

Project Report





MTH208A - Data Science Lab I



Analysis of the Network of Question-and-Answer (QnA) websites on topics in varied fields

18 th November, 2022

By Group No. 13:

-  Aritra Basak (221282)
-  Satmeet Singh (200890)
-  Vala Yash (211142)
-  Dasari Charithambika (210302)

Under the guidance of Dr. Dootika Vats

Acknowledgement

We, the students of group-13 would like to express our profound gratitude towards *Dr. Dootika Vats*, our academic and project instructor for MTH208A (Data Science Lab I), for her guidance and constant supervision throughout the process and providing creative ideas and necessary information regarding the project which led to the completion of this project.

It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course lecture.

Introduction

Let's ask a basic question. **‘What do we do when we have any queries?’** The query may relate to any topic, like education, job, extra curricular activities, travelling, movie etc. We may have some idea regarding the answer of our query. But, to be confident we always prefer to recheck our knowledge from somewhere, that may be from the person nearby or from ‘Google’. If we don't know anything about the query then we definitely try to know the answer from somewhere. Now a days internet is available in every hands. So, why should bother someone, when we can know the answer within just a few clicks. Then, We just search in google and answers of our questions are easily available from different websites. You may have a very stupid question or very technical question. Google always have some answers irrespective of whatever it is.

We did the same thing while we thought of our course project and landed upon those technical sites and amidst we were going through it. We thought why cannot we just scrap this site over where we are searching for our answers. So, we decided to scrap those sites. In most of the times the answers are available on the question and answer (QnA) websites like *Quora*, *Stack-Exchange* etc.

Now what are these ‘Network of QnA websites’? Let's know about three leading QnA websites briefly.



- **Quora** -> Quora is a social question-and-answer website. Here, users can help each other by answering each others' questions on a variety of topics. Quora was launched in 2009. According to wikipedia "as of 2020, the website was visited by 300 million users a month".
- **Stack-Exchange** -> Stack-Exchange is itself a network of question-and-answer (Q&A) websites, where more technical questions and answers are discussed by the users. It was founded in 2009. According to wikipedia "Stack Exchange publicly launched in January 2011 with 33 Web sites; it had 1.5 million users at the time, and it included advertising. At that time, it was compared to Quora, founded in 2009, which similarly specializes in expert answers".
- **Stack-overflow** -> Stack-overflow is a site of the Stack-Exchange network, which mainly focuses on programming related topics. It was launched in 2008. According to wikipedia, "As of March 2021 Stack Overflow has over 14 million registered users and has received over 21 million questions and 31 million answers".

Data Description

- **Primary Data:** To analyse the performance of the Network of question-and-answer(QnA) websites, we have collected our primary data from three leading QnA websites named Quora, Stack-exchange and Stack-overflow, on some relevant topics in various fields. We have collected top 10 queries asked by people along with their number of followers and number of answers for each selected topic from quora and top 15 queries asked along with their number of votes, number of answers, number of views, asked year and modified year for each selected topic from Stack-exchange and Stack-overflow each.
- **Obtaining the Data:**
 - i. **Data collection procedure:** The data was collected from the official websites of Quora, Stack-exchange and Stack-overflow via web scraping in R programming language. Libraries like tidyverse and rvest and functions like `html_elements()` and `html_attr()` were used to scrape the required data from the webpage.

Links to the official Websites ->

- *Quora.com*
- *Stack-Exchange.com*
- *Stack-Overflow.com*

ii. **Challenges faced to obtain the data:**

1. Html code of Quora written in Javascript: The source code of the 'Quora' webpage is written in Javascript. So, the `html_text()` function was returning a large string of text from where getting the required amount of text was very troublesome.
2. Very few number of datapoints: If someone search for a particular topic (say, 'technology') in the selected QnA websites then they can see a numerous number of questions which goes on increasing as he/she scroll down through the webpage. But at the time of scraping questions we could scrape only top 10 questions for eaach topic from 'Quora' and top 15 questions for each topic from 'Stack-Exchange' and 'Stack-overflow' each which were very less number of datapoints to analyse.

iii. **Steps taken to tackle the challenges:**

1. Using the library 'stringr': By observing that the queries asked in 'Quora' were of a similar pattern, the functions `str_match_all()` and `str_locate()` were used. Moreover, the `gsub()` and `substr()` functions were used to extract the final list of required questions.
2. Number of datapoints: To make the analysis stronger, a comparatively large number of relevent topics were considered so that the total number of datapoints for each QnA platform gets increased despite of the less number of datapoints for each topic.

- **Relevent Features of the Data:**

- There are 170 data points in quora data set and 225 data points each in stack-exchange and stack-overflow data set.
- Quora data set contains questions related to 17 topics named,

```

[1] "technology"      "movie"      "health"
[4] "food"             "science"    "music"
[7] "visiting and travel" "sports"     "fashion and lifestyle"
[10] "politics"         "python"     "c++"
[13] "java"             "fortran"    "matlab"
[16] "IIT Kanpur statistics" "IIT Bombay statistics"

```

and stack-exchange and stack-overflow contains questions related to 15 topics, same as quora data set except two topics “IIT Kanpur statistics” and “IIT Bombay statistics”.

- Quora data set contains 4 columns named ‘questions’, ‘topics’, ‘followers’ and ‘answers’.

-> Detailed Description of the columns

1. questions: It contains the questions asked from the relevant topics in quora.
2. topics: It contains the relevant topic where is the question from.
3. followers: It contains the number of people who follow a particular question.
4. answers: It contains the number of answers available for each question.

