# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic) CIS5200 Term Project Tutorial

#### Authors: [Yash](https://www.linkedin.com/in/tom-cruise) Choksi, Sweta Tripathi, Karolina Stachurska, Sowmya Mareedu, Sindhura Vemuri

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 05/02/2020

**Lab Tutorial**

Yash Choksi ([ychoksi@calstatela.edu](mailto:ychoksi@calstatela.edu))

Sweta Tripathi ([stripat@calstatela.edu](mailto:stripat@calstatela.edu))

Karolina Stachurska([kstachu@calstatela.edu](mailto:kstachu@calstatela.edu))

Sindhura Vemuri ([svemuri@calstatela.edu](mailto:svemuri@calstatela.edu))

Sowmya Mareedu ([smareed@calstatela.edu](mailto:smareed@calstatela.edu))

05/02/2020

**Reddit comments analysis**

**Platform Spec**

* Required space: 25 GB
* CPU Memory size: 2.20 GHz
* # Nodes: 3 nodes
* Hadoop cluster

Step 1: Get data from Kaggle

1. Found dataset from Kaggle.com

2. Dataset link: <https://www.kaggle.com/reddit/reddit-comments-may-2015>

3. Original size of data set is 30 GB, because of compression they reduced it to 16 GB. Dataset format is sql table so sqlite. So converted data from sqlite to csv using sql workbench. Data set has more than 15 columns and 21199015 rows.

4. Uploded data to Hadoop cluster using scp command: scp ./May2015.csv [ychoksi@129.150.71.75](mailto:ychoksi@129.150.71.75):~/

5. Uploded to hdfs using put command.

6. Created external hive table and get data from csv file:

**CREATE EXTERNAL TABLE if not exists lookup ( created\_utc BIGINT,ups BIGINT,subreddit\_id String, link\_id String,name String, score\_hidden BIGINT, author\_flair\_css\_class String, author\_flair\_text String, subreddit String, id String ,removal\_reason String,gilded BIGINT,downs BIGINT, archived BIGINT,author String,score BIGINT,retrieved\_on BIGINT,body String, distinguished String,edited BIGINT,controversiality BIGINT,parent\_id String )**

**ROW FORMAT DELIMITED FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE LOCATION '/user/ychoksi/check/';**

A picture containing bottle, sitting, white

Description automatically generated

7. As the table is created we can check size of database: select count(\*) from lookup;

A picture containing clock, drawing

Description automatically generated

8. All column name with data types: describe lookup;

A close up of text on a black background

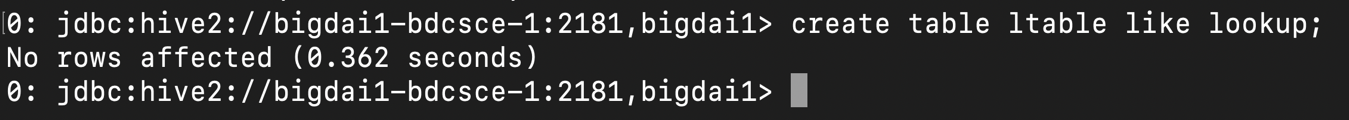
Description automatically generated

9. Two most important columns are author and name. If you go and check structure of reddit you can see that author is the person who asked the actual question, and name means person who made the comments. Now, to ensure privacy they **anonymized** data of name column. But they kept original form of author data.

As part of data cleaning we are removing all rows which do not have any value for name:

We are creating new table where author name NULL will be removed, table name is lookup:

Create table ltable like lookup;



A screenshot of a cell phone

Description automatically generated

Now insert data in newly created table which do not have null values:

Insert into ltable from lookup where author is not null;

After data cleaning our number of rows will be:

A picture containing drawing

Description automatically generated

10. Top 10 authors who are asking most questions:

select author, count(\*) as total from fcleaned where author != "[deleted]" and author != "" and author!="0" group by author order by total desc limit 10;

A close up of text on a black background

Description automatically generated

11. Which authors asked longest questions so we can able to know who is asking how long questions:

A close up of text on a black background

Description automatically generated

12. There is one column named as utc and which is universal time at which comment generated. Let’s see top 10 times at which most of the comments are generated:

Select created\_utc, count(\*) as total from ltable group by created\_utc order by total desc limit 10;

A close up of text on a black background

Description automatically generated

UTC means coordinated universal time

To convert from this strange looking time strings there is tool using programming you can convert to local time. Here is the link:

<https://stackoverflow.com/questions/179940/convert-utc-gmt-time-to-local-time>

13. Subreddit categories type of channels or community they want to target. Like sports or many others:

select subreddit, count(\*) as total from fcleaned group by subreddit order by total desc limit 10;

A close up of text on a black background

Description automatically generated

Let’s explore more this community based channels:

14. Let’s see top 10 authors who most asked questions related to nfl:

select author, count(subreddit) as nfl from fcleaned where subreddit = "nfl" and author != "[deleted]" group by author order by nfl desc limit 10;

A close up of text on a black background

Description automatically generated

As we can see in subreddit channels Dota2 is very popular game so llet’s see how many users comment about that game:

select count(\*) as Dota2\_count from fcleaned where subreddit = "DotA2";

A picture containing drawing

Description automatically generated

15. Top authors who received comments regrading dota2:

select author, count(subreddit) as dota2\_count from fcleaned where subreddit = "DotA2" and author != "[deleted]" group by author order by dota2\_count desc limit 10;

A close up of text on a black background

Description automatically generated

16. Relationship between controversy points and number of questions asked:

select author, SUM(Controversiality) as total\_contro, COUNT(\*) as sum from fcleaned where subreddit="politics" group by author order by total\_contro desc limit 100;

A close up of text on a black background

Description automatically generated

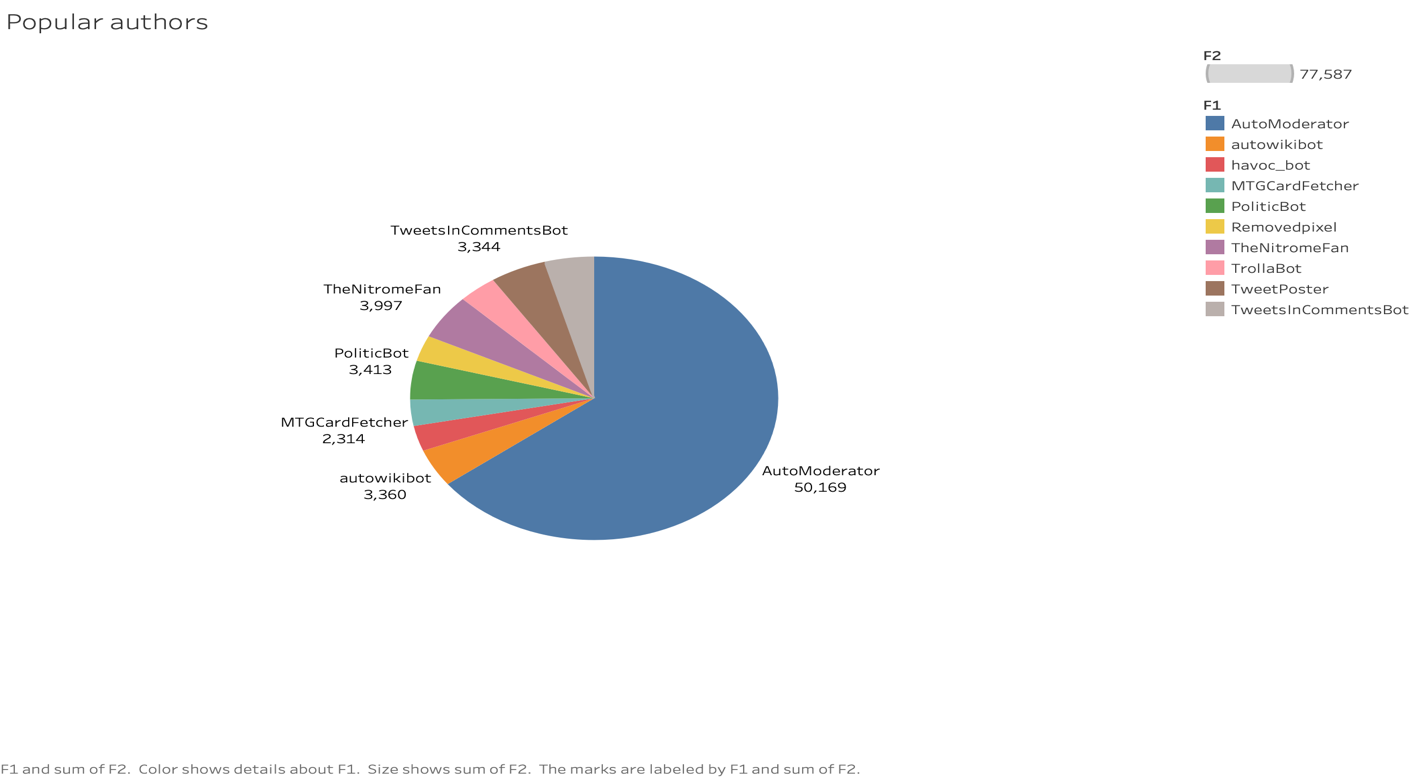
17. Edit rates to controversial posts:

select author, sum(controversiality) as contro, sum(edited) as edit from fcleaned where author!="0" and author!="[deleted]" and author != "" group by author order by contro desc limit 50;

