Stack Overflow Statistical Analysis

Areej Althubaity, Malak Alsabban, Misbah Khan
Department of Computer Science and Engineering
University of Connecticut, Storrs, USA
{areej.althubaity, malak.alsabban, misbah.khan}@uconn.edu

***Abstract*— 'Stack Overflow' is a question and answer site for professional and enthusiast programmers. This paper involves the analysis of the Stack Overflow website statistics, which mainly deals with measures such as questions, number of answers and comments, scores, views, tags, details about users, user's reputation, etc.**

***Index Terms*— Mining software repositories, statistical analysis, online community, Stack Overflow.**

I. Introduction

Stack Overflow[1] (SO) allows programmers to ask and answer questions related to their work. The site discourages opinion-based questions that are likely to generate discussions rather than objective answers. Questions that do not meet the community's criteria will be down voted and may be removed by moderators. Questions are tagged by the asker as being related to at least one and up to five topics. Answers accumulate votes and can also be accepted by the original question asker independently of how many up or down votes it has accumulated. Users gain reputation points when their questions and answers are up-voted. As users gain reputation they gain more and more privileges until they become site moderators.

SO has become an indispensable reference for professional and hobbyist programmers alike. A search for any programming related question will inevitably lead you to a SO post and when it does not, you can post your own question which will likely be answered very quickly by a member of the community. This makes SO a great resource for anyone interested in studying programmers and programming across disciplines.

In this paper, we are interested in analyzing the behavior of the site's users to gain insights into the behavior of programmers in general. Hence, we will perform statistics methodologies to describe measures like number of questions, answers, comments, views, details about users, users' ratings, etc.

II. Data and approach

Our study aims at answering the following seven research questions (RQ):

- **RQ$_1$**: *What is the distribution for number of answers and views per question, number of comments per answer and question, score and reputation per question, answer and comment, and the answerer's reputation per accepted and non-accepted answer?* Our interest with this RQ is to plot and discuss the best-fit distribution for the mentioned attributes.
- **RQ$_2$**: *Do questions and answers have similar distribution of score and number of comments?* This RQ aims to investigate if both data have the same shape of distribution that is if they both have common distribution.
- **RQ$_3$**: *Is there a relationship between a question's score and the number of its answers? Is there a relationship between a question's score and the number of comments it receives?* RQ$_3$ is mainly more about exploring if there exists a relationship between these attributes.
- **RQ$_4$**: *Does the reputation of users who ask questions, answer questions, or leave comments tend to be significantly different?* Our goal here is to find out whether the reputations among questions, answers and comments have the same or common distribution.
- **RQ$_5$:** *Compare and discuss the distributions of answers that were accepted and those that were not?* The focus here is to explore and to compare the distributions of number of comments, number of views, score, and users' reputation in both accepted and non-accepted answers separately.
- **RQ$_6$**: *Is user reputation and score strongly correlated with an answer being accepted?* This RQ is to find the correlation and the relationship between the user reputation and answer's score for accepted answers.
- **RQ$_7$**: *Are there certain tags that tend to result in more answers or comments? Are there any that tend to attract users with particularly high or low reputation?* Our goal here is to discuss any interesting patterns we saw in the different tags.

To answer every RQ we ran some tests and studied in depth the findings we acquired from every analysis.

A. Data Extraction Process

We used posts, comments, users and tags schemas from the official Stack Overflow dump that was available as a PostgreSQL dump for the MSR 2013 Mining challenge [1]. Based on the type of posts, we extracted questions and answers as tables. We ended up dealing with four main .CSV files, which are questions, answers, comments, and tags since our

[1] http://stackoverflow.com/

analysis is based on studying these criteria. Every file has the following attributes/variables- number of views, number of answers, number of comments, user's reputation, and tags. We chose to perform the statistical and graphical analysis over the collected samples through R programming language. We coded the statistical analysis script with the help of RStudio[2], which is an open source, and free IDE for R language. We enhanced our analysis by removing all the noisy data that contain missing values that is all the rows that include Not Applicable (NA) entries were not considered in our analysis. However, we found that the overall number of questions in SO is 3453741 with 6858132 answers and 13252466 comments.