- Stack-exchange and Stack-overflow each contains 7 columns named ‘questions’, ‘topics’, ‘votes’, ‘answers’, ‘views’, ‘asked year’ and ‘modified year’.

-> Detailed Description of the columns

1. Description of the ‘questions’ and ‘topic’ column is same as ‘Quora’.
2. votes: It contains the number of people who voted (upvote or downvote) for a question.
3. answers: It contains the number of answers available for each question.
4. views: It contains the number people who viewed the question.
5. asked year: It contains the year in which a particular question was asked.
6. modified year: It contains the year in which the question was last modified.

• A Glimpse of the Data sets (Showing only first 6 data points)

****Quora Data Set****

	questions	topic
1	What is the latest technology 1?	technology
2	What is blockchain technology 1?	technology
3	What are the best technology and gadget blogs?	technology
4	How do you define technology?	technology
5	Is there such thing as Lost Technology or is that a myth?	technology
6	What are some mind blowing facts related to technology?	technology
	followers	answers
1	700	324
2	656	459
3	765	111

4	164	111
5	791	221
6	817	238

****Stack-Exchnage Data Set****

		questions				
1	Edge of factoring technology?					
2	Difference between Topological Data Analysis and Graph Technology					
3	Technology Behind BaKoMa Tex					
4	Good Technology Resources for a Pre-Algebra Class?					
5	How to incorporate meaningful and purposeful technology into math lessons.					
6	Hide Technology Usage					
	topic	votes	answers	views	asked_year	modified_year
1	technology	3	4	318	2011	NA
2	technology	5	2	1101	2014	NA
3	technology	5	0	510	2017	NA
4	technology	3	2	127	2012	NA
5	technology	1	1	31	2018	NA
6	technology	1	1	179	2014	NA

****Stack-overflow Data Set****

	questions	topic	votes	
1	What good technology podcasts are out there?	technology	526	
2	How to get URL parameter using jQuery or plain JavaScript?	technology	760	
3	How to center a "position: absolute" element	technology	911	
4	How does push notification technology work on Android?	technology	259	
5	WebSockets vs. Server-Sent events/EventSource	technology	1113	
6	Using Git with Visual Studio [closed]	technology	1463	
	answers	views	asked_year	modified_year
1	97	158705	2008	2011
2	34	1603456	2013	2022
3	31	1256127	2011	2021
4	5	89694	2012	2020
5	6	288278	2011	2022
6	16	474002	2009	2020

Possible Biases in the Data

- Here we have considered some selective limited number of topics to compare different QnA websites. This will result bias in the data to some extent.
- There may exist some confounding variable which affects the relationship between the number of followers and number of answers (E.g. asking year of the question) on which data as not available for the quora data set, which may result bias in the analysis.

Interesting questions to ask from the data

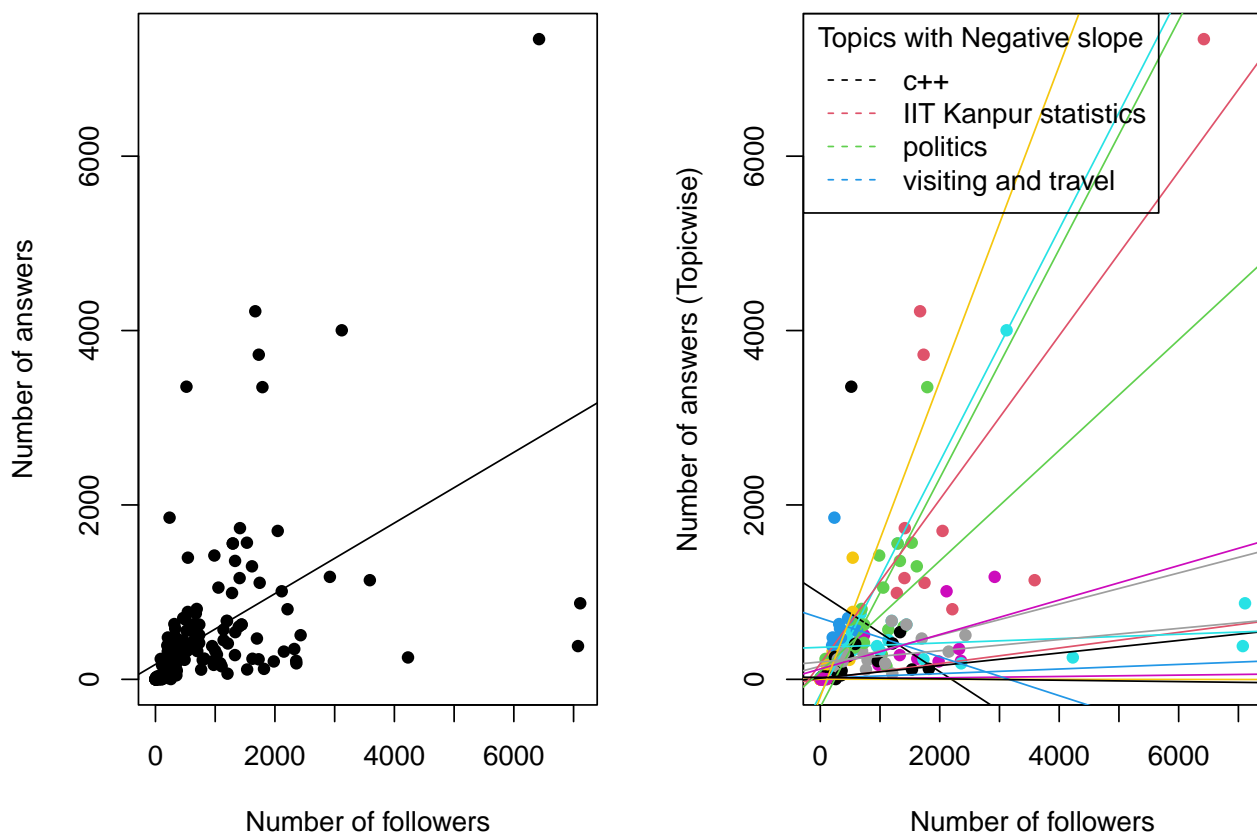
1. Which topic is more popular (or less popular) throughout different QnA websites?
2. Is there any relationship between followers and answers? Is there any discrepancy?
3. Which coding languages have more followers/votes?
4. Which QnA platform is the best to ask the questions?

