Career Village Documentation

DSBA 6211 - Advanced Business Analytics

MAY 2022

PREPARED BY
Carson Groulx
Naomi Thammadi
Saurabh Vijay
Shina Dhingra
Will Patten

Executive Summary

Career Village is an online platform that intends to provide students and professionals across the world a forum to discuss career-related topics. Many students are often at a crossroads to decide which career path is best for them. They need advice regarding their field of study, internships, jobs, career advice, etc. Career Village offers them a platform to ask questions related to a certain topic and have professionals from that field of industry answer these questions. Professionals can subscribe to the topics they want to be notified about when new questions are posted. They can also join group memberships and indicate their school memberships. Questions and Answers are scored by users based on their popularity, usefulness, and how they have solved the purpose.

This project explored the different dynamics of professionals based on their location, industry, and subscriptions. It also explored dynamics related to student location and activity levels. The questions and answers posted were analyzed based on time and text dynamics to further look into how each of them was being scored.

This project hopes to provide enough insights for the Management of Career Village to understand how users communicate on the platform and provide ways to improve the overall customer satisfaction - students finding answers faster, professionals notified about the right topics to respond to. It also hopes to educate about the activity levels of different students in different locations and time-to-response of questions related to a variety of topics. This can help promote the platform to certain professionals of particular groups to improve the user base.

This project applied some advanced Exploratory Data Analysis techniques using Tableau, Python, and R. Models were then created using algorithms like K-means Clustering, Topic Modeling, Sentiment Analysis, Decision Trees, Linear Regression and Logistic Regression.

We as a team have learned a lot through this project and have developed some advanced data analysis skills and improved our interpersonal skills by working in a team.

Introduction & Background

CareerVillage.org is a nonprofit that crowdsources career advice for underserved youth. Founded in 2011 in four classrooms in New York City, the platform has now served career advice from 25,000 volunteer professionals to over 3.5M online learners. The platform uses a Q&A style similar to StackOverflow or Quora to provide students with answers to any question about any career. The main goal is to help recommend questions to appropriate volunteers. CareerVillage.org has supplied five years of data.

The U.S. has almost 500 students for every guidance counselor. Underserved youth lack the network to find their career role models, making CareerVillage.org the only option for millions of young people in America and around the globe with nowhere else to turn. To date, 25,000 volunteers have created profiles and opted in to receive emails when a career question is a good fit for them. To help students get the advice they need, the team at CareerVillage.org needs to be able to send the right questions to the right volunteers. The notifications sent to volunteers seem to have the greatest impact on how many questions are answered.

Project Objectives

1. Find the topics most popular among students.

2. Find the industry of professionals who respond the most to questions.

3. Find the role email notifications play in getting a question answered

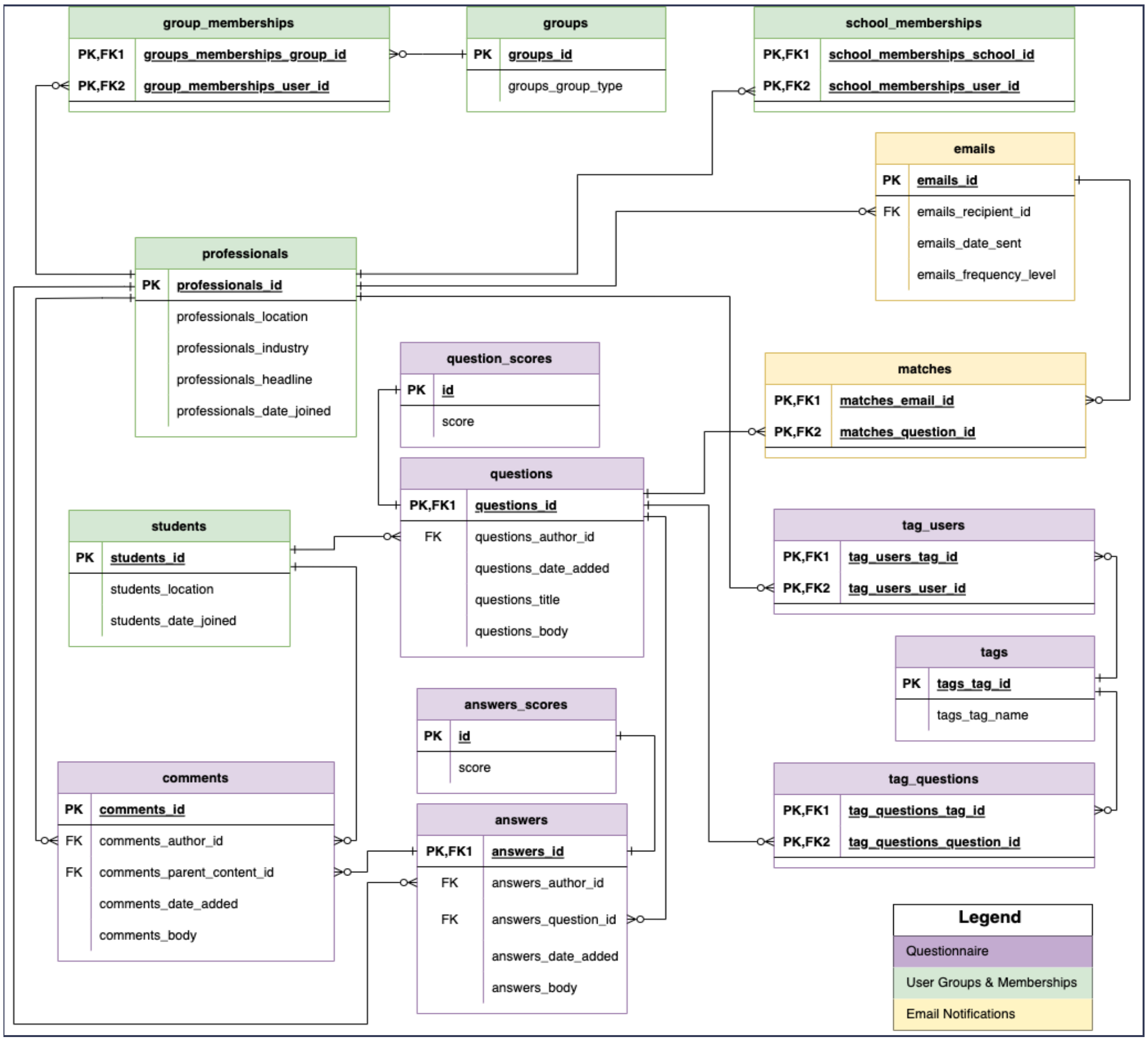
Data Cleaning and Structure

Data Tables

1. Answers: Answers get posted in response to questions. Answers can only be posted by users who are registered as Professionals. However, if someone has changed their registration type after joining, they may show up as the author of an Answer even if they are no longer a Professional.
2. Comments: Comments can be made on Answers or Questions. We refer to whichever the comment is posted to as the "parent" of that comment. Comments can be posted by any type of user.
3. Emails: Each email corresponds to one specific email to one specific recipient. The frequency_level refers to the type of email template which includes immediate emails sent right after a question is asked, daily digests, and weekly digests.
4. Group memberships: Any type of user can join any group. There are only a handful of groups so far.
5. Groups: Each group has a "type". For privacy reasons, the group names are off.
6. Matches: Each row tells you which questions were included in emails. If an email contains only one question, that email's ID will show up here only once. If an email contains 10 questions, that email's ID would show up here 10 times.
7. Professionals: Volunteers are called "Professionals" and volunteer their time to answer questions on the site.
8. Questions: Questions get posted by students. Sometimes they're very advanced. Sometimes they're just getting started. It's all fair game, as long as it's relevant to the student's future professional success.
9. School memberships: Just like group memberships, but for schools instead.
10. Students: Students are the most important people on CareerVillage.org. They tend to range in age from about 14 to 24 and are spread across the world.
11. Tag questions: Every question can be hashtagged. Hashtag-to-question pairings are put into this table.
12. Tag users: Users of any type can follow a hashtag. This shows you which hashtags each user follows.

13. Tags: Each tag gets a name.
14. Question scores: "Hearts" scores for each question.
15. Answer scores: "Hearts" scores for each answer

Data Model



Data Pre-Processing

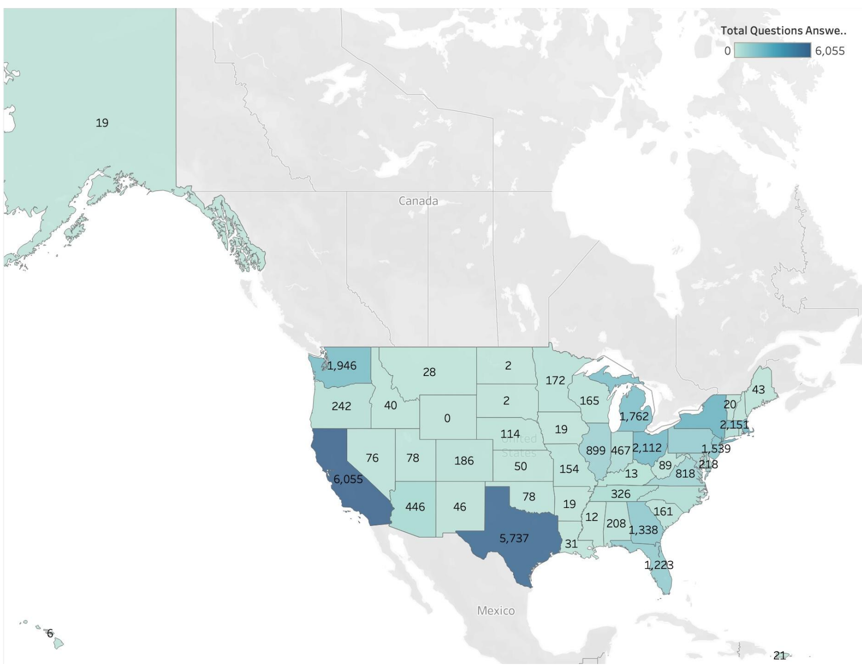
Data from common tables are combined and included into consolidated files for further analysis. Most important pre-processing was performed in consolidating and summarizing tags on questions, tags professionals subscribe to, combining scores on questions and answers, and summarizing emails sent on an immediate, daily, and weekly basis.

