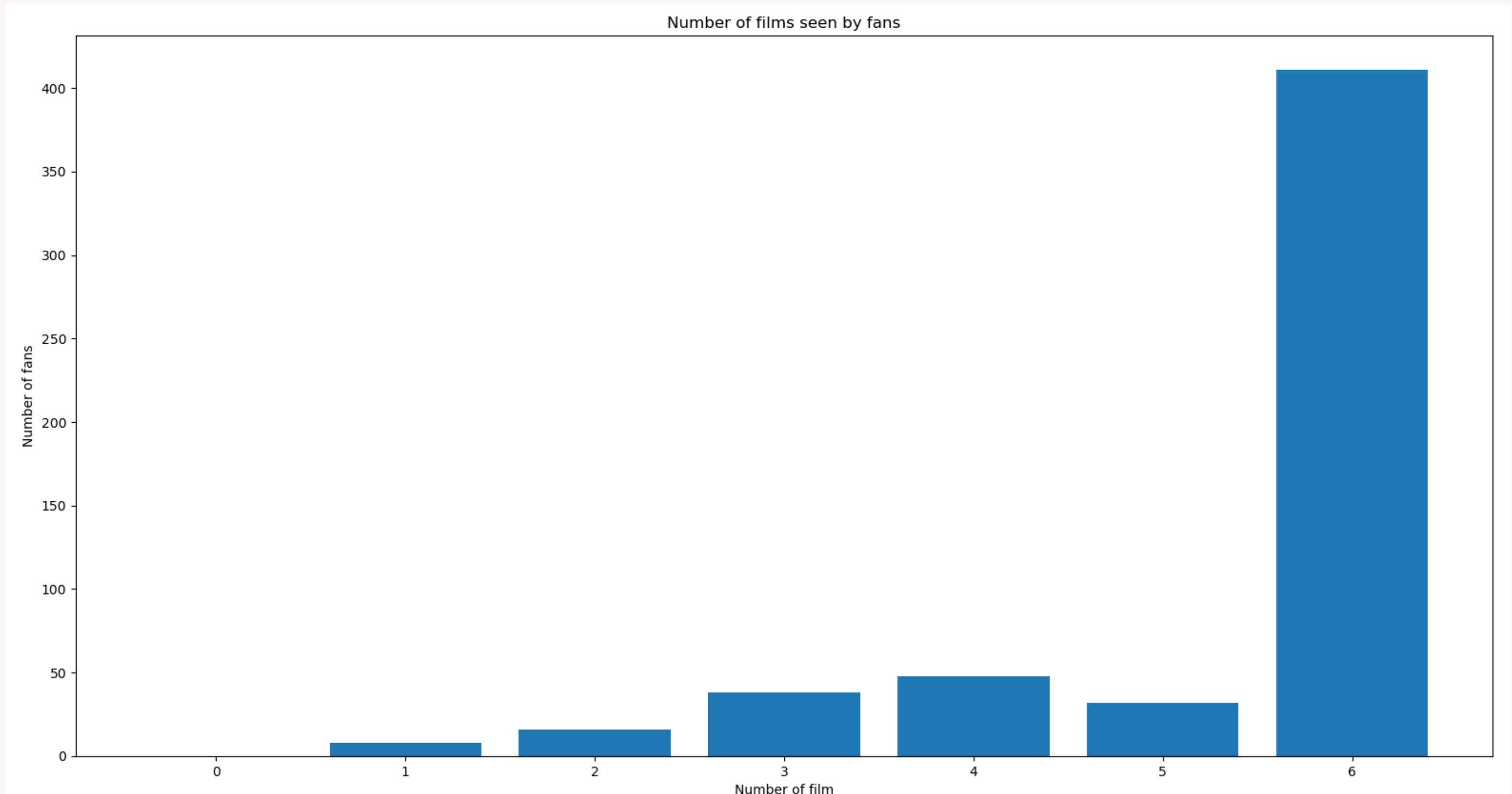
Data analysis of the respondents

Number of films seen by respondents



Data analysis of the fans

Number of films seen by allegedly fans





■ **High polarization**

Be it for the fans or the respondents, there is a high number of people who watched the 6 movies. It is also interesting to see that a large number of people answered even though they have never seen a Star Wars movie.