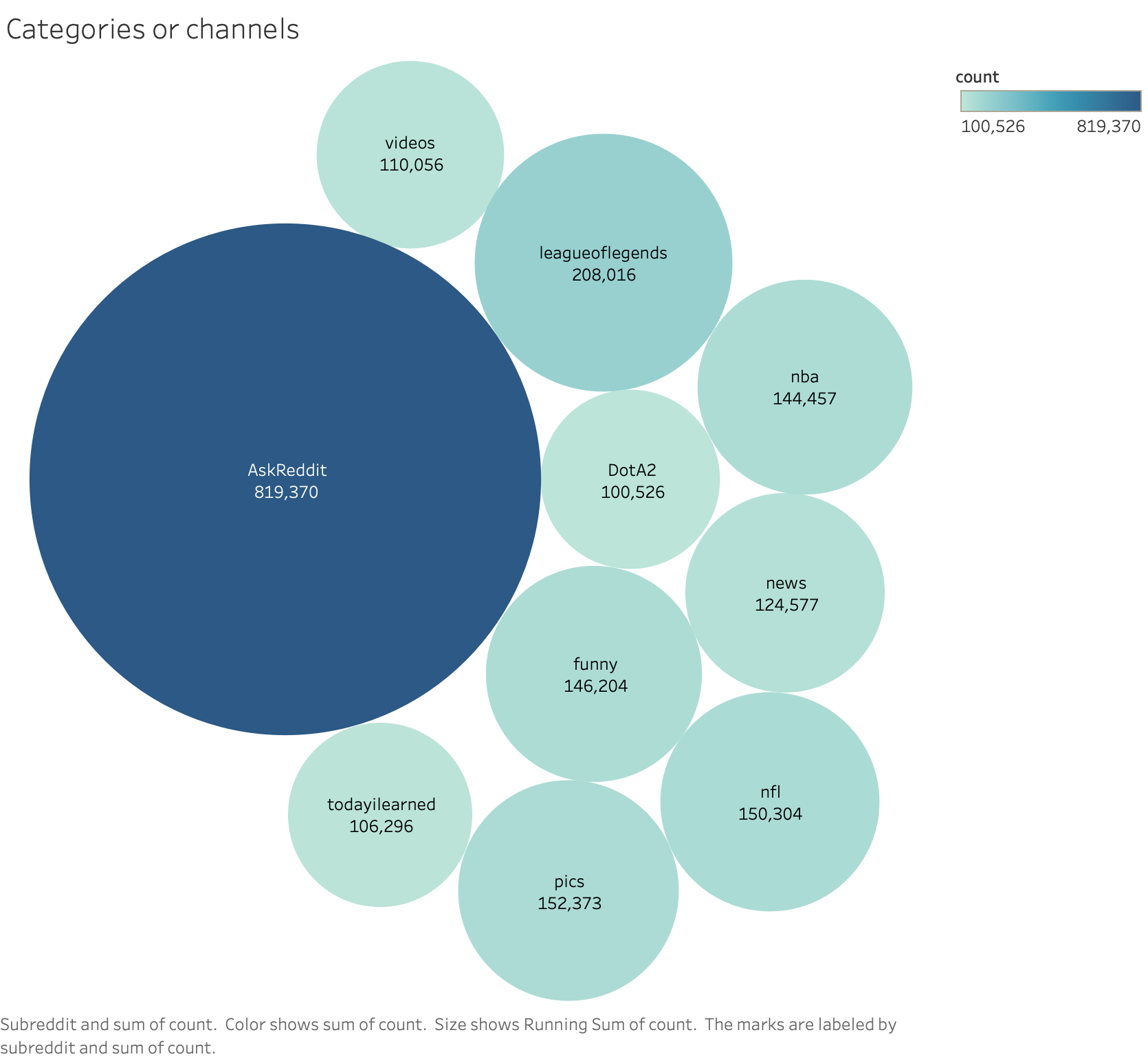
To get this data efficiently we have converted whole different small tables and then downloaded with scp command. And by doing that we can download data and can perform visualization.

Visualizations:

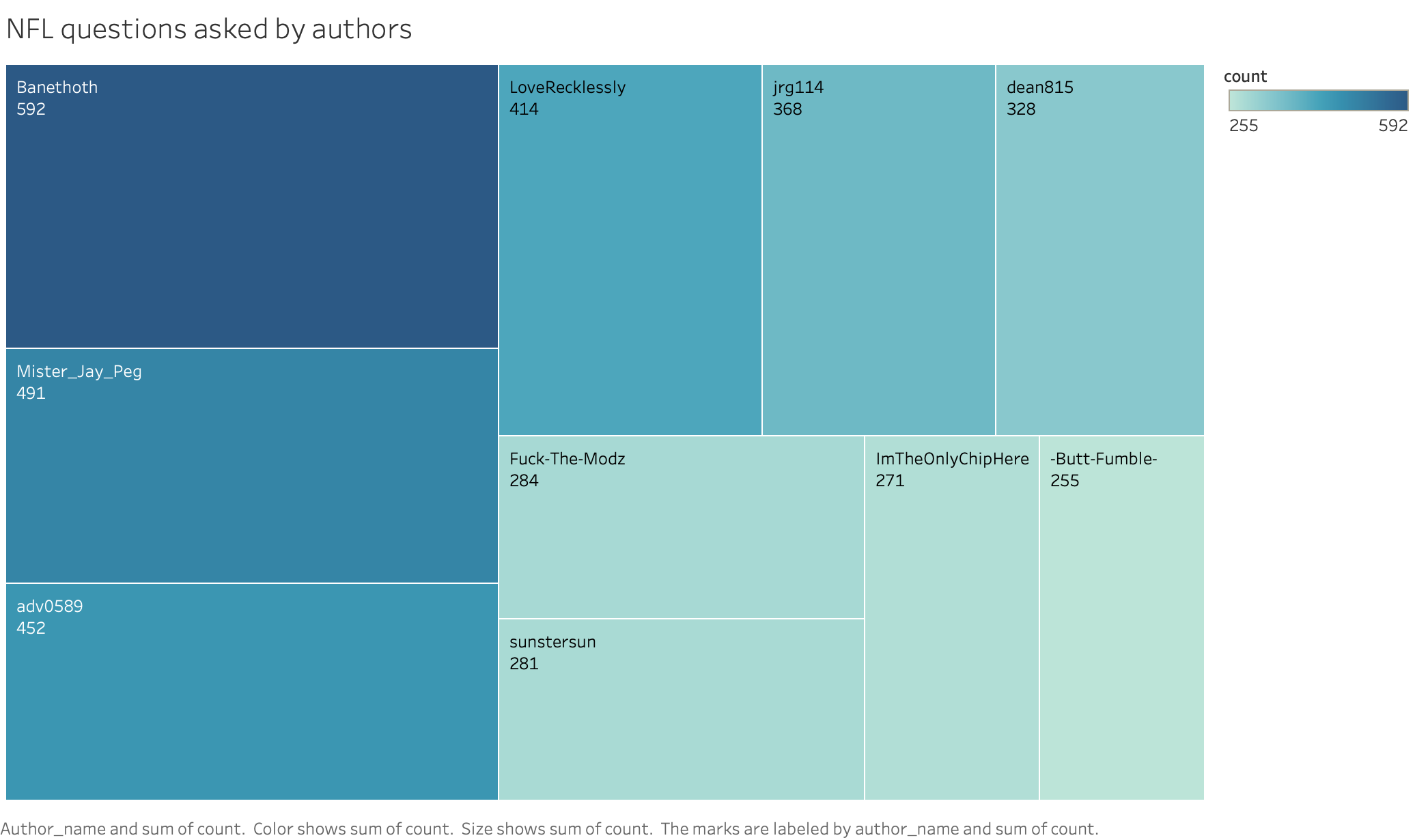
We have used tableau for our visualization work.



In the above figure the top 10 authors who asked most questions. The data is labelled with author name and how many questions he posted.