Exploratory Data Analysis

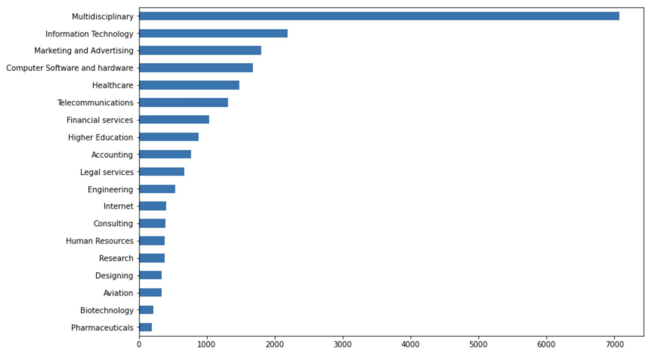
Professional Dynamics
Here professionals were filtered based on the number of answers, wherein only professors with more than 10 answers were included in the analysis. This helps us understand the dynamics of most active professionals only.

Here analysis contained 3 different kinds of dynamics as follows:.

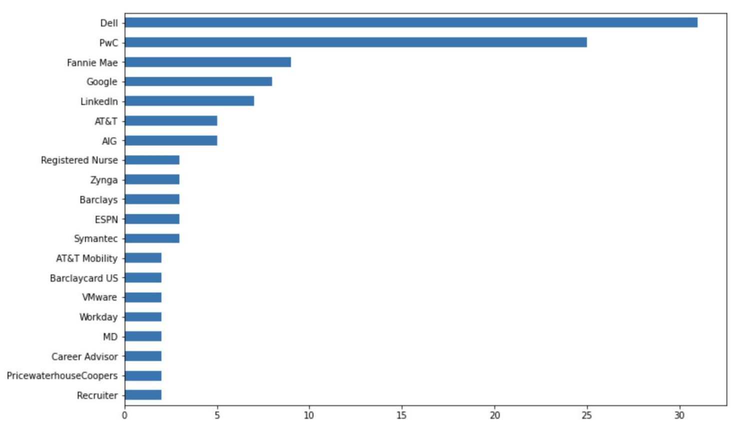
1. Location: A map was used for this analysis. As more than 90% of professionals are from the United States thus location analysis was done on U.S. based professionals only. We determine that most of the professionals are from California or Texas.



2. Industry: After grouping in 20 industries we find that most professionals are from multidisciplinary industries followed by information technology.



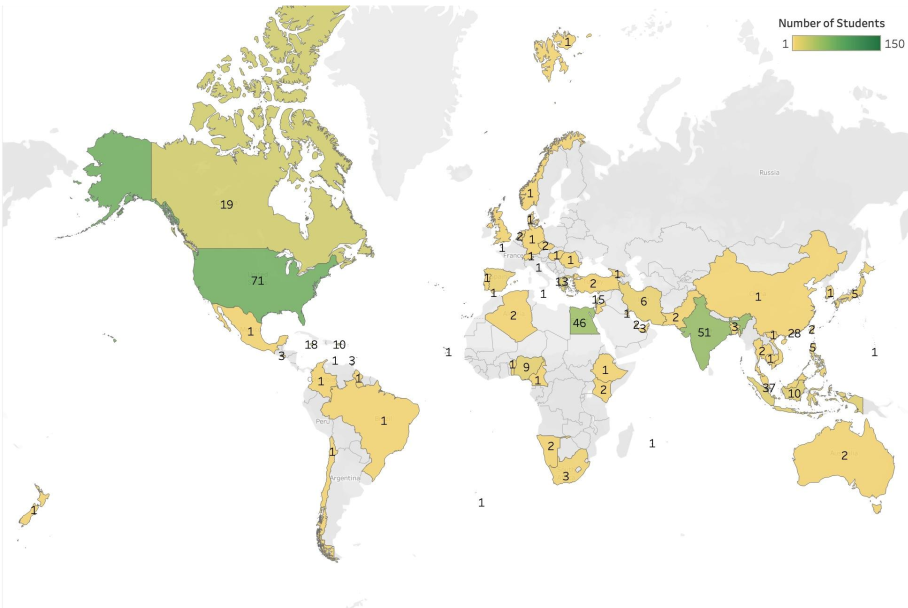
3. Company: Based on the professionals heading column we did text analysis and made a histogram based on frequency. We conclude that most professionals relate to consultancy based firms and IT companies. We also created a job position word cloud for further analysis.



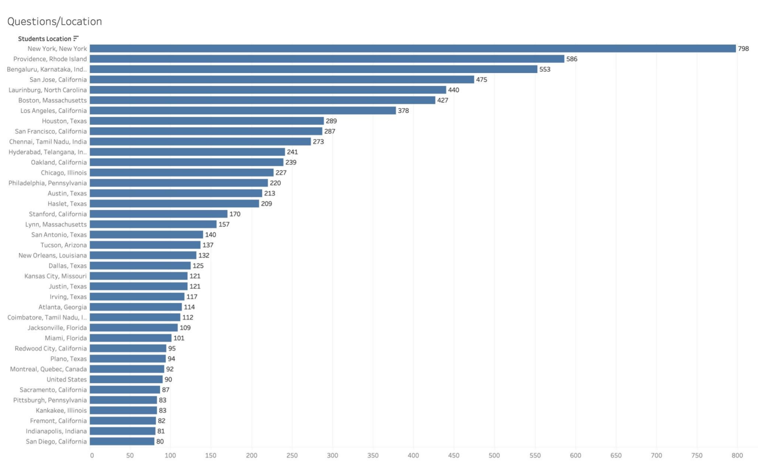
Student Dynamics

Different columns related to students were used like their location of membership and question author location. We wanted to analyze if the students of a particular location were more active than others.

1. Location of membership: It signifies the membership of the student. Results indicated that students are spread around the world with most students from the USA, followed by India and Egypt.



2. Activity: Here we analyzed what areas are most active on the platform irrespective of their membership. We concluded that even though memberships are spread around the world still most active users are in the eastern states of the U.S.

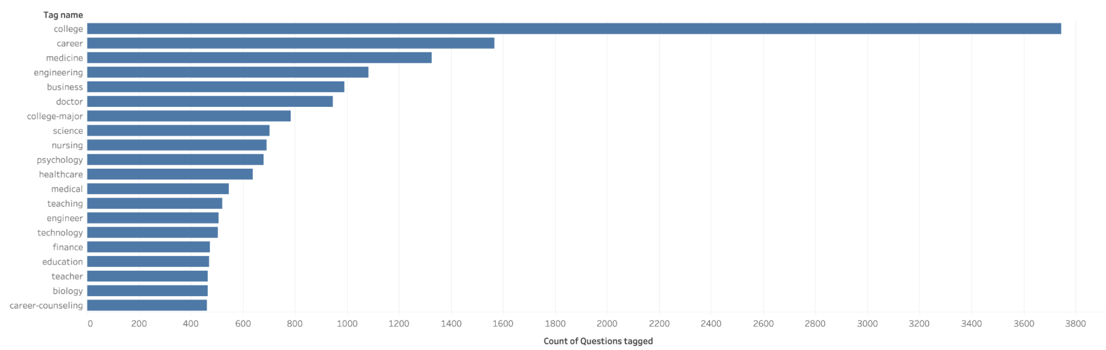


Tag Dynamics

The most popular and most subscribed tags show us which tags have a lot of activity on them. They further show us the time-to-response for different tags.

We see that students ask a majority of questions about their career and college, especially in the fields of medicine, engineering, and business.

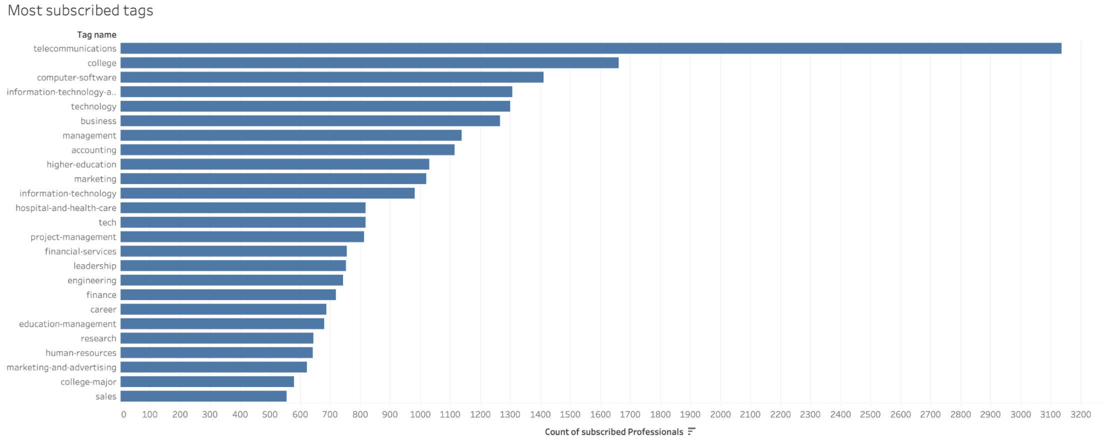
Most popular Tags on Questions



Sum of Count of Questions tagged for each Tag name. The view is filtered on sum of Count of Questions tagged, which ranges from 450 to 3,744.