■ **False fans?**

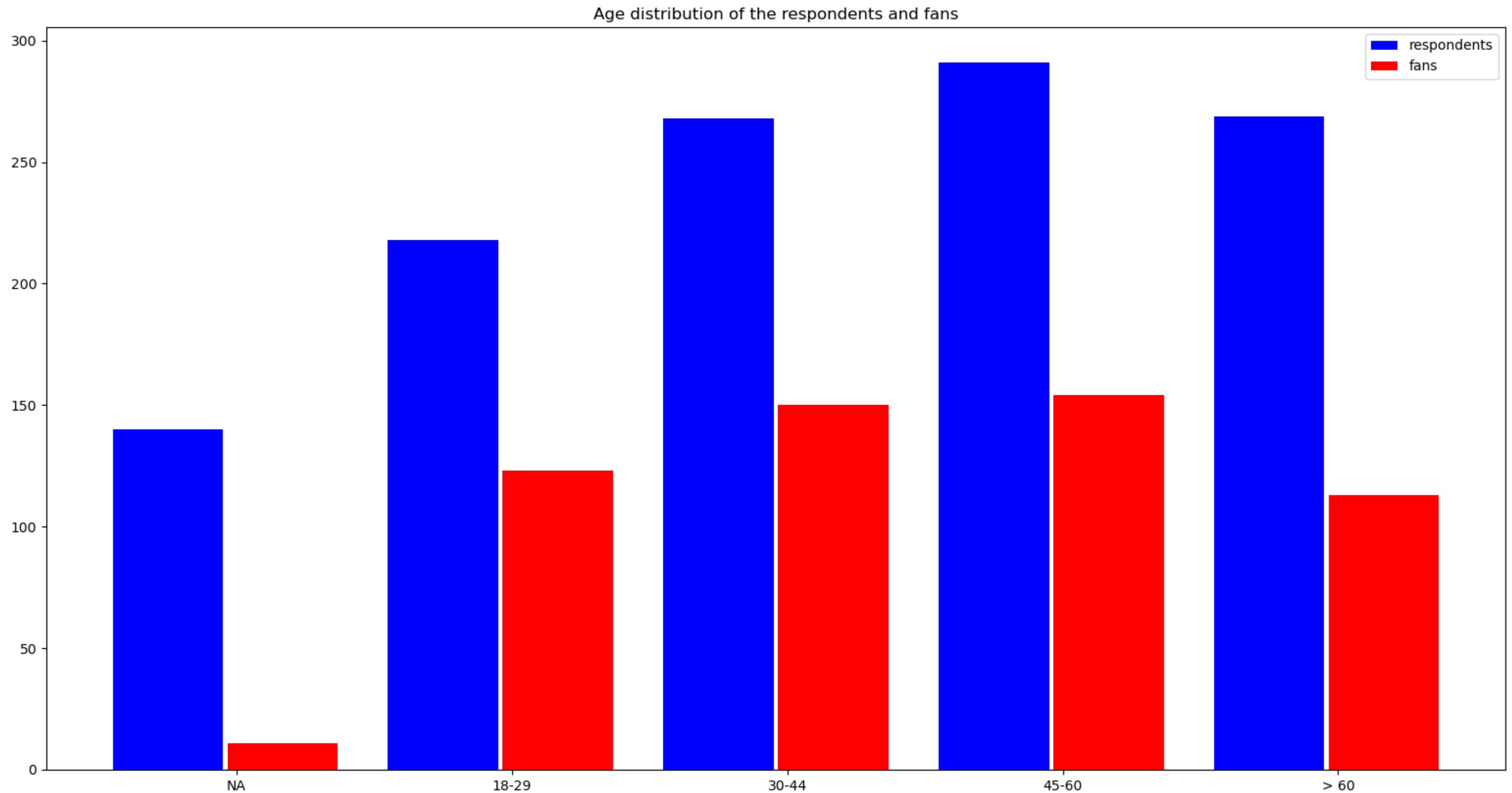
110 allegedly fans have never seen the 6 movies, what represents almost 20% of the fans (551).