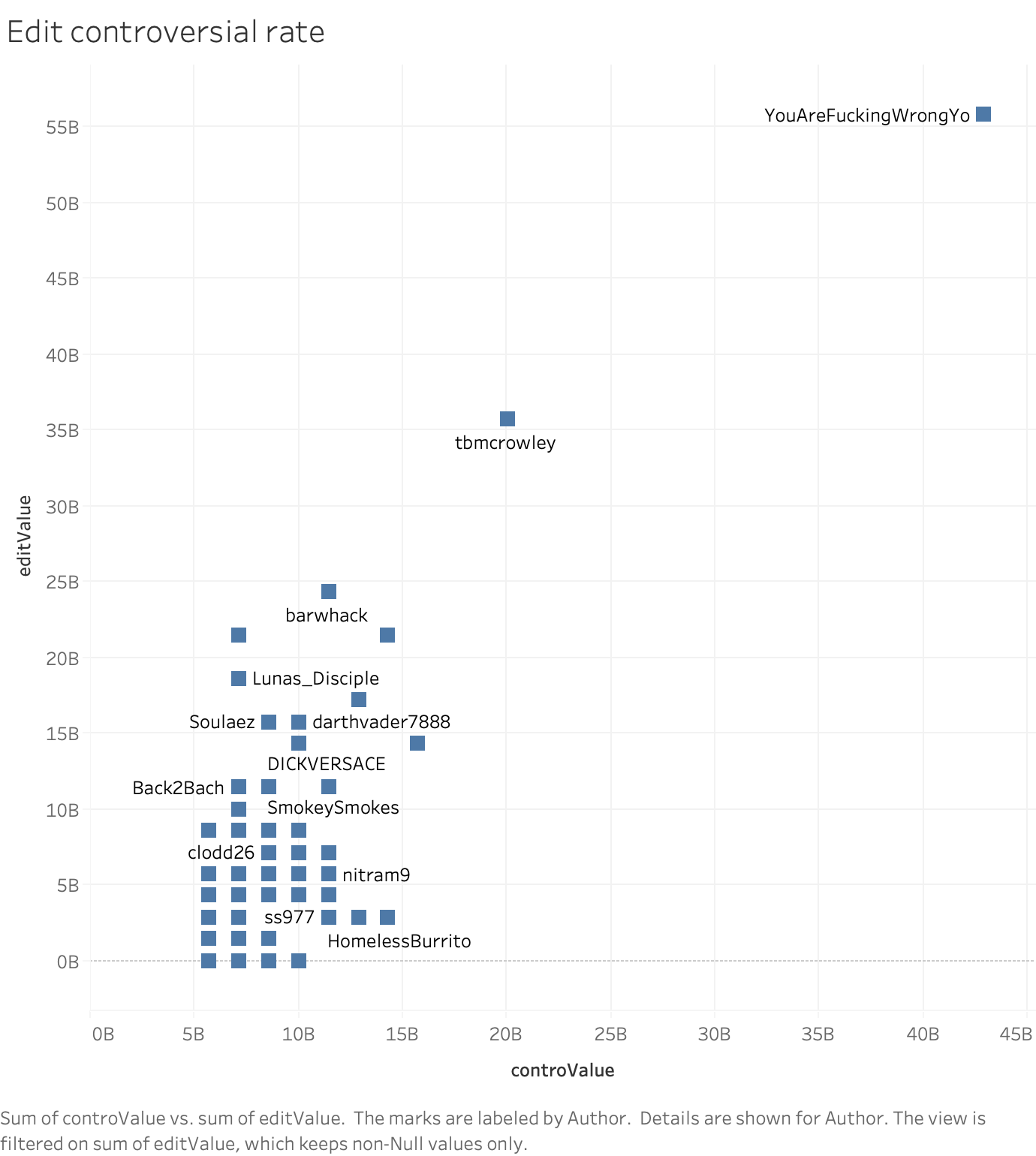
There are so many different channels but out of those amny channels these are top 10 channels.



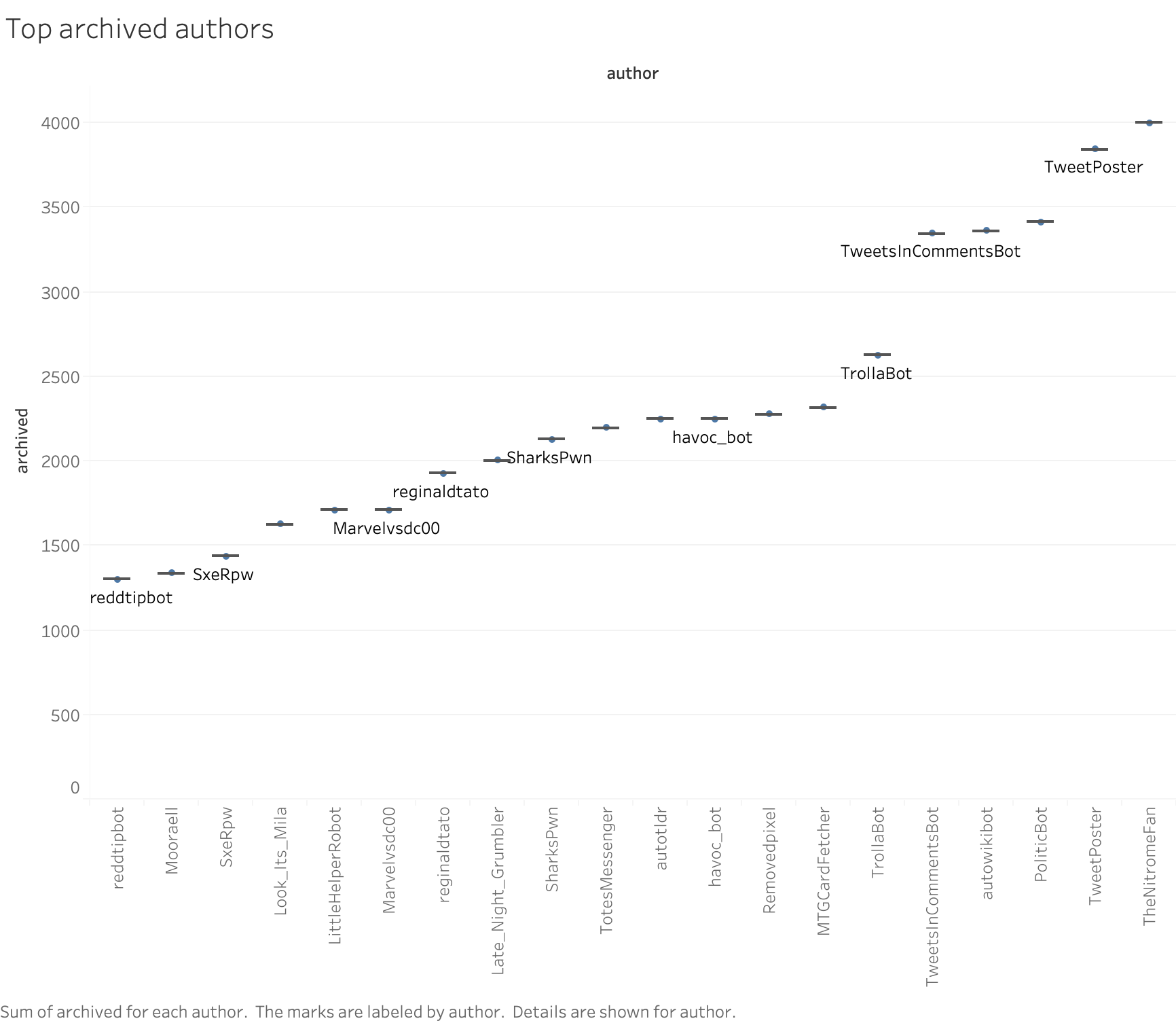
One of the important channel is NFL(National Football League) and these are top 10 authors who asked most of the questions. Data is labelled with names and how many times each person asked question.