Professionals however have a higher subscription to tags related to Telecommunications, Colleges, Computer Software, Information Technology, Business, and Management. This aligns with the popular tags and going forward we understand how professionals associated with these tags respond faster.

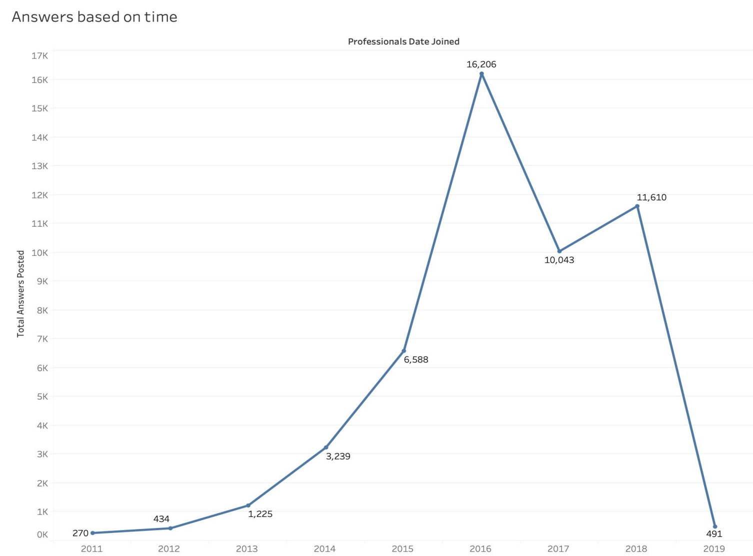
Most subscribed tags

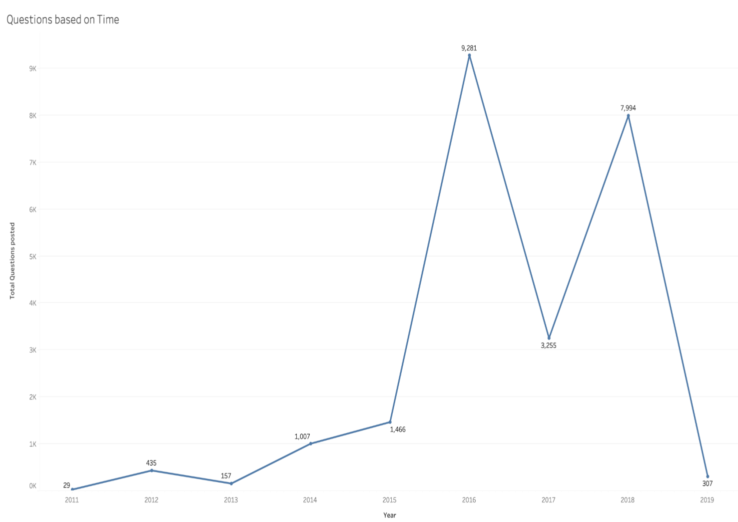


Sum of Count of subscribed Professionals for each Tag name. The view is filtered on sum of Count of subscribed Professionals, which ranges from 550 to 3,135.

Time Dynamics of Activity Levels

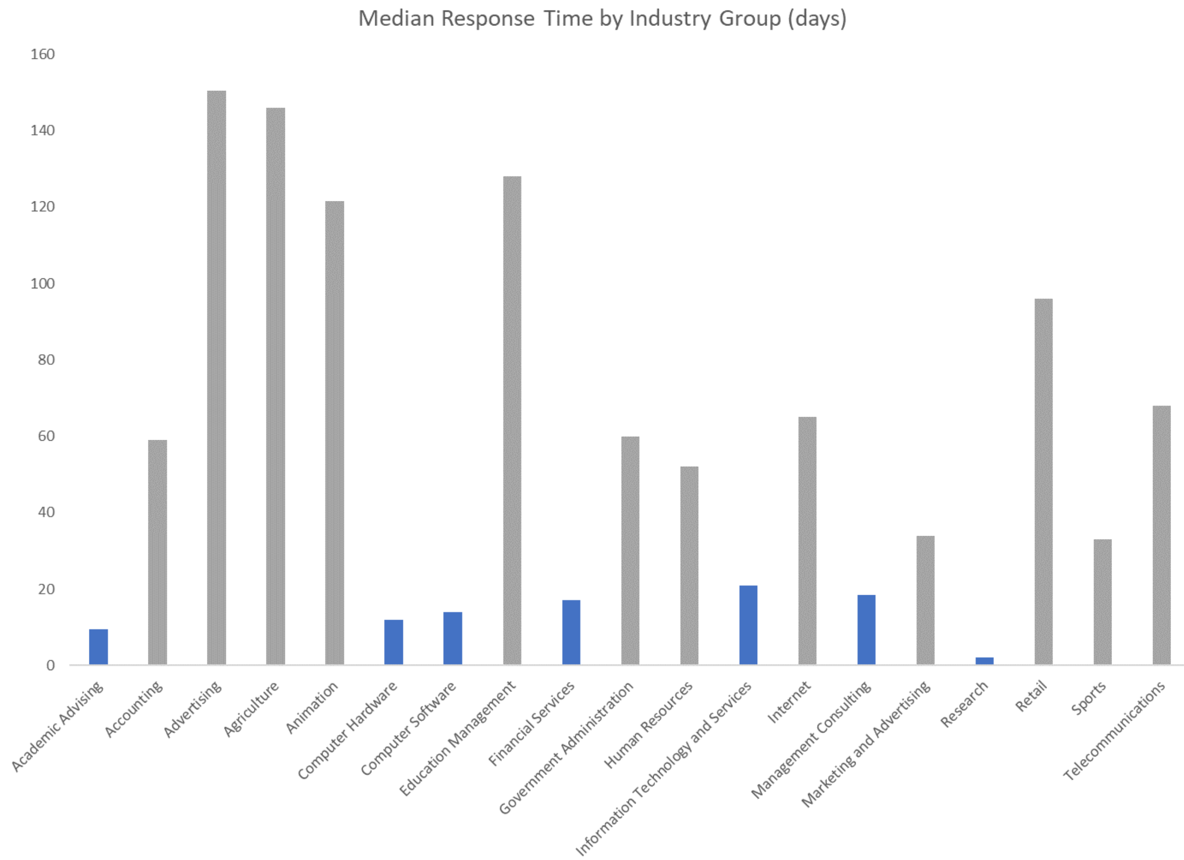
Time series analysis of both question and answer frequency was needed to determine if there is any trend or seasonality associated. We calculated the number of questions posed and answers posted on a year by year basis and found that there is an upward trend in both metrics with a peak in the year 2016. However, a downward trend in activity levels is visible post 2016. This change can be attributed to the growth of other such platforms like Unacadedy in major markets.





The trend of count of Questions Id for Question Date - Year Year. The marks are labeled by count of Questions Id.

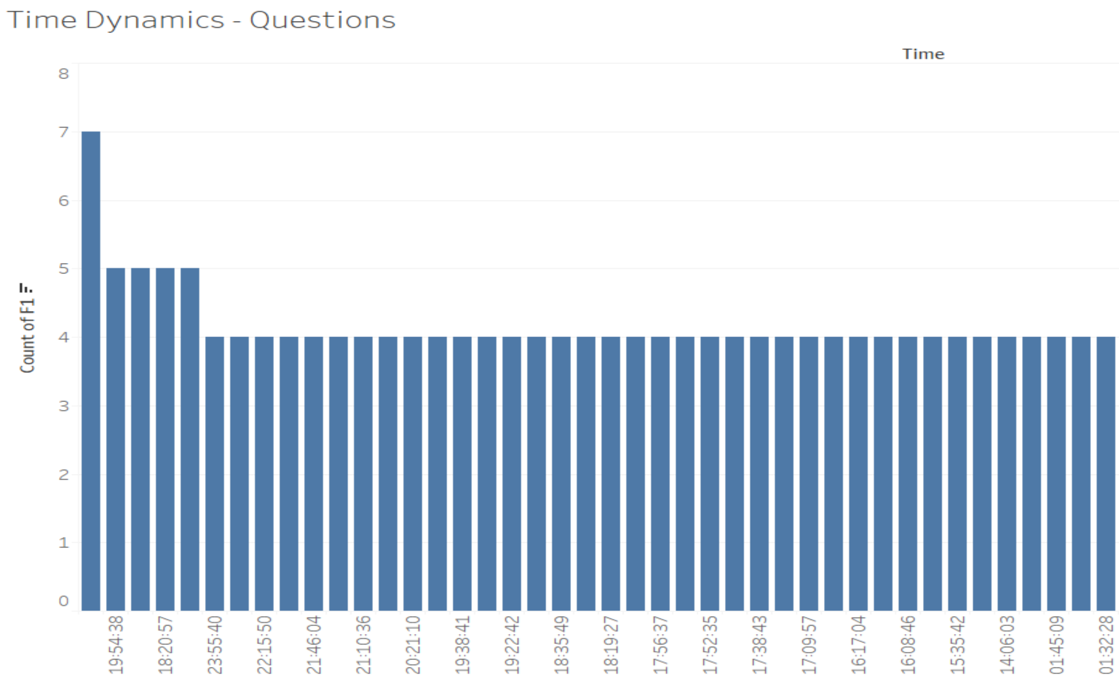
Response times (difference between when a question was posted and when the question was answered) for professionals show that some of our most active groups are the quickest to respond, such as academic advising, research, and technology related professionals. Those in and around universities appear to be a part of the fastest answering professionals.