■ **Interesting parallel**

Be it for the respondents or the fans, there is an interesting parallel between the curves that increases at first and drop for $x=5$. An explanation could be that viewers who saw 4 movies or more were interested in Star Wars and wanted to see them all.

Age distribution

Comparison of the age distribution of the respondents and the fans



Who shot first?

This is probably the question that divides the fans the most.

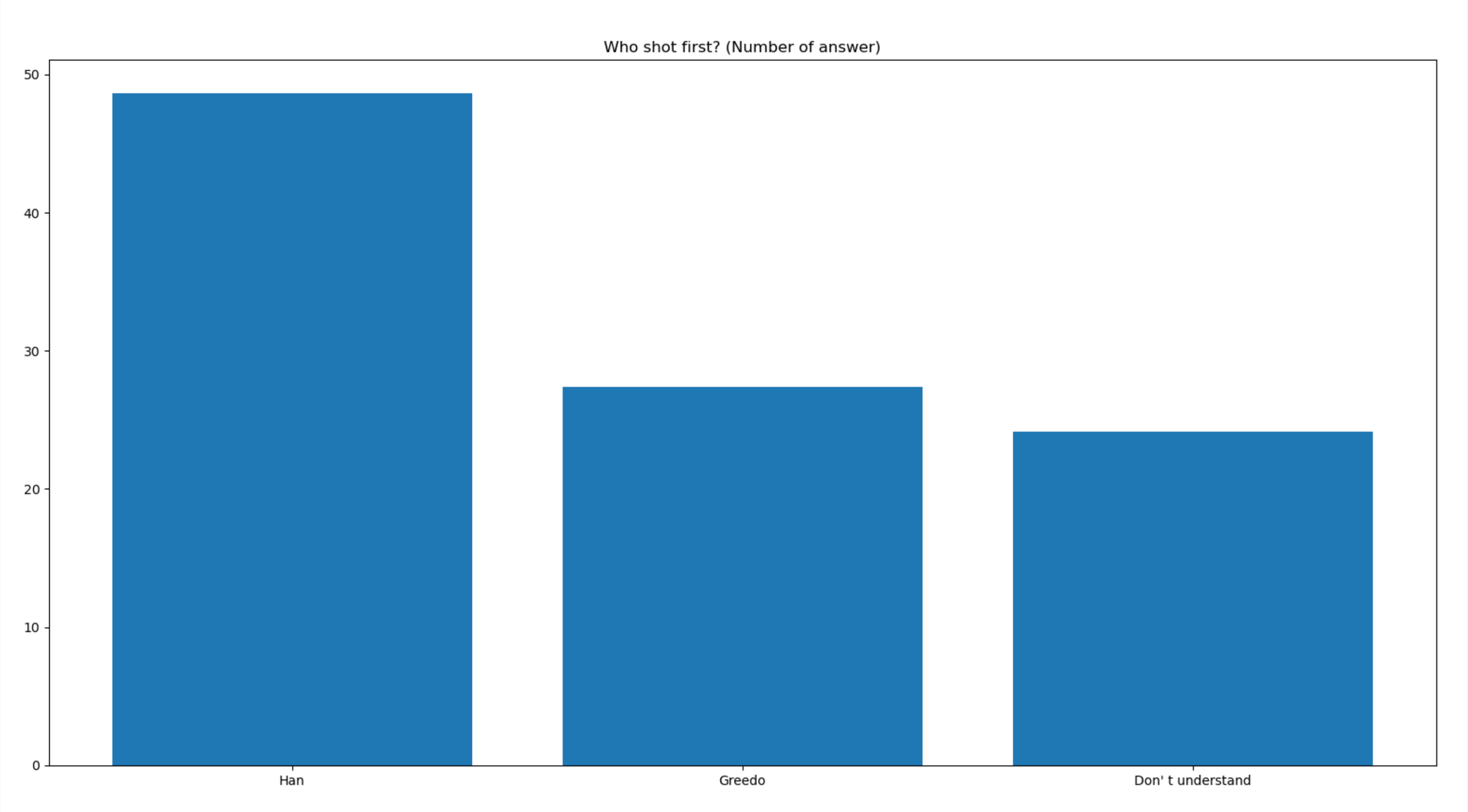
This division might appear within the poll.