Important Visualizations

- **Scatterplot of answers vs followers from Quora Data set:**

In general, if most of the people wants to know the answer of a particular question then they follow that question in 'Quora' so that they get a notification as soon as some solution to that question appears. Now, as the number of followers increases, the question reaches out to more people and the chance of the question being answered increases. Let's try to visualize this through the following scatterplots,

Scatterplot of answers vs followers



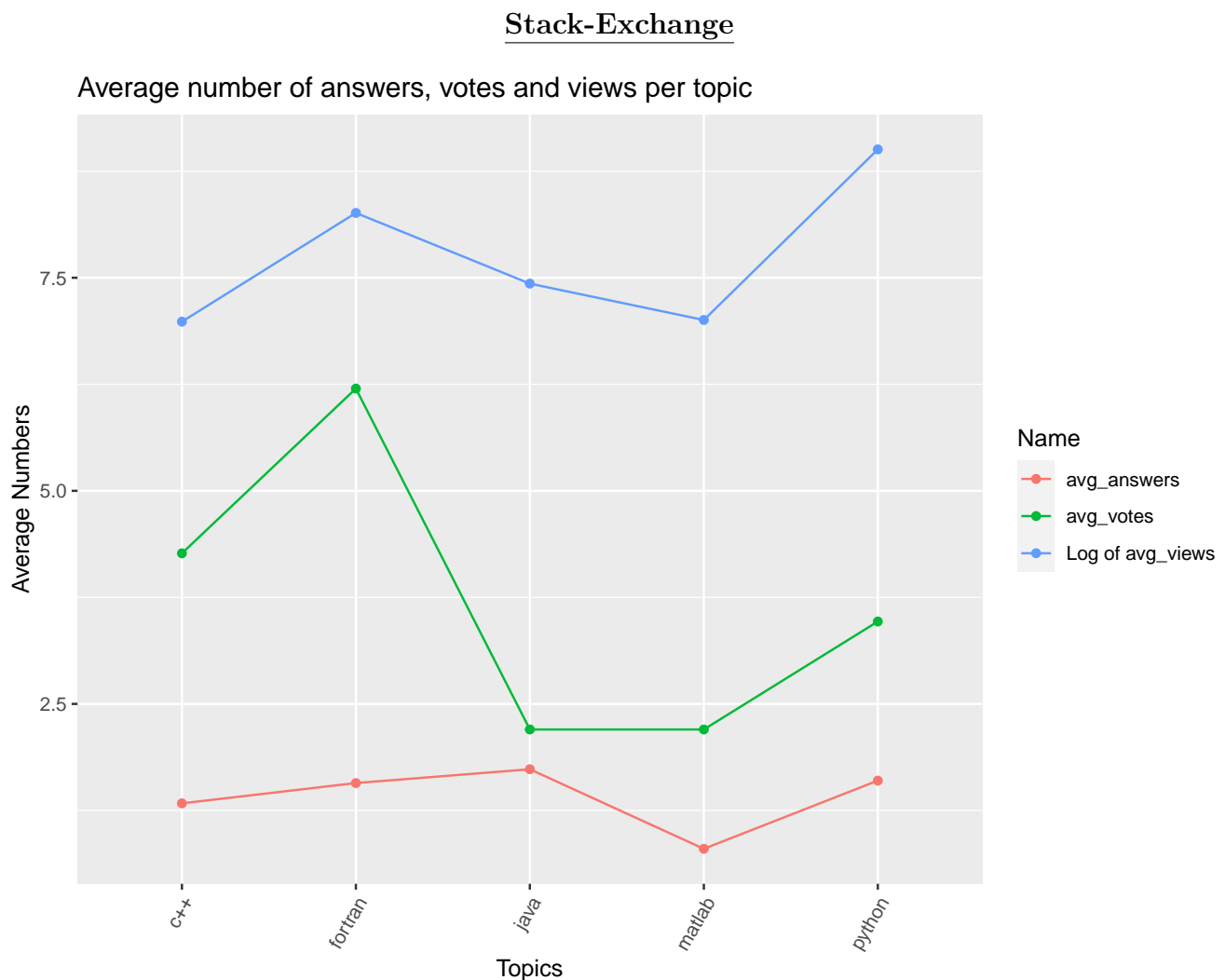
Clearly, from the first plot we can observe that the number of answers and the number of followers are positively correlated (General trend). But, if we do the plot topic wise then for 4 topics the relation between the number of answers and the number of followers are negatively correlated.

Specially, c++ which is a very popular coding language, has the estimated regression line with a slope of very high negative value. Why this unnatural behaviour?

Now, 'Quora' is a general platform to ask questions. So, the reason may be c++ being a very basic and important programming language most people want to learn it and that's why they have many questions but they don't know the answers of them which results in decreasing number of answers to the questions as the number of followers increases.