A group of people sitting at a beach

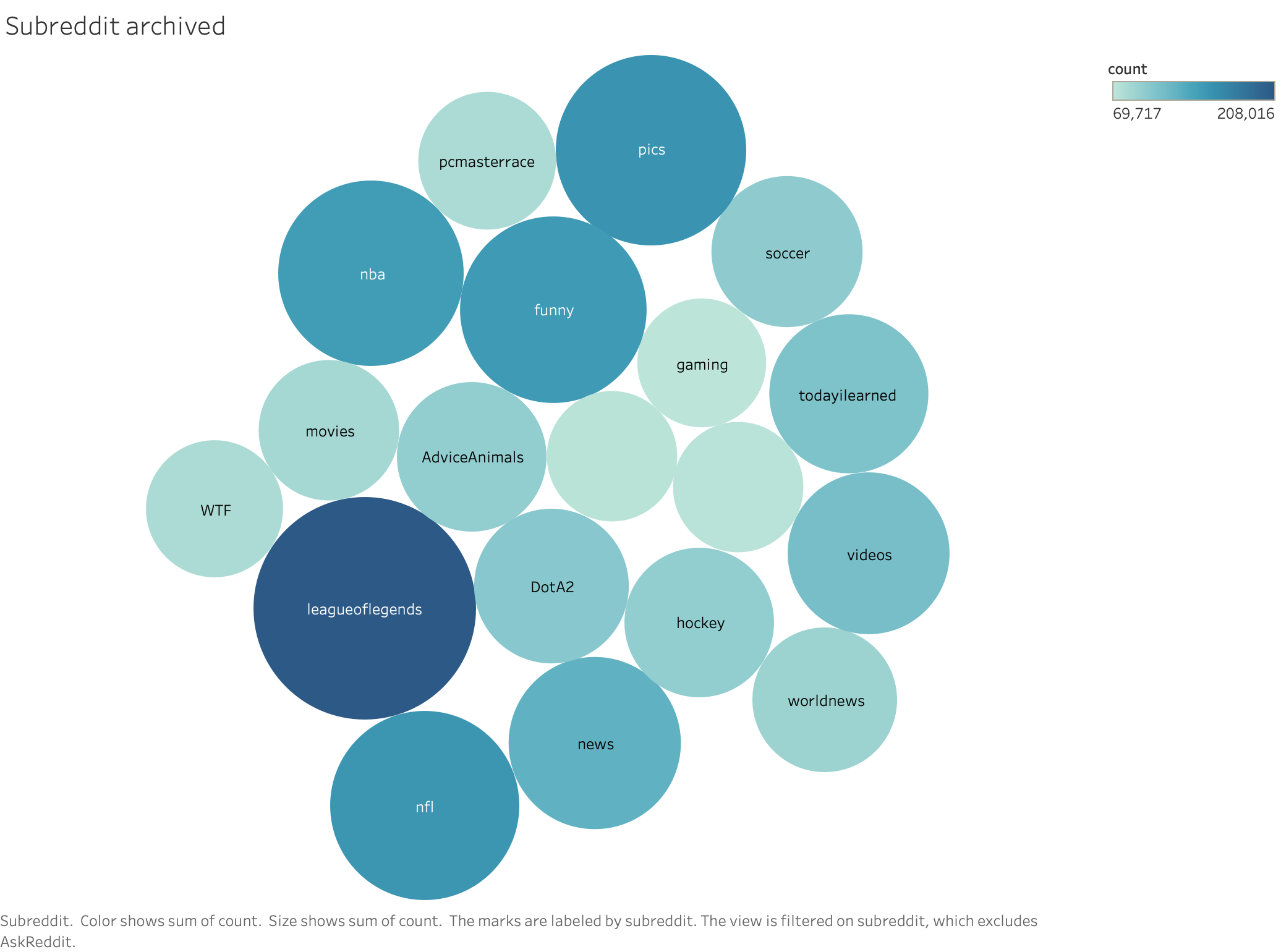
Description automatically generatedAs we can see above that number of questions and controversial points has no relation in general, so by keeping that in reference wee can say that some questions are so damaging and controversial that even few questions can storm whole social network.