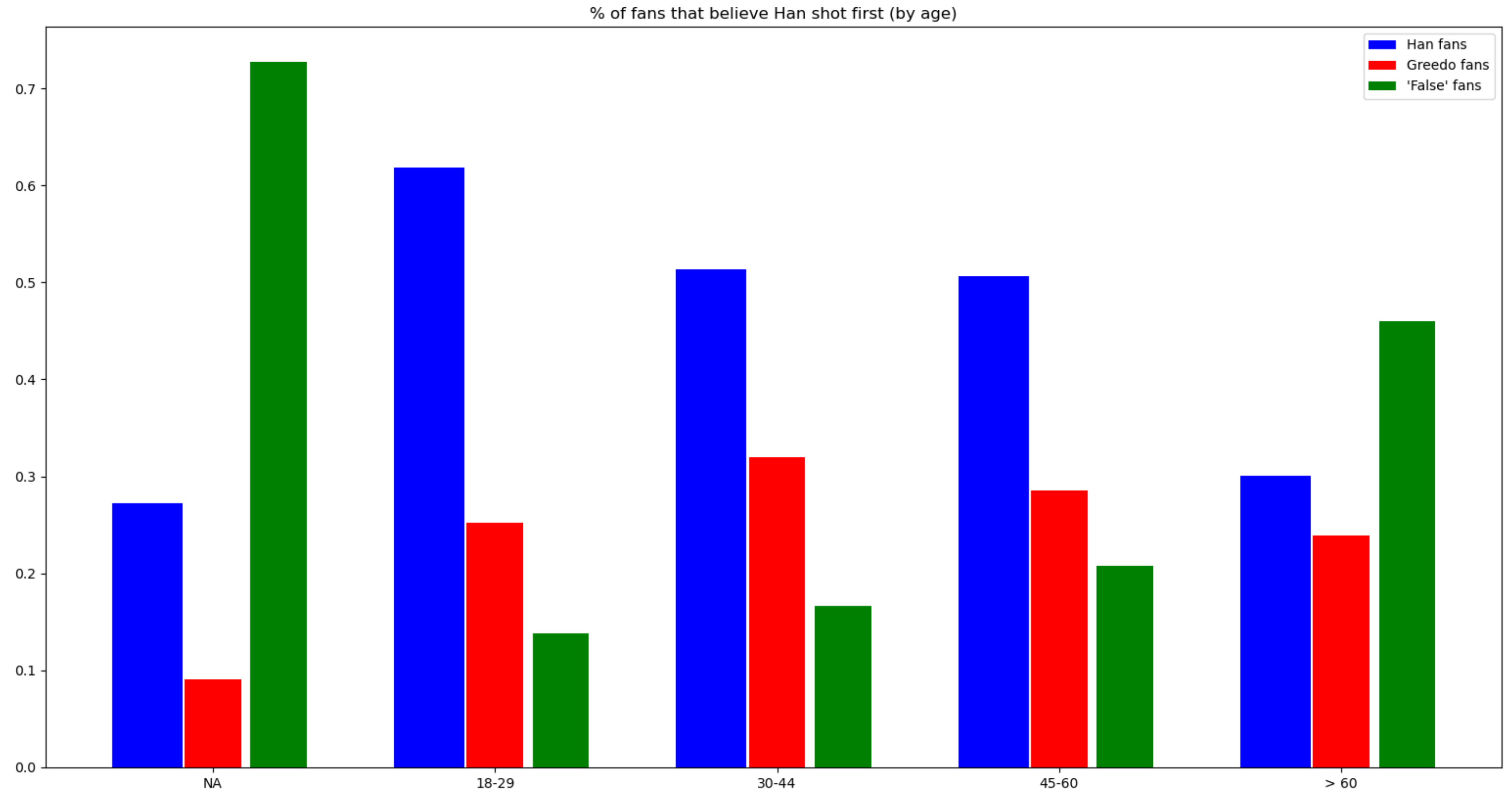
Who shot first?

Number of fans' answers by character



Who shot first?

Number of fans' answers by character and by age



Who shot first?

Number of fans' answers by character and by education