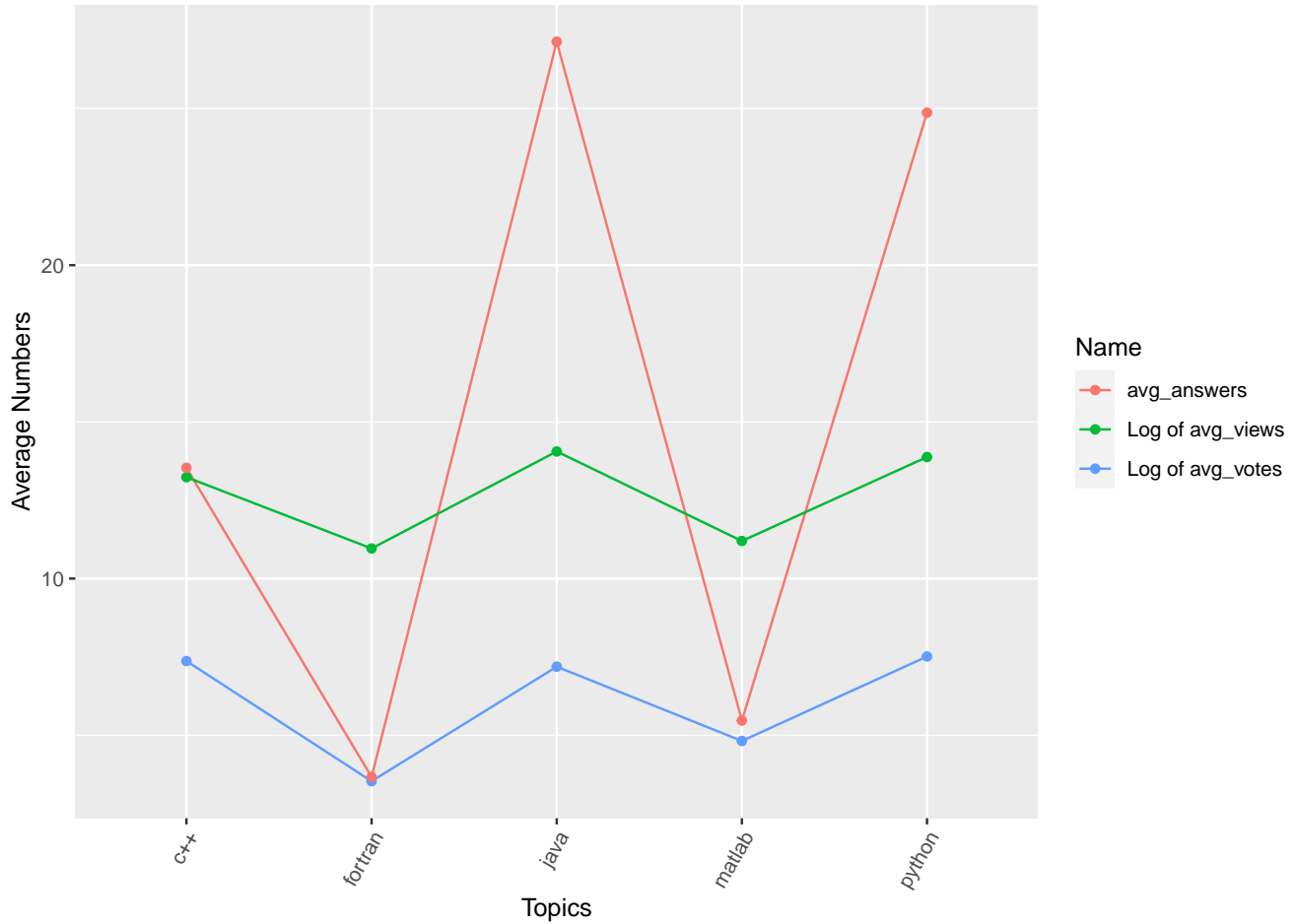
- **Line diagram of average number of votes, views and answers per topic for Stack-Exchange and Stack-overflow Data Set:**

Stack-Exchange and Stack-overflow are more technical sites. So, we have plotted the average number of votes, views and answers per coding topics (python, java etc.) for both the platforms for comparison purpose,