B. Analysis Methods for RQ_1

As mentioned previously, RQ_1 aims to find the best-fit distribution for several attributes. To answer this RQ, we first found the most applicable distribution for every variable/attributes in every data frame by using the fit distribution function (fitdistr) in R. Then, we compared the log-likelihoods values of different distributions; the distribution with the largest value is the best fit one for the data. Then, we tested if the attributes have normal distribution or not by running several tests. Starting with the Skewness test that is mainly a measure of the symmetry of the probability distribution of a real-valued variable about its mean. The skewness value can be positive, negative or zero. A positive skewness denotes that the mean of the data is larger than the median, thus the distribution of the data is right-skewed. A negative skewness value shows that the mean of the data values is less than the median, thus the data distribution is left-skewed. A zero value refers to normal distribution. Moreover, we plotted the Quantile-to-Quantile distribution (Q-Q) for each attribute, which reveals if all the points in the addressed data follow the normality line or they are skewed to either direction. Another test we performed is the Kurtosis, which is used as a measure of peakedness (or flatness) of the data relative to normal distribution. A positive kurtosis value indicates a peaked distribution and a negative kurtosis indicates a flat distribution. One of the easiest ways to check if the data has normal distribution is by checking if the histogram has the smooth bell curve shape. Table 1 summarizes all the findings for RQ_1 attributes along with their percentiles values. We can see that all the attributes have high peaked with right-skewed distribution; thus none of them follow a normal distribution. Unfortunately, we could not include all the Q-Q plots and the histograms figures due to the limited number of pages.

C. Analysis Methods for RQ_2

In RQ_2, we tried to find the similarity in the distributions of questions and answers, in terms of their score and number of comments. The first hypothesis test we applied is the Kolmogorov-Smirnov Test, which calculates the null distribution under the null hypothesis that the two data samples are drawn from the same distribution. The test provided two significant values. The first one is the maximum difference (D) between the two samples' probability mass function. The second value is p-value, which helps in determining whether we can reject or accept the null or the alternative hypothesizes. Hence, if the p-value is greater than or equal to the significance level of 0.05, then we failed to reject the null hypothesis that is the two data have a common distribution. Otherwise, the alternative hypothesis is applied and they both drawn from different distribution. The second test is called Wilcoxon Rank Sum or known as Mann-Whitney that has been used to determine if two datasets are having the same distribution. Again, this test provides the p-value and another value W that corresponds to the sum of ranks assigned to the differences with positive sign. Like Kolmogorov-Smirnov Test, the p-value helps in deciding if the null hypothesis could be applied to the given two data sets. As presented in Table 2, the p-value for both tests is less than our assumed significance level of 0.05; thus we fail to reject the alternative hypothesis and concluded that there is evidence in the data to suggest that score and number comments in questions and answers have different distributions.

TABLE I. SUMMARY OF RQ_1 RESULTS FOR ALL THE ATTRIBUTES

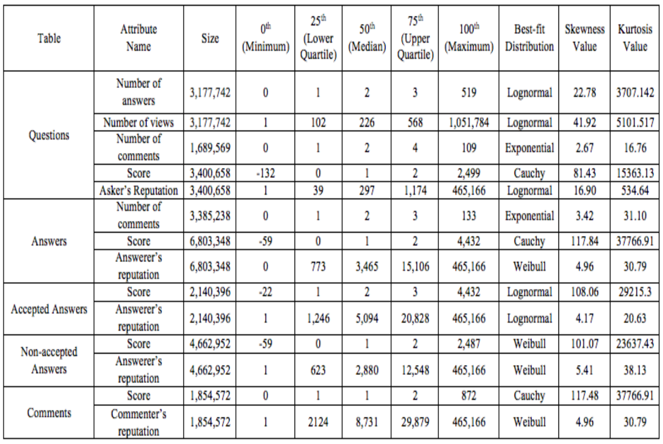
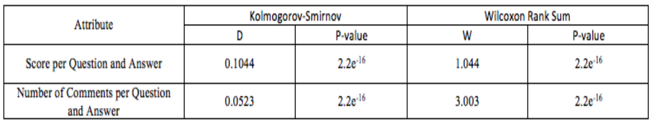


Table	Attribute Name	Size	0th (Minimum)	25th (Lower Quartile)	50th (Median)	75th (Upper Quartile)	100th (Maximum)	Best-fit Distribution	Skewness Value	Kurtosis Value
Questions	Number of answers	3,177,742	0	1	2	3	519	Lognormal	22.78	3707.142
	Number of views	3,177,742	1	102	226	568	1,051,784	Lognormal	41.92	5101.517
	Number of comments	1,689,569	0	1	2	4	109	Exponential	2.67	16.76
	Score	3,400,658	-132	0	1	2	2,499	Cauchy	81.43	15363.13
	Asker's Reputation	3,400,658	1	39	297	1,174	465,166	Lognormal	16.90	534.64
Answers	Number of comments	3,385,238	0	1	2	3	133	Exponential	3.42	31.10
	Score	6,803,348	-59	0	1	2	4,432	Cauchy	117.84	37766.91
	Answerer's reputation	6,803,348	0	773	3,465	15,106	465,166	Weibull	4.96	30.79
Accepted Answers	Score	2,140,396	-22	1	2	3	4,432	Lognormal	108.06	29215.3
	Answerer's reputation	2,140,396	1	1,246	5,094	20,828	465,166	Lognormal	4.17	20.63
Non-accepted Answers	Score	4,662,952	-59	0	1	2	2,487	Weibull	101.07	23637.43
	Answerer's reputation	4,662,952	1	623	2,880	12,548	465,166	Weibull	5.41	38.13
Comments	Score	1,854,572	0	1	1	2	872	Cauchy	117.48	37766.91
	Commenter's reputation	1,854,572	1	2124	8,731	29,879	465,166	Weibull	4.96	30.79