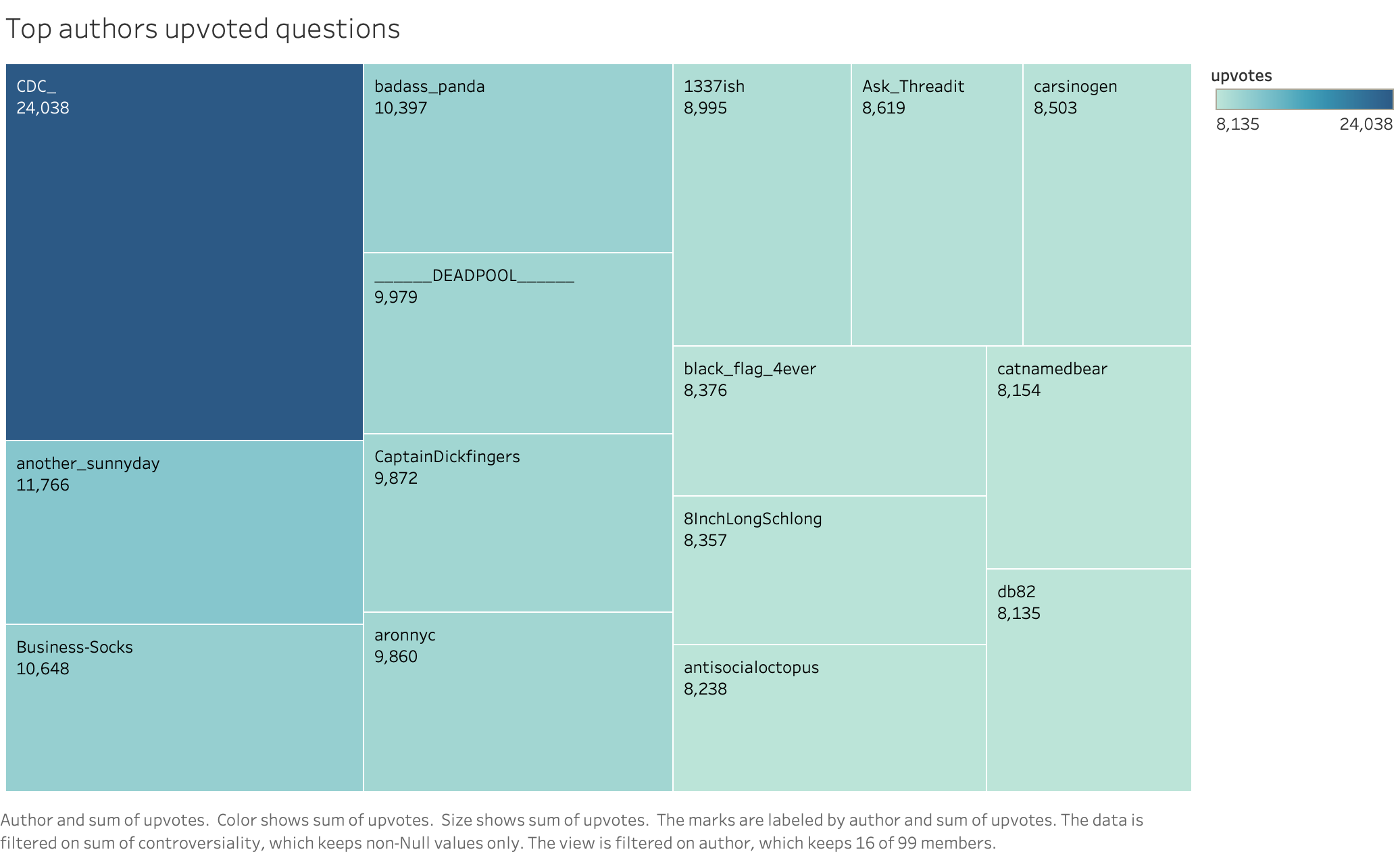
As we can see above that as the edit rate increases the controversial that comments gets more and more. Because there is policy from reedit that if you can’t improve your comments and don’t make it less controversial than they will block it. So, naturally edit rate is higher for such commentns.