Stack-overflow

Average number of answers, votes and views per topic

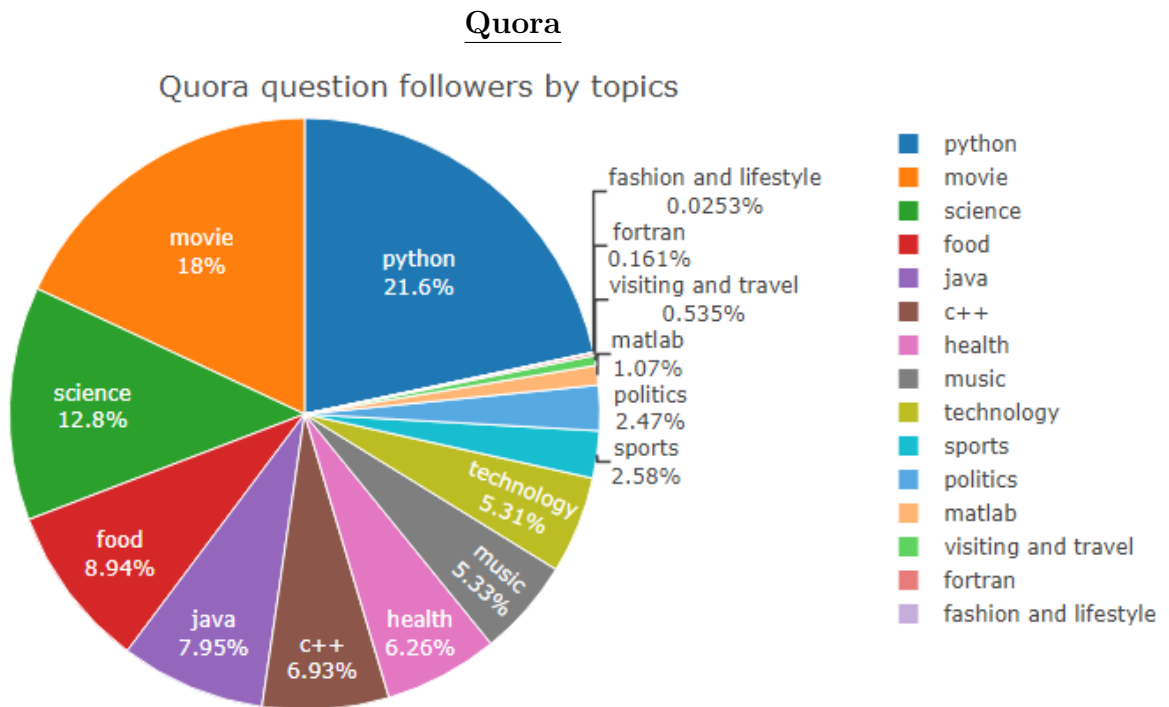


Here, we have scaled down the attributes by taking $\log()$ for comparison purpose.

Here also, though the average number of views and votes for the c++ language are comparatively larger than the rest of the languages, but still the average number of answers are very less in that comparison. This strengthens our observation in the previous visualization that **c++** being a very **basic and important programming language** most people want to learn it and that's why they have many questions but they don't know the answers.

Also, we can observe that python programming language may be in most demand now a days with a huge number of views and votes to the questions.

- Pie chart showing followers/views by topics

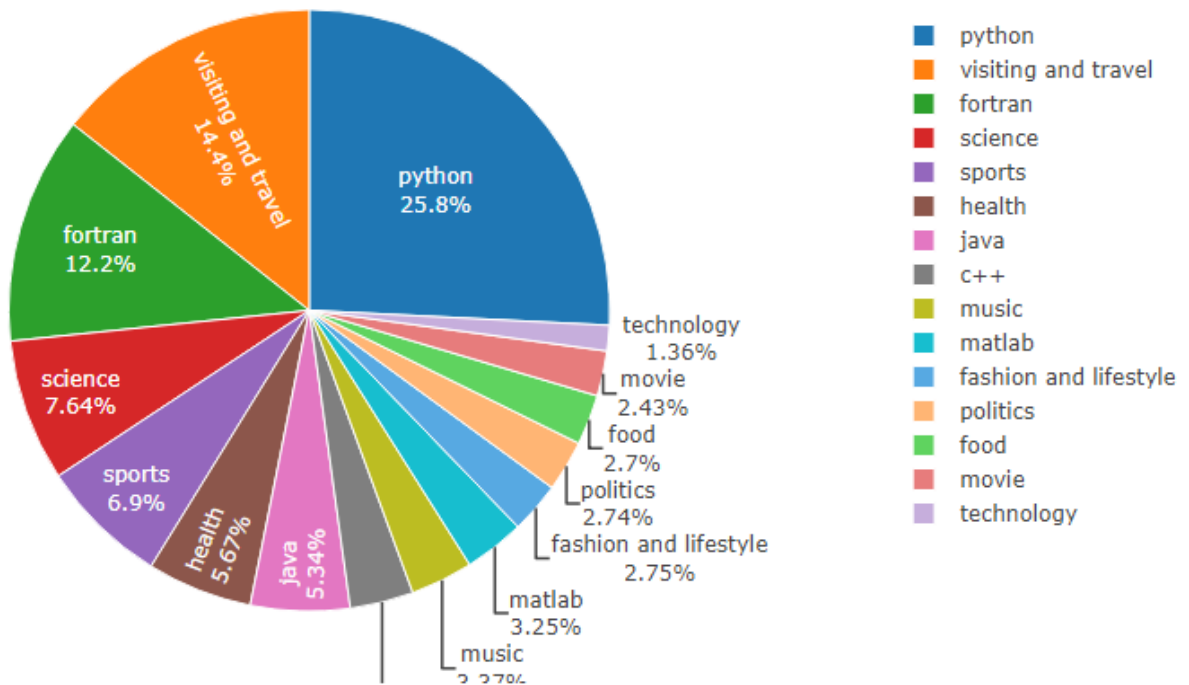


We know, Quora is a social QnA (Question and Answer) website where people ask questions related to every field. Still, python questions have maximum number of followers compared to other topics like music, movie, food etc. From this we can guess the popularity of the 'Python' programming language.