TABLE II. SIMIARITY OF QUESTIONS AND ANSWERS DISTRIBUTION OF SCORE AND NUMBER OF COMMENTS



Attribute	Kolmogorov-Smirnov		Wilcoxon Rank Sum	
	D	P-value	W	P-value
Score per Question and Answer	0.1044	$2.2e^{-16}$	1.044	$2.2e^{-16}$
Number of Comments per Question and Answer	0.0523	$2.2e^{-16}$	3.003	$2.2e^{-16}$

D. Analysis Methods for RQ_3

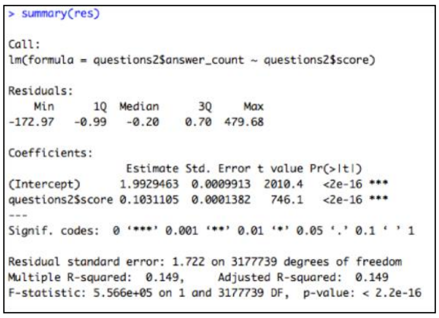
To find the relationship between the distribution of score per question and its number of answers and its number of comments, we first studied the nature of this relation by finding the Pearson Correlation Coefficient between the two variables. The correlation illustrates whether the two variables have a linear relation that is we wanted to see in which degree they are moving together, if applicable. When the correlation is equal to +1 that means we have a strong positive linear relation. If it is 0 then there is no relationship at all. When it is equal to -1 that means there is a negative linear relationship. Regarding the question's score and its number of answers, the correlation is equal to 0.386 that means their relationship is

[2] https://www.rstudio.com/

weak and positively linear since it is close to 1; we could conclude that the variables are linearly related. Also, the correlation value of score and number of comments per question is a positive small value equals to 0.0563; thus they have a very weak linear relationship. Moreover, we found out from the covariance how the distributions of score and number of answers in questions are linearly related. In other words, the covariance value was a positive number (= 5.394) which indicates that the relationship is a positive one; hence if the questions have a high score, mostly it will have a high number of answers, and vice versa. Similarly, the distributions of score and number of comments in questions have a positive linear relationship according to the covariance value (= 1.151). Lastly, we performed the linear regression model to examine the relationship between the distributions by using the linear model function (lm) in R. In this RQ, we would like to find how the score of a question could affect the number of its answers and its comments. Thus, we setup the explanatory variable to be the score of a question and the number of answers or comments to be the response variable. From the previous model, the regression equations that represent such relation are:

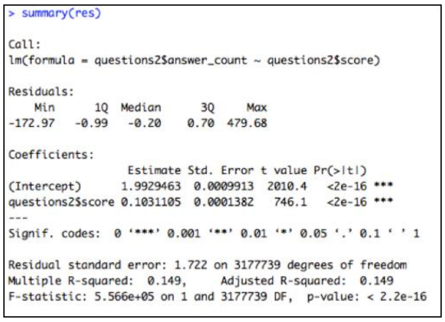
$$\text{Number of Answers/Question} = 0.10\ (\text{Score}) + 1.99 \qquad (1)$$

$$\text{Number of Comments/Question} = 0.015\ (\text{Score}) + 2.68 \qquad (2)$$



```
> summary(res)

Call:
lm(formula = questions2$answer_count ~ questions2$score)

Residuals:
    Min      1Q  Median      3Q     Max 
-172.97   -0.99   -0.20    0.70  479.68 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.9929463  0.0009913  2010.4   <2e-16 ***
questions2$score 0.1031105  0.0001382   746.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.722 on 3177739 degrees of freedom
Multiple R-squared:  0.149,	Adjusted R-squared:  0.149 
F-statistic: 5.566e+05 on 1 and 3177739 DF,  p-value: < 2.2e-16
```

Fig. 1. Statistical analysis of the distribution question’s score and its number of answers



```
> summary(res)

Call:
lm(formula = questions2$answer_count ~ questions2$score)

Residuals:
    Min      1Q  Median      3Q     Max 
-172.97   -0.99   -0.20    0.70  479.68 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.9929463  0.0009913  2010.4   <2e-16 ***
questions2$score 0.1031105  0.0001382   746.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.722 on 3177739 degrees of freedom
Multiple R-squared:  0.149,	Adjusted R-squared:  0.149 
F-statistic: 5.566e+05 on 1 and 3177739 DF,  p-value: < 2.2e-16
```