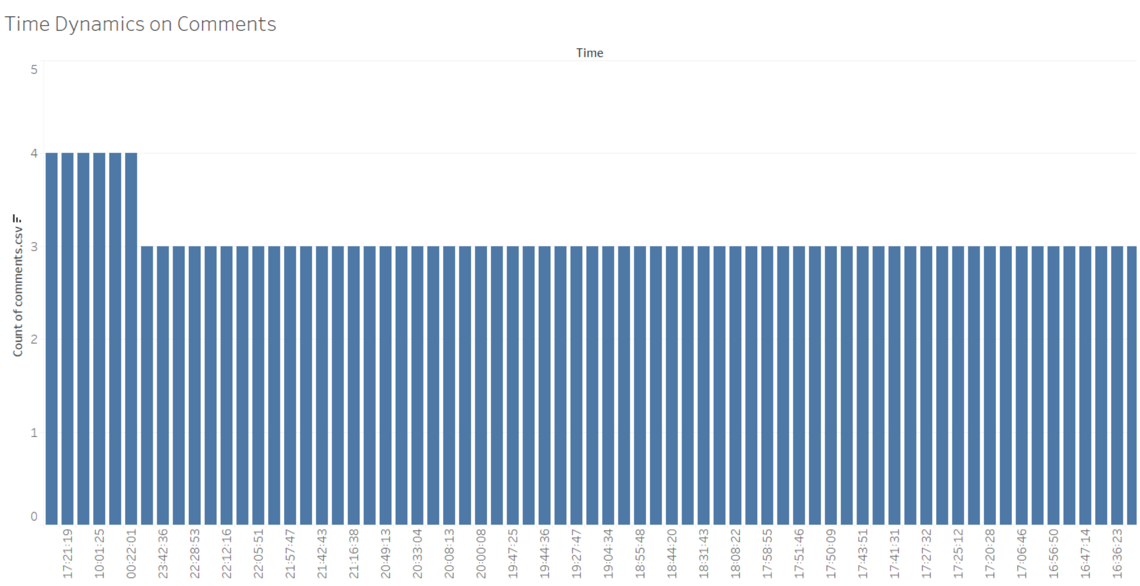
Activity Levels on Questions and Comments

We try to track the activity levels of students and professionals to find out at what times during the day are they most active to post questions and comments. We also analyzed the same for answers but that does not display a substantial insight and the activity is spread across the day.

Most questions posted by the students are after 2 pm.

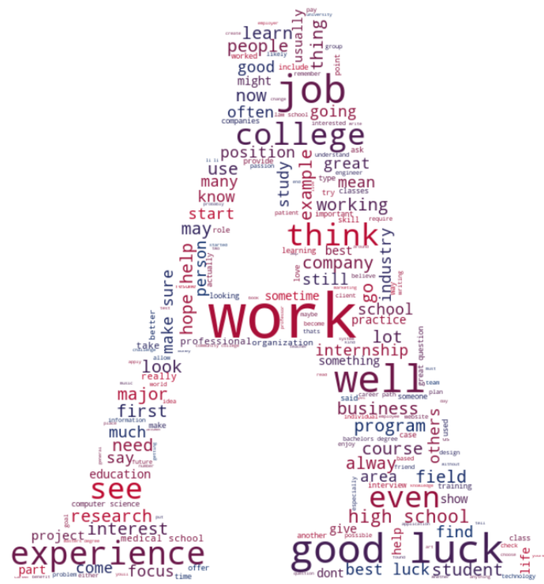


Most comments posted are after 4 pm.



Text Dynamics of Questions and Answers

Both questions and answers talk a lot about shaping a student's career. We see students mostly asking questions about choosing their career paths, advice to apply for jobs, internships, skills required for them, major to choose, etc.



Modeling Methods

Different Modeling Techniques were used to explore trends and perform Predictive Modeling. The programming languages used are R and Python.

Cluster Analysis

One issue we faced was related to multiple types of professional industries and tags used. Due to this, it became harder to do further analysis and modeling. We resolved this issue by using a similarity matrix and cluster analysis. We wanted to create 15-20 clusters of different industries and tags which are similar and can be grouped together so that further dummy coding can be done without increasing much dimensionality. The process followed included creating tf-idf similarity matrix and using it for cluster analysis.

1. Industries: We aimed at creating 20 clusters for industries and got good clustering results. Based on these results we gave industry categories to each cluster manually. This approach helped us to analyze professional dynamics.

2. Tags: Following the above approach we created 15 clusters for tags and labeled each cluster manually so as to categorize them. These categories were further used by dummy coding for each category and using it in the analysis.

Text Analysis

Three text analysis techniques were applied to questions and answers and metrics from these analyses were used further in logistic and linear regression.

1. Text Similarity Metrics of Question/Answer body with Tags
2. Sentiment Analysis of Questions and Answers
3. Topic Modeling of Questions and Answers

String distance calculations are performed to find similarities in tags and question and answer text. This is to check if students are tagging questions correctly. “Jaccard” distance measure was used to create a similarity index.

The R package “sentimentr” was used to perform sentiment analysis on both Questions and Answers.

Topic Modeling was performed on both questions and answers using Python’s “Gensim” library. Since 15 clusters of tags were created earlier, 15 topics were generated during the topic modeling phase to see how well the model categorizes the questions.

Resultant topic clusters are added as a categorical variable to the questions and answers datasets. A summary of all metrics is created which is used further in logistic and linear regression.

Logistic Regression

A logistic regression model was run using the dummy coded tag clusters to determine which topics of questions were most and least likely to be answered within 24 hours. Date variables were converted from characters to POSIXlt using the strptime function, and then the dates the questions were added were subtracted from the dates of the first answers using the difftime function to determine response times. A logical condition was used to create a binary variable which equals 0 if the response times were null, meaning they were unanswered, or were greater than 24 hours and equals 1 otherwise. Tables with question_id were then merged so that all variables were in one table. VIF was checked, data was partitioned, and numeric and binary variables were subsetted in training and validating datasets so that only the desired variables were run in the regression. A glm binomial model was run and a confusion matrix was produced with validating data.

Linear Regression

A linear regression model was run to determine the total number of answers based on factors like question score, number of emails sent, sentiment score, number of tags and the response time to answer a question after an email notification is sent.

Data Preprocessing before creating the model included the following -
- Formatting date variables from string to a datetime format
- Creating another variable to calculate the response time to answer a question after the first email notification is sent.
- Removing data where questions were answered before an email was sent, since we are analyzing answers in relation to the emails sent.
- Removing null values

Ordinary least squares model is run using the sklearn package in Python. Results show that Adjusted R square is 0.508 for 37118 observations.

Variables question score, and response time after email is sent are statistically significant.

Decision Trees

Regression Decision Trees are used to predict the score on questions and answers using the various metrics created earlier.

The metrics used for Question scores are:

1. TOTAL ANSWERS
2. QUESTION SIMILARITY INDEX WITH TAGS
3. COUNT OF TAGS
4. SENTIMENT CATEGORY
5. TAG CATEGORY
6. DOMINANT TOPIC

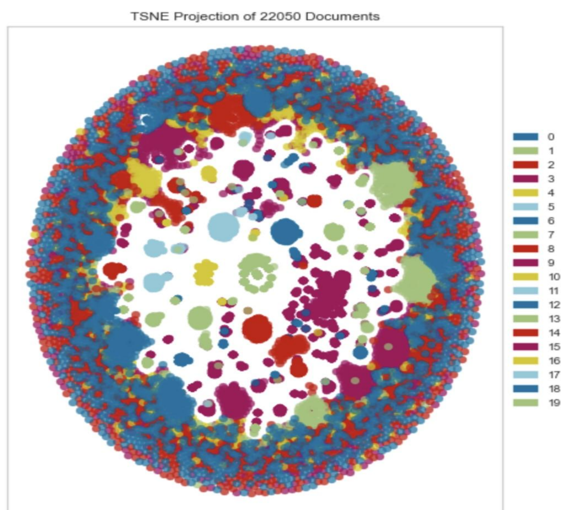
The metrics used for Answer scores are:

1. TOTAL COMMENTS
2. ANSWER SIMILARITY INDEX WITH TAGS
3. COUNT OF TAGS
4. SENTIMENT CATEGORY
5. PROFESSIONAL INDUSTRY CATEGORY
6. DOMINANT TOPIC