Whereas, Stack-Exchange and Stack-overflow being more technical websites has obviously maximum number of views for the python programming language. We can visualize those through the following plots,

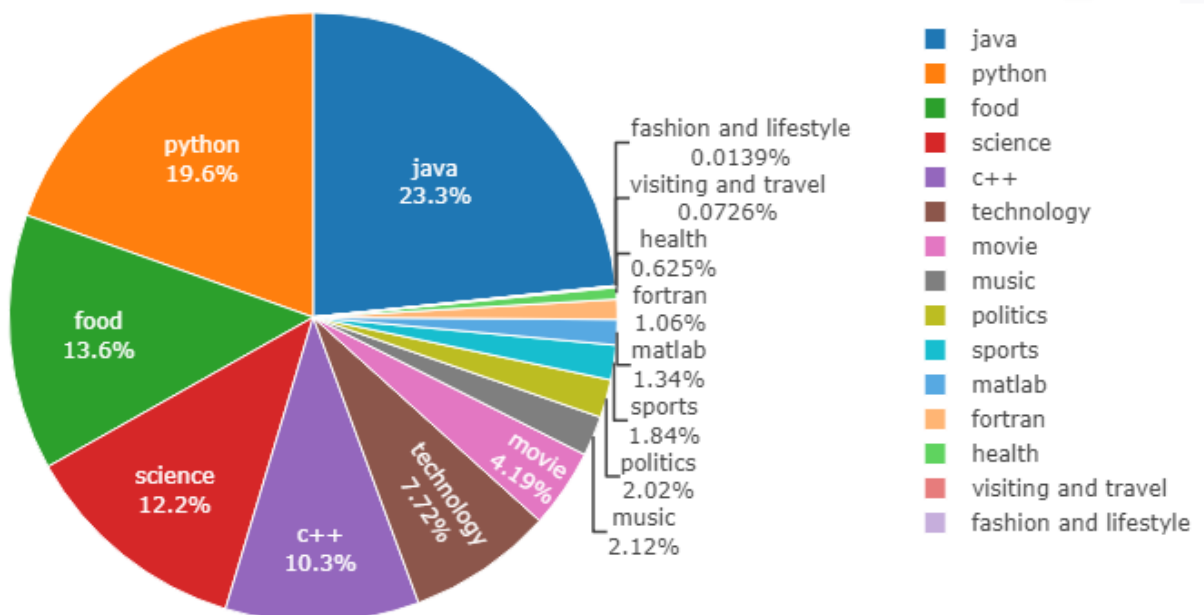
Stack-Exchange

Stack-exchange question views by topics



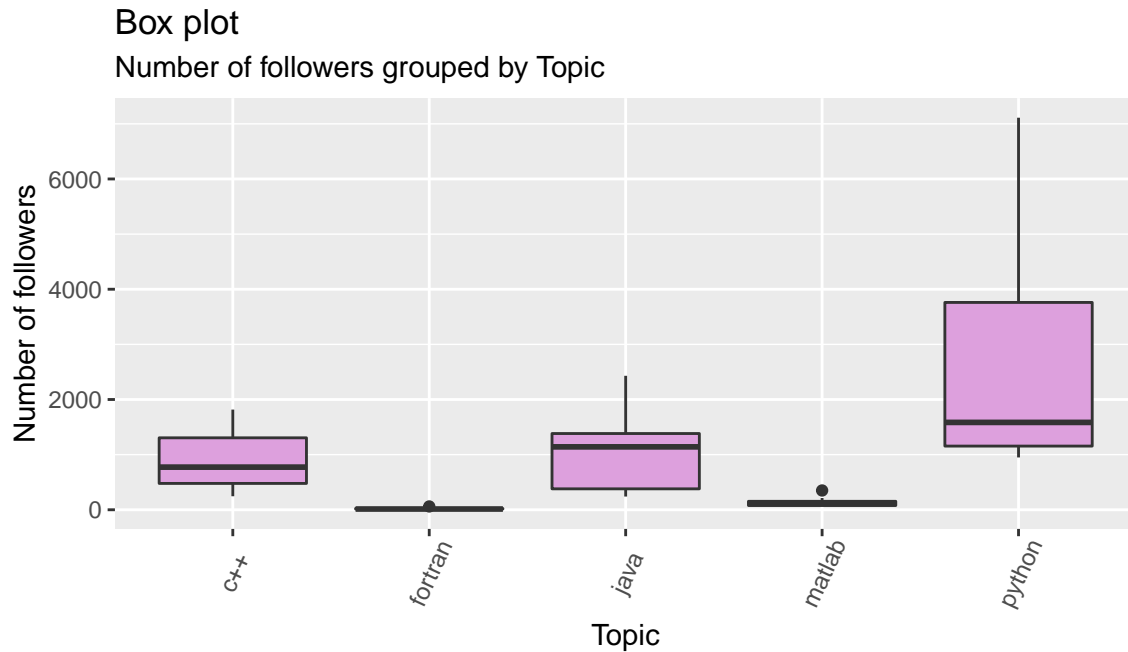
Stack-overflow

Stack-overflow question views by topics



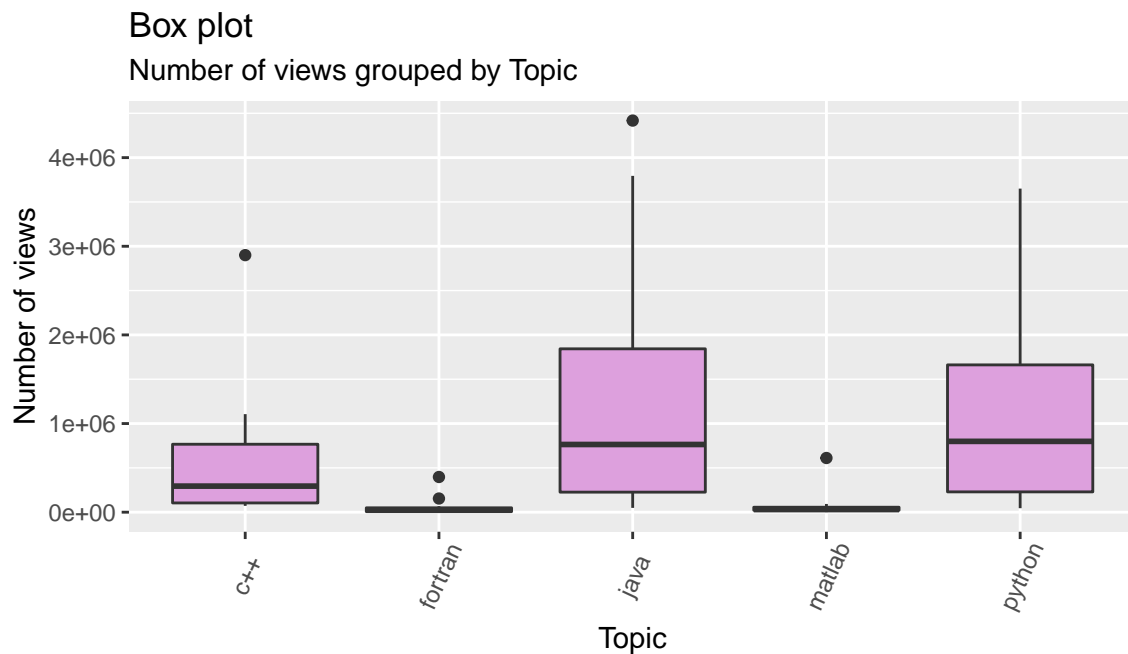
- Boxplot showing number of followers/views grouped by topics:

Quora



From the plot it is clear that, though python has on an average most number of followers but there is more variability in the number of followers compared to the other programming languages i.e. some questions have very less number of followers and some have very large number of followers.