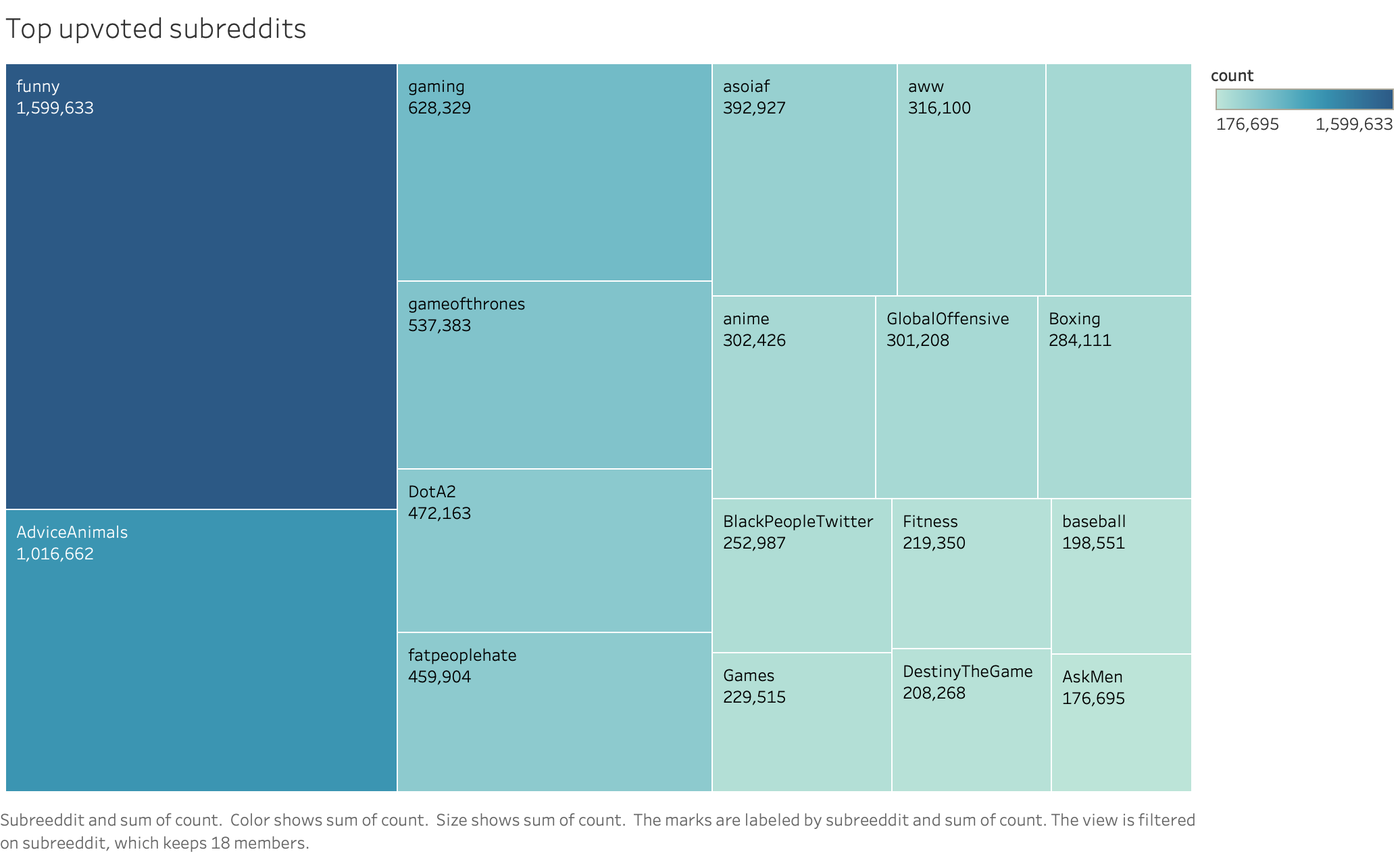


The above figure is for top 20 archived authors. Archived means when commenter gave the final fact based answer and there is nothing more can be done in that case. So, it’s better idea to see which author’s question is satisfactory answered.



As the authors we have to see for sub redditors who have most archive rate this is very important because we have to see which redditors are answered properly.



This chart is for top authors who are upvoted all the time high. Each author above here is counted on the basiscs of total number of upvote they received over the time. As the authors received upvotes each sub redditors also received upvotes and here in this way funny is always higher than any other thing.

**This is end of this tutorial.**