Results

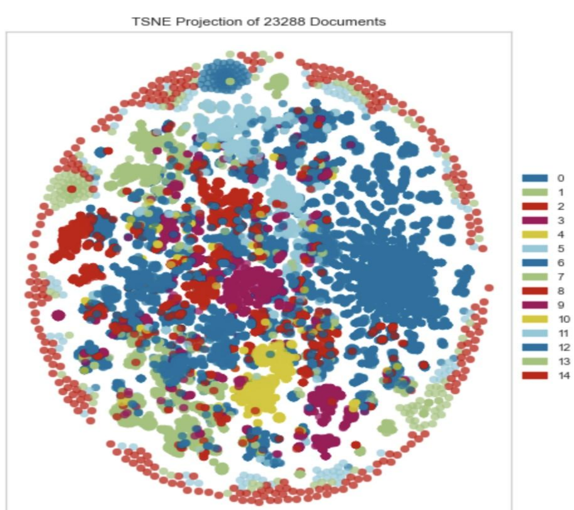
Cluster Analysis

1. Industries: The visualized clusters in multiple dimensions(SVD) along with their labels are given below:



Multidisciplinary
Consulting
Computer Software and hardware
Internet
Designing
Healthcare
Marketing and Advertising
Financial services
Higher Education
Engineering
Research
Information Technology
Aviation
Human Resources
Biotechnology
Legal services
Telecommunications
Accounting
Pharmaceuticals

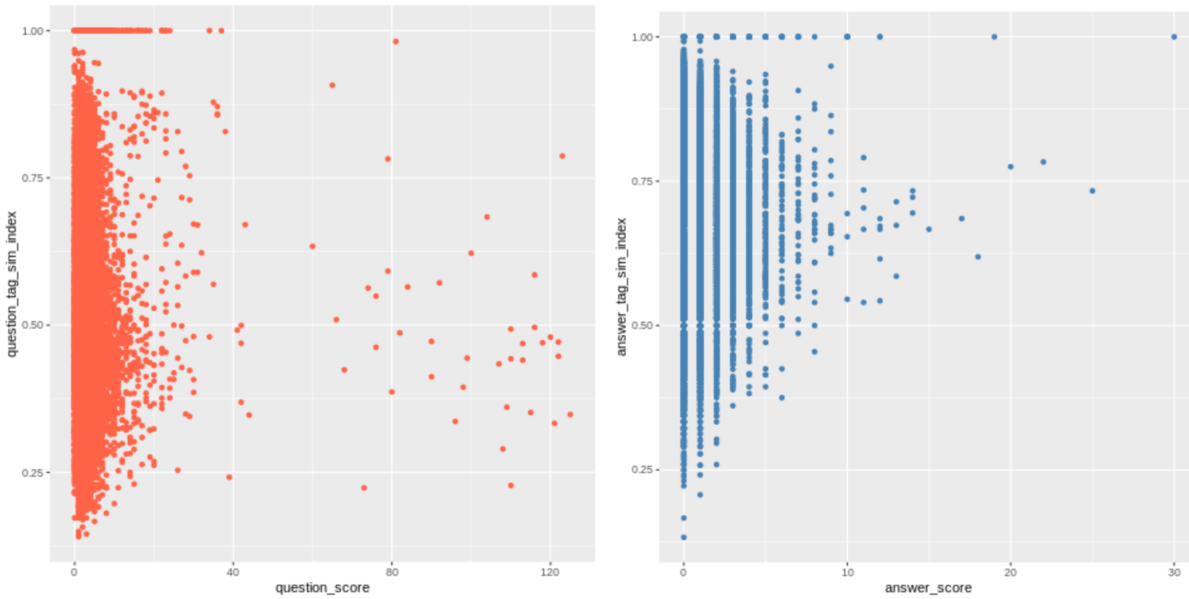
2. Tags: The visualized clusters in multiple dimensions(SVD) along with their labels are given below:



Professional Career
Business
Nursing and Medicine
Engineering
Law enforcement
Psychology
Career Counseling
College
Teaching
Scientist
Social Work
Technology
school
Professional

Text Analysis

- Text Similarity Metrics of Question/Answer body with Tags

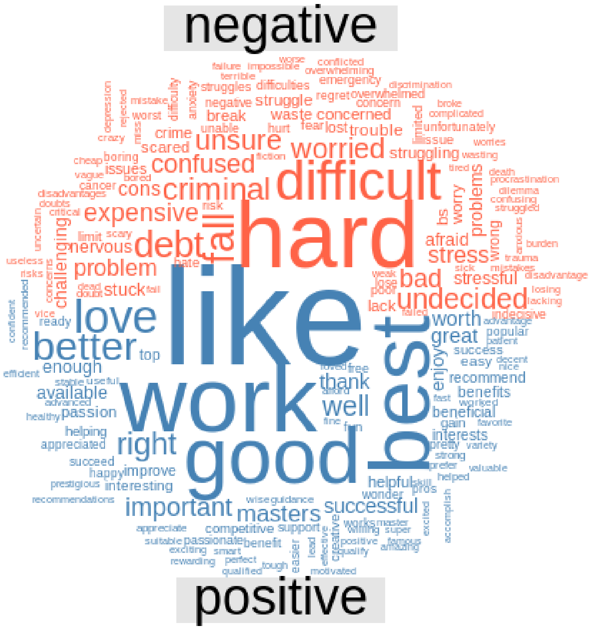


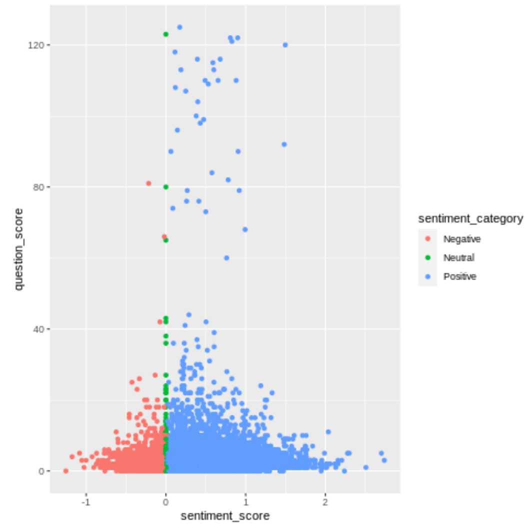
No strong correlations were found between similarity measures of question/answer text and tags.

- Sentiment Analysis of Questions and Answers

We observe that moderately positive and neutral sentimental tone in questions and answers tend to get scored higher.

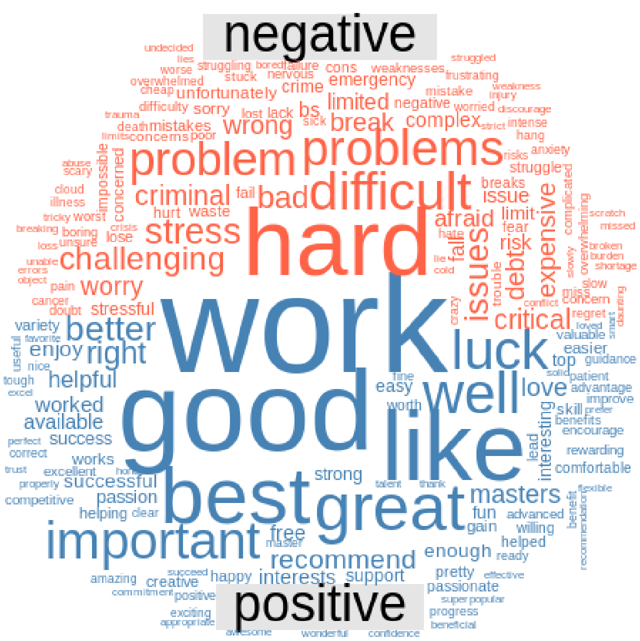
Questions:
Students talk positively about finding the best career or job but they show negative emotion while talking about student debt, the difficult or hard career choices, stress during studies, etc.

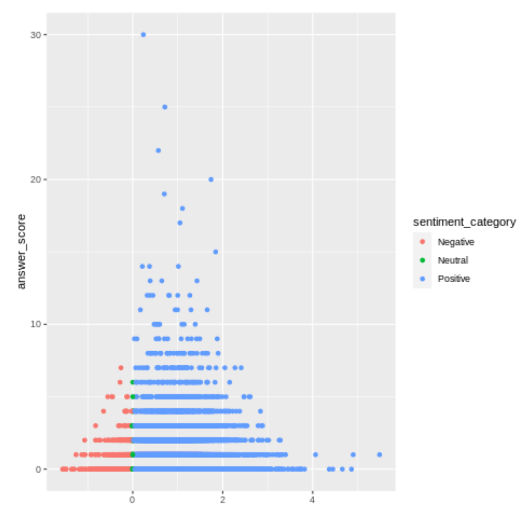




Answers:

We see similar sentiment in answers and professionals respond with empathy. Positive responses tend to receive higher scores.