Stack-overflow



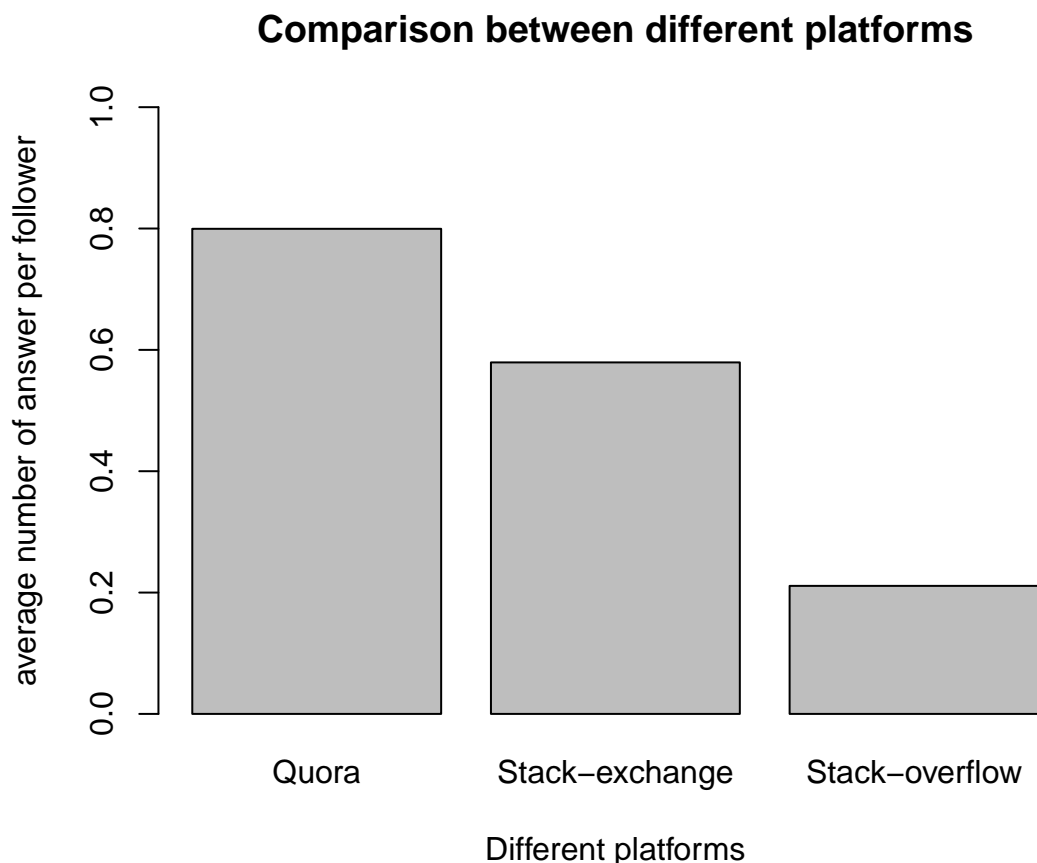
Stack-overflow being a more coding related question answer platform is showing a more or less similar pattern in 'Java' and 'Python'.

- **Comparative study between the three Question and Answer websites:**

The three ‘Question and Answer’ websites ‘Quora’, ‘Stack-exchange’ and ‘Stack-overflow’ has different types of users. Quora is a more general site. Whereas, Stack-exchange and Stack-overflow are more technical sites. So, we can’t directly compare these websites in terms of any of the available variables. Rather, we can compare these QnA sites in terms of the number of answer per follower, being the ratio of two such variables (It becomes unit free).