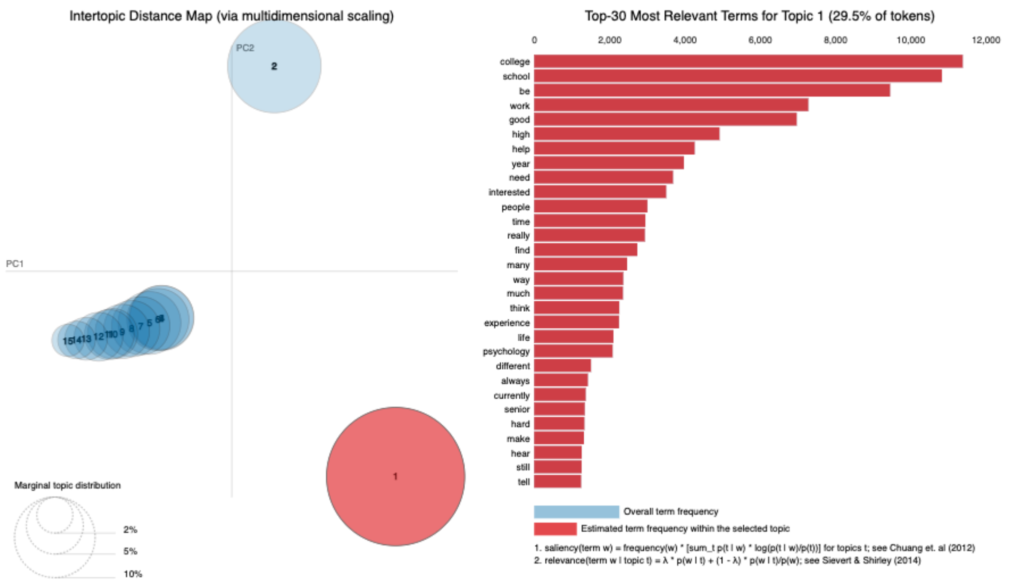




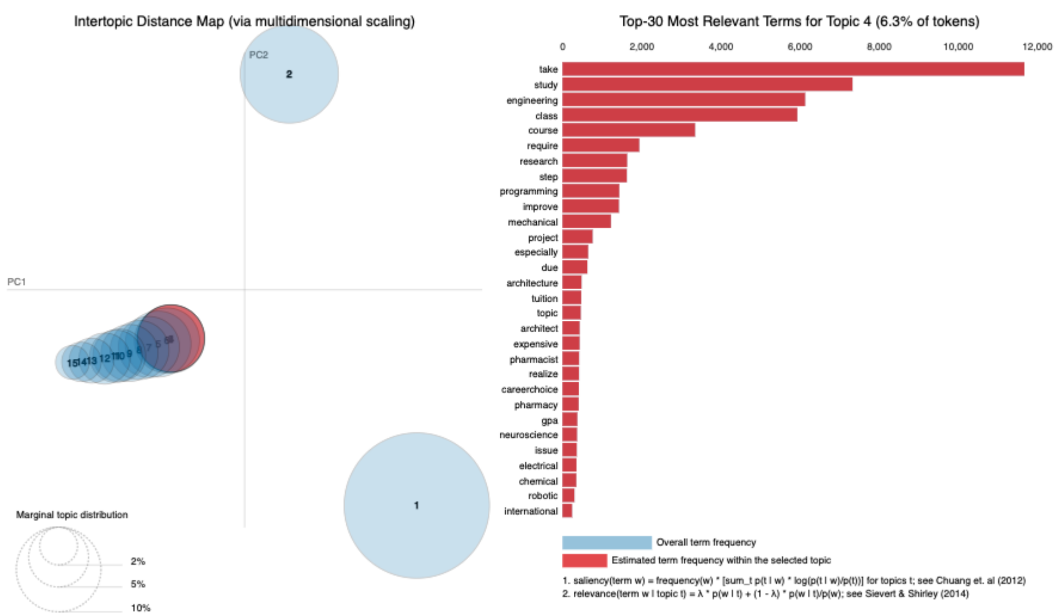
- Topic Modeling of Questions and Answers

We observe that the results of topic modeling have some overlapping contexts of topics, but each of them talks about a separate context.

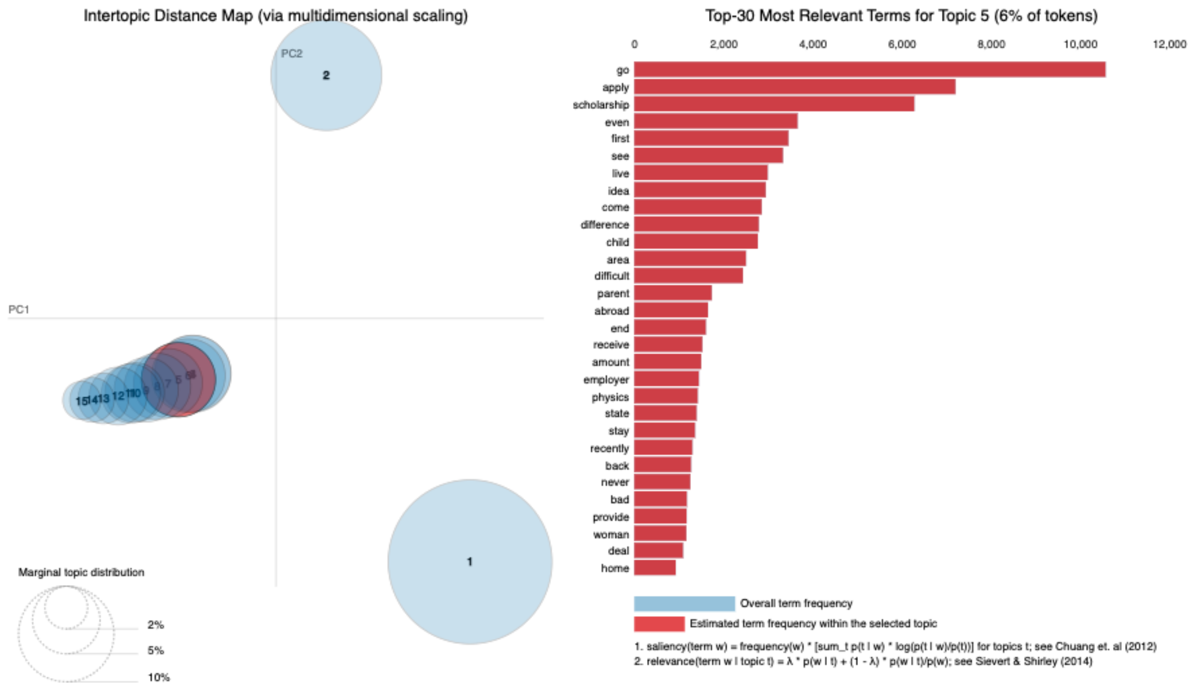
Questions:



Topic 1: Starting a career in business, finance, or management. The top terms in this topic are listed above.



Topic 4: Applying for jobs. The top terms are related to the field of study.



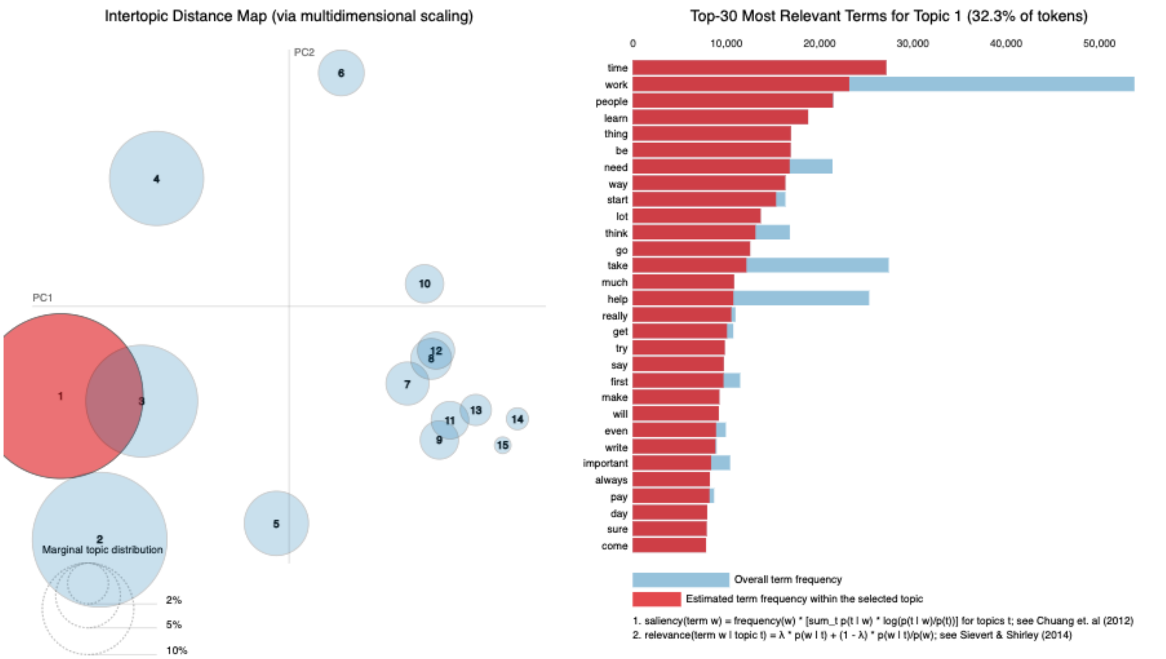
Topic 5: Computer Science related topics and need for scholarships. The top terms are listed above and applying for scholarships is one of them.

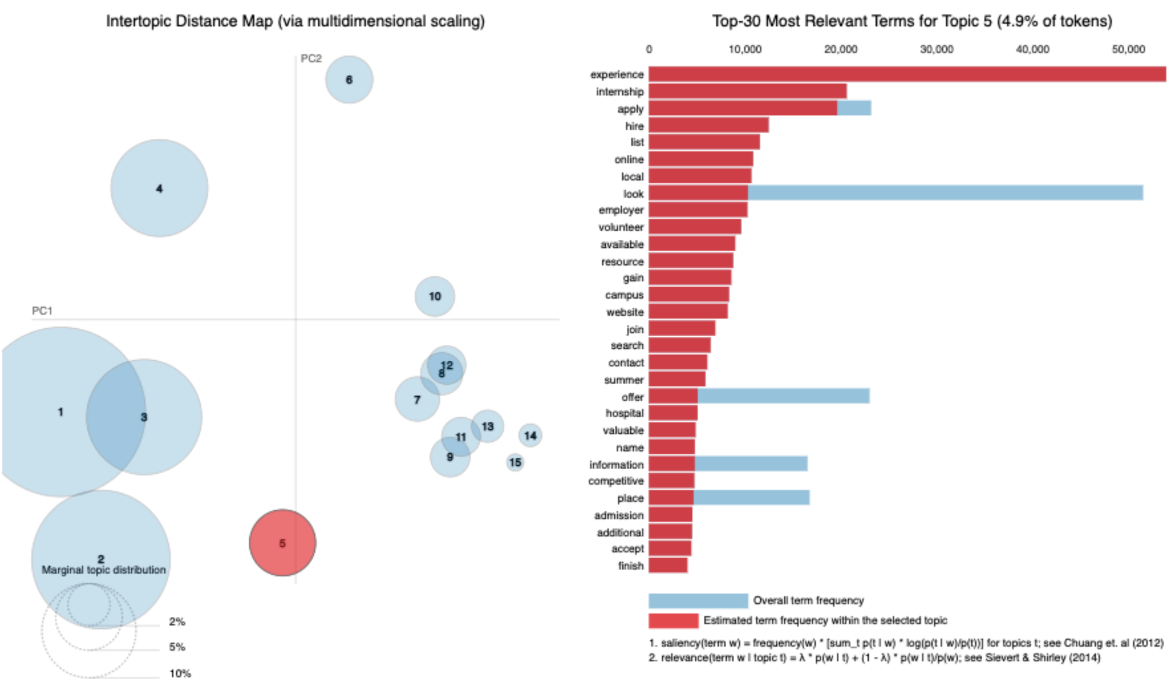
Main terms that define each topic are:

```
[(0,
  '0.134*"business" + 0.130*"start" + 0.112*"program" + 0.053*"management" + '
  '0.043*"finance" + 0.035*"summer" + 0.034*"subject" + 0.028*"accept" + '
  '0.027*"account" + 0.026*"accounting"'),
 (1,
  '0.500*"job" + 0.115*"well" + 0.057*"use" + 0.041*"long" + 0.028*"know" + '
  '0.028*"relate" + 0.027*"actually" + 0.026*"follow" + 0.026*"hire" + '
  '0.024*"salary"'),
 (2,
  '0.116*"go" + 0.079*"apply" + 0.069*"scholarship" + 0.040*"even" + '
  '0.038*"first" + 0.037*"see" + 0.033*"live" + 0.032*"idea" + 0.031*"come" + '
  '0.031*"difference"'),
 (3,
  '0.121*"thing" + 0.090*"may" + 0.084*"able" + 0.061*"important" + '
  '0.049*"order" + 0.047*"community" + 0.038*"resume" + 0.036*"answer" + '
  '0.033*"application" + 0.032*"position"'),
 (4,
  '0.296*"major" + 0.131*"science" + 0.117*"computer" + 0.042*"opportunity" + '
  '0.038*"offer" + 0.034*"specific" + 0.031*"minor" + 0.027*"chance" + '
  '0.025*"language" + 0.020*"collegemajor"'),
 (5,
  '0.088*"college" + 0.084*"school" + 0.073*"be" + 0.056*"work" + 0.054*"good" '
  '+ 0.038*"high" + 0.033*"help" + 0.031*"year" + 0.028*"need" + '
  '0.027*"interested"'),
 (6,
  '0.105*"engineer" + 0.092*"love" + 0.067*"learn" + 0.053*"technology" + '
  '0.051*"math" + 0.044*"company" + 0.040*"skill" + 0.038*"art" + 0.030*"game" '
  '+ 0.027*"software"'),
 (7,
  '0.161*"internship" + 0.121*"lot" + 0.079*"teach" + 0.078*"say" + '
  '0.073*"professional" + 0.051*"better" + 0.049*"communication" + '
  '0.040*"teaching" + 0.026*"website" + 0.023*"hour"'),
 (8,
  '0.183*"career" + 0.065*"degree" + 0.057*"look" + 0.040*"graduate" + '
  '0.040*"wonder" + 0.035*"get" + 0.033*"sure" + 0.033*"want" + 0.028*"plan" + '
  '0.028*"pursue"'),
 (9,
  '0.167*"field" + 0.092*"become" + 0.071*"medical" + 0.066*"nursing" + '
  '0.062*"biology" + 0.059*"nurse" + 0.058*"interest" + 0.050*"medicine" + '
  '0.049*"path" + 0.042*"travel"'),
 (10,
  '0.119*"sport" + 0.112*"law" + 0.108*"prepare" + 0.086*"lawyer" + '
  '0.051*"play" + 0.051*"expect" + 0.042*"old" + 0.039*"certain" + '
  '0.032*"begin" + 0.029*"base"'),
 (11,
  '0.128*"student" + 0.069*"money" + 0.069*"future" + 0.063*"ask" + '
  '0.057*"pay" + 0.057*"question" + 0.043*"give" + 0.029*"would" + '
  '0.029*"family" + 0.029*"marketing"'),
 (12,
  '0.211*"take" + 0.132*"study" + 0.111*"engineering" + 0.107*"class" + '
  '0.061*"course" + 0.035*"require" + 0.030*"research" + 0.029*"step" + '
  '0.026*"programming" + 0.026*"improve"'),
 (13,
  '0.101*"teacher" + 0.083*"education" + 0.070*"s" + 0.067*"thank" + '
  '0.064*"day" + 0.041*"curious" + 0.040*"part" + 0.039*"world" + 0.038*"kid" '
  '+ 0.031*"helpful"'),
```

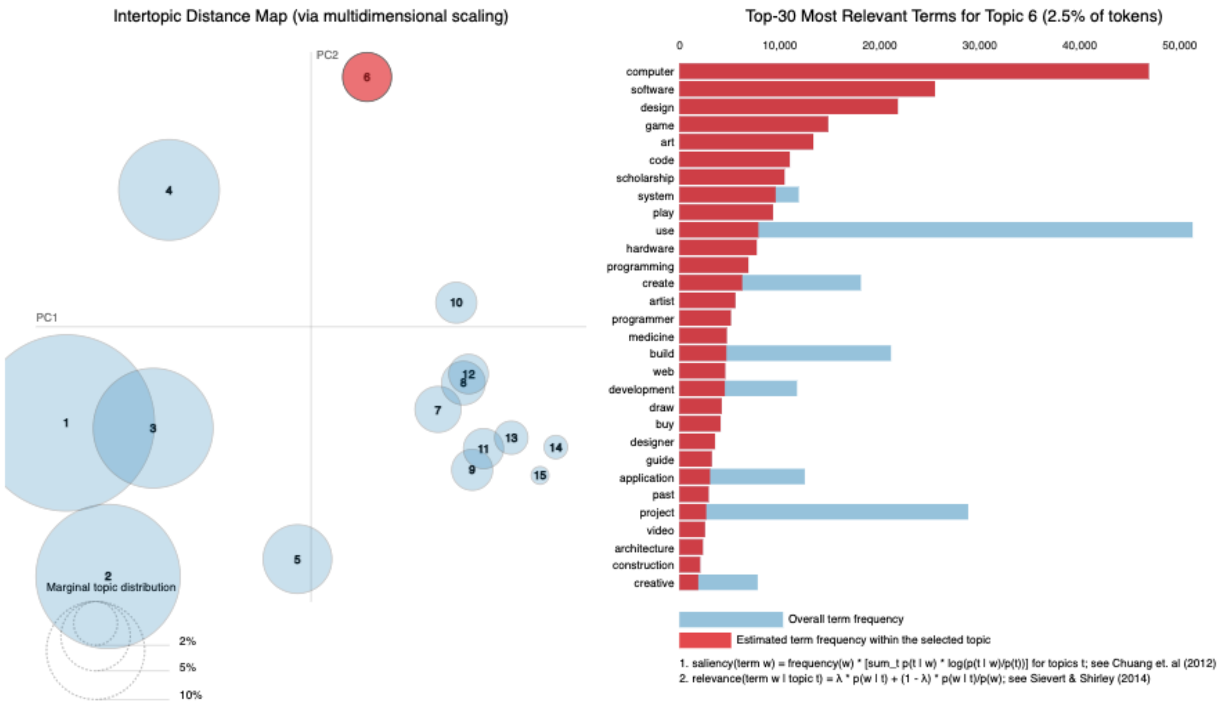
Answers:

Answers were also categorized into 15 topics. We see some overlap in the topics discussed with questions, but terms for answers are more towards providing help.