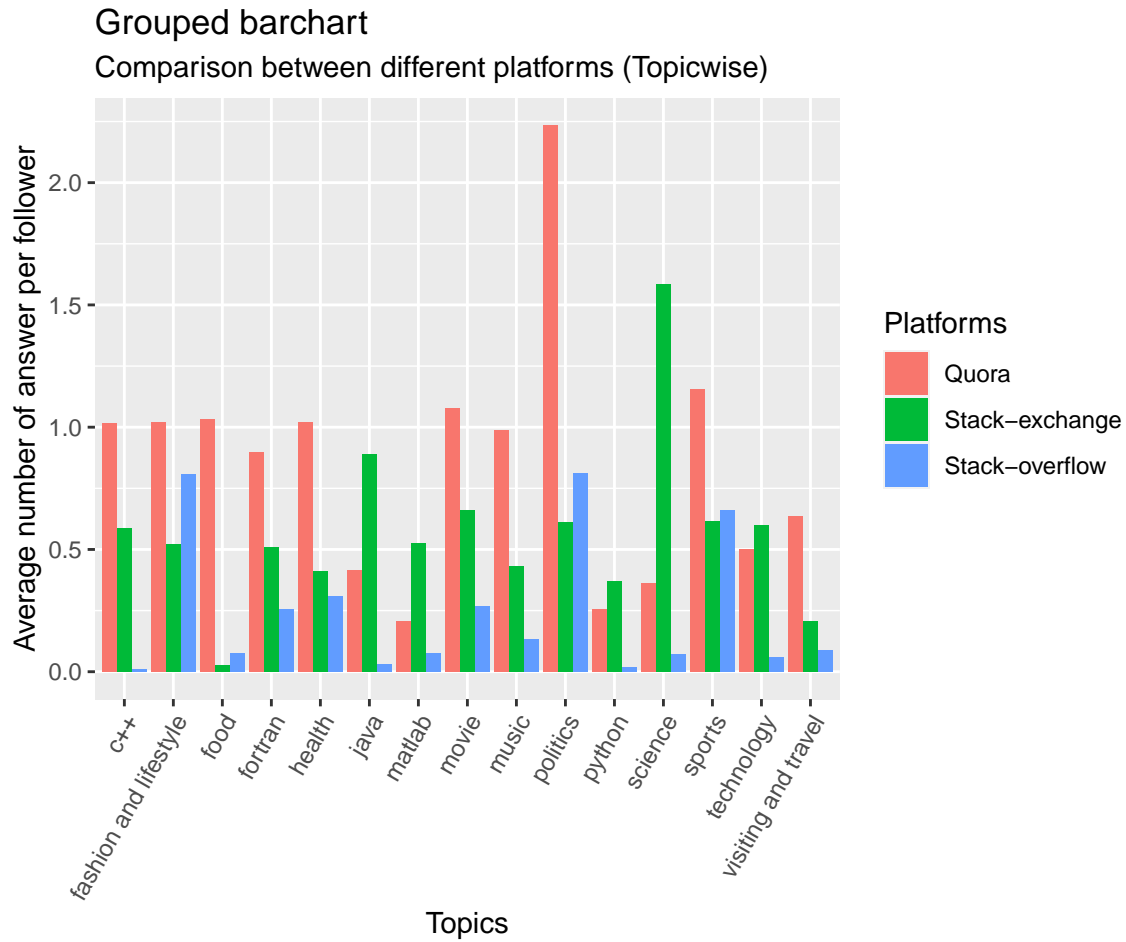
(In quora, for being a follower of a question you have to press the follow button. This is somewhat similar of pressing the upvote (/downvote) button for a question in Stack-exchange and Stack-overflow. That’s why we have selected the ‘votes’ column from the stack-exchange data set (Rather than number of views) in parallel to the ‘followers’ column in the quora data set.)

To compare the performance of different Question and Answer platforms we have plotted vertical bars for each site where the number of answer per follower is plotted on the y-axis.



It seems that Quora is the best site to ask a question regarding any topic as the number of answer per follower is the largest for quora. So, if we ask our question in quora, then we will have more number of answers to our question from which we can select the desired one.

Now, we will repeat the same plot where the bars are grouped by topics,



Now, it is clear that the average number of answer per follower is largest for quora for all the topics except ‘java’, ‘matlab’, ‘python’, ‘science’ and ‘technology’. For these topics Stack-exchange has the most average number of answers per follower. This is quiet obvious as, these topics are more technical topics than the rest of the other topics.

Clearly, this plot rejects our previous hypothesis that ‘Quora’ may be the best place to ask any type of questions regarding any topic.

Hence, we may infer that for general topics ‘Quora’ is the best platfrom to ask questions. Whereas, for the technical or coding topics ‘Stack-exchange’ and ‘Stack-overflow’ are the best to use.

R Shiny App

- The side panel of the R shiny consists of selections of QnA platforms. Then multiple group check boxes are provided for 17 topics. Also, there are some conditional panel based on the selected dataset.
- Corresponding to the user selected topics, various images will load on the main panel which are explained thoroughly in the report.
- For the comparative study of the three QnA websites the user have to select the ‘All’ option in the dataset.

Conclusion

- **c++** being a very **basic and important programming language** most people want to learn it and that’s why they have many questions but they don’t know the answers.
- **Python programming language** is the most popular coding language (may be popular than other topics also) now a days.
- For general topics (movie, music etc.) ‘**Quora**’ is the best platfrom to ask questions. Whereas, for the technical or coding topics ‘**Stack-exchange**’ and ‘**Stack-overflow**’ are the best to use.

References

- <https://en.wikipedia.org/wiki/Quora>
- https://en.wikipedia.org/wiki/Stack_Exchange
- https://en.wikipedia.org/wiki/Stack_Overflow
- <https://shiny.rstudio.com/articles/layout-guide.html>
- <https://plotly.com/r/pie-charts/>
- <https://r-graph-gallery.com/48-grouped-barplot-with-ggplot2.html>
- <https://rmarkdown.rstudio.com/>