Topic 5 talks about answering to scholarship and internship application process



Topic 6 talks about software skills

```
[(0,
  '0.217*"engineer" + 0.162*"engineering" + 0.081*"law" + 0.044*"firm" + '
  '0.042*"salary" + 0.038*"electrical" + 0.037*"accounting" + '
  '0.037*"mechanical" + 0.033*"determine" + 0.023*"lawyer"'),
 (1,
  '0.134*"experience" + 0.051*"internship" + 0.049*"apply" + 0.031*"hire" + '
  '0.029*"list" + 0.027*"online" + 0.027*"local" + 0.026*"look" + '
  '0.025*"employer" + 0.024*"volunteer"'),
 (2,
  '0.082*"problem" + 0.060*"basic" + 0.055*"client" + 0.049*"issue" + '
  '0.038*"material" + 0.032*"transfer" + 0.032*"necessary" + 0.032*"scientist" '
  '+ 0.031*"process" + 0.029*"talent"'),
 (3,
  '0.100*"math" + 0.099*"medical" + 0.088*"music" + 0.086*"test" + '
  '0.050*"doctor" + 0.043*"nursing" + 0.039*"exam" + 0.034*"specialty" + '
  '0.028*"practice" + 0.028*"physics"'),
 (4,
  '0.161*"computer" + 0.088*"software" + 0.075*"design" + 0.051*"game" + '
  '0.046*"art" + 0.038*"code" + 0.036*"scholarship" + 0.033*"system" + '
  '0.032*"play" + 0.027*"use"'),
 (5,
  '0.056*"good" + 0.050*"job" + 0.042*"may" + 0.042*"career" + 0.034*"many" + '
  '0.030*"work" + 0.024*"field" + 0.023*"look" + 0.022*"help" + '
  '0.022*"different"'),
 (6,
  '0.082*"school" + 0.063*"college" + 0.051*"degree" + 0.048*"year" + '
  '0.042*"program" + 0.034*"take" + 0.032*"student" + 0.032*"class" + '
  '0.031*"course" + 0.025*"high"'),
```

```
 (7,
  '0.093*"live" + 0.091*"friend" + 0.075*"family" + 0.054*"add" + 0.050*"home" '
  '+ 0.034*"post" + 0.033*"away" + 0.033*"close" + 0.031*"bad" + '
  '0.029*"woman"'),
 (8,
  '0.048*"company" + 0.044*"skill" + 0.036*"work" + 0.022*"position" + '
  '0.021*"project" + 0.020*"professional" + 0.017*"provide" + 0.017*"role" + '
  '0.016*"interview" + 0.016*"team"'),
 (9,
  '0.268*"question" + 0.177*"ask" + 0.137*"answer" + 0.048*"have" + '
  '0.043*"figure" + 0.040*"share" + 0.034*"link" + 0.018*"simple" + '
  '0.015*"thank" + 0.015*"core"'),
 (10,
  '0.028*"time" + 0.024*"work" + 0.022*"people" + 0.020*"learn" + '
  '0.018*"thing" + 0.018*"be" + 0.017*"need" + 0.017*"way" + 0.016*"start" + '
  '0.014*"lot"'),
 (11,
  '0.188*"business" + 0.078*"management" + 0.073*"marketing" + '
  '0.067*"technology" + 0.057*"communication" + 0.051*"product" + '
  '0.041*"finance" + 0.033*"run" + 0.030*"aspect" + 0.028*"tool"'),
 (12,
  '0.105*"resume" + 0.077*"kid" + 0.056*"intern" + 0.056*"schedule" + '
  '0.048*"nurse" + 0.038*"semester" + 0.031*"communicate" + 0.029*"wait" + '
  '0.028*"plenty" + 0.026*"walk"'),
 (13,
  '0.208*"psychology" + 0.091*"note" + 0.080*"datum" + 0.079*"perform" + '
  '0.050*"psychologist" + 0.045*"user" + 0.041*"feedback" + 0.038*"evaluate" + '
  '0.035*"app" + 0.034*"return"'),
```

Logistic Regression

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2370  -0.9541  -0.8532   1.3191   2.1672

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.270e+00  9.887e-02 -12.844  < 2e-16 ***
tag_count               4.708e-02  1.013e-02   4.647 3.37e-06 ***
emails_sent_count      -5.462e-04  8.966e-05  -6.092 1.11e-09 ***
Business1               7.271e-01  9.183e-02   7.918 2.41e-15 ***
Career_Couns1           4.768e-01  6.823e-02   6.989 2.77e-12 ***
College1                2.482e-01  5.485e-02   4.525 6.04e-06 ***
Engineering1            4.043e-01  7.755e-02   5.214 1.85e-07 ***
Law_enf1                2.658e-01  1.010e-01   2.632 0.008477 **
Nursing_and_Med1       -5.016e-01  7.273e-02  -6.896 5.34e-12 ***
Professional1          -1.447e+00  2.108e-01  -6.866 6.59e-12 ***
Psych1                  4.531e-01  1.094e-01   4.142 3.45e-05 ***
Social_Work1            2.828e-01  1.228e-01   2.303 0.021294 *
Scientist1             -1.070e+00  3.177e-01  -3.368 0.000757 ***
school1                 1.778e-01  8.922e-02   1.993 0.046303 *
Teaching1              -3.140e-01  1.123e-01  -2.795 0.005182 **
Technology1             7.246e-01  9.495e-02   7.631 2.33e-14 ***
question_tag_sim_index  6.455e-01  1.461e-01   4.420 9.89e-06 ***
question_score          7.693e-02  6.720e-03  11.447  < 2e-16 ***
sentiment_score         7.716e-02  4.593e-02   1.680 0.092984 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21458  on 16302  degrees of freedom
Residual deviance: 20860  on 16284  degrees of freedom
AIC: 20898
```

- All variables are significant except for sentiment score and all significant non-dummy variables have positive coefficients except for emails sent
- Business, technology, and engineering questions are answered fastest, while science, nursing and medicine, and teaching response times and answer rates should be improved

Linear Regression

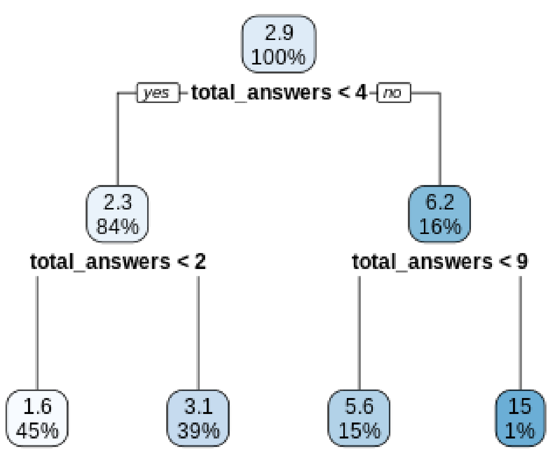
Total number of answers for a given question is dependent on the question score and the response time taken to answer a question according to the results of the Linear Regression.

Adjusted R square for this model is 0.508. P value for number of tags, number of emails sent and sentiment score is greater than 0.05, hence these variables are statistically insignificant in predicting the number of answers for a particular question.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          total_answers   R-squared:                       0.508
Model:                            OLS   Adj. R-squared:                  0.508
Method:                 Least Squares   F-statistic:                     7670.
Date:                Mon, 09 May 2022   Prob (F-statistic):               0.00
Time:                        14:19:18   Log-Likelihood:                -89360.
No. Observations:               37118   AIC:                         1.787e+05
Df Residuals:                   37112   BIC:                         1.788e+05
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 1.8624      0.034     54.155      0.000       1.795       1.930
question_score        0.4039      0.002    190.440      0.000       0.400       0.408
tag_count             0.0018      0.006      0.317      0.751      -0.009       0.013
emails_sent_count -1.672e-05   5.99e-05     -0.279      0.780      -0.000       0.000
sentiment_score       0.0052      0.027      0.191      0.848      -0.049       0.059
diff_hours           -0.0001    5.2e-06    -20.210      0.000      -0.000   -9.5e-05
==============================================================================
Omnibus:                    36013.552   Durbin-Watson:                   0.466
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         18033300.931
Skew:                          -3.922   Prob(JB):                         0.00
Kurtosis:                     110.697   Cond. No.                     8.35e+03
==============================================================================
```

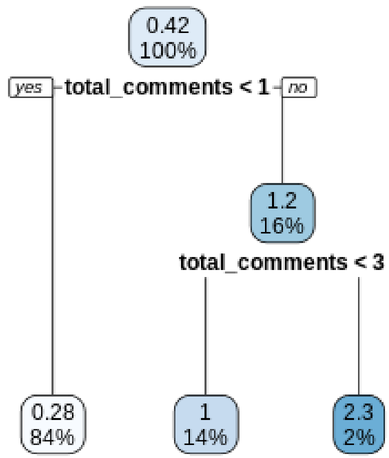
Decision Trees

The decision tree for Question score has an RMSE of 4.688.



We observe that more answers received on the question tend to give it a higher score.

The decision tree for Answer score has an RMSE of 0.85.



Similar to questions, we see a higher engagement on answers through comments tends to give it a higher score. However, a majority of answers might not have comments (84%) which is shown in the first node. This indicates that comments count might not be the best predictor of answer score and further analysis should be performed.

Conclusions

Summarizing the Exploratory Data Analysis and Modeling, we conclude that :

- Scoring on questions and answers is dependent on the activity level
- Professionals who joined recently are more likely to post to answers
- Business, technology, and engineering questions are answered fastest, while science, nursing and medicine, and teaching response times and answer rates should be improved
- Questions are sent in emails after they go a certain period of time unanswered
- The number of answers received is dependent on the question score and the time taken to respond after an email is sent.

Suggestions

We have the below suggestions for CareerVillage.org to improve their platform :

- Promote the platform to other industry professionals and encourage them to subscribe to immediate notifications to have faster time-to-response for those fields.
- Promote the platform in locations like India and Egypt that have a high student base to get faster response times.
- Optimize the platform and have more professionals available during the high activity time.
- Improve social media presence to increase the student and professional user base.

Resources & Code

GitHub Repository: https://github.com/nthammadi-uncc/CareerVillage