Fig. 2. Statistical analysis of the distribution-question’s score and its number of comments

To test the ‘statistical significance’ of the previous equations, we used the summary function in R. By analyzing the information given Fig.1, the residuals are the difference between the actual values of the variable we are predicting and predicted values from our regression. The significant stars represents the significance level with the number of asterisks displayed according to the p-value computed (where * stands for low and *** for high). In our case, *** assures that there is a relationship between the question’s score and the number of its answers. The estimated coefficient is the slope, which is calculated by the regression. The standard error of the coefficient estimate is used to measure of the variability in the coefficient estimate, and hence should be as less as possible, as compared to the magnitude of the coefficient estimate. In our case, the standard error value of score is 0.00013, which is 1000x less than the estimate coefficient. The T-value of the estimate coefficient is used to calculate the p-value and the significance level. The variable p-value, is the probability that the variable is not relevant; the smaller this value the better it means. In our scenario, 2.2e^{-16} means that the odds that score is meaningless is less than $^{1}/_{5000000000000000}$. Residual estimated error is the standard deviation of our residuals. Degree of freedom is the difference in the number of observations in our dataset and the number of variables used in our model. R-squared is a measurement to evaluate the goodness of fit of our model, the more this value is closer to 1, the better. Hence, in our case ~14% of the increases in number of answers for a particular question is because of its score. The F-statistic means that the more parameter we have in the model, the higher p-value we get. However, in our case we only have one parameter (score) and thus we got low p-value. The degree of freedom (DF) shows the number of parameters in the model and we only have one variable (score) in our case study. However, the interpretation of Fig.2 is the same as Fig.1 with some slight changes regarding the following values: the standard error of the coefficient estimate value of score is 0.00020, which is 100x less than the estimate coefficient. The R-squared value means that ~0.37% of the increases in number of comments for a particular question are because of its score.

E. Analysis Methods for RQ$_4$

To answer the question if user’s reputation among questions, answers and comments has the same or common distribution, we superimposed the density distributions of users’ reputation in questions, answers, and comments one on top of the other as in Fig.3, in order to better understand the range of the data with respect to each other. Then, we collected statistical information about the user reputation for each of the three sets. The results obtained are represented in Table 3.

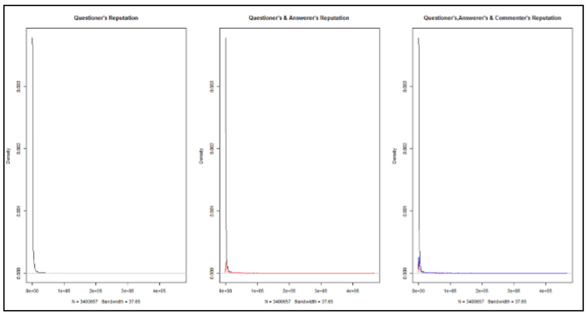


Fig. 3. The density distributions of users’ reputation in questions, answers, and comments plotted over each other.

TABLE III. STATISTICAL SUMMARY FOR USER REPUTATION IN QUESTIONS, ANSWERS, AND COMMENTS

Attribute	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
User reputation/question	1	39	297	1764	1174	465200
User reputation/answer	1	773	3465	19970	15110	465200
User reputation/comment	1	408	2258	17220	12440	465200

We came up with the following findings:

- The reputation of users ranges from about 1 to 465,200.
- The reputation of users who ask questions tends to be generally on the lower end of the spectrum. This could correspond to those users who are not active members of stackoverflow.com, but rather use this website only to seek help from the community.
- Furthermore, it can also be seen that the reputation of users who answer questions tends to be higher than the users who comment.

Lastly, we performed the Kolmogorov-Smirnov Test and we concluded from the tests values in Table 4 that the probability of the three attributes belonging to the same distribution is very minimal. That is, the three attributes come from different distributions. Moreover, the user reputation of answerers and commenters is comparatively closer to each other, and are far from questioners.

F. *Analysis Methods for RQ_5*

For RQ_5, we divided the answers into two sets, one set representing the accepting answers and another representing the unaccepted answers. We applied the same similarity tests done in RQ_4. First, we plotted the distribution of variables of accepted answers in one row, and their corresponding equivalents in unaccepted answers below it as in Fig.4. Then, we analyzed the statistical summary as in Fig.5 on both data sets and found that:

- The reputation of users in case of accepted answers is much better than those of unaccepted answers.
- The score in case of accepted answers is much better than those of unaccepted answers.
- Other numerical statistics are shown in Fig.5.

TABLE IV. THE KS-TEST RESULTS FOR USER REPUTATION IN THE THREE DATA SETS (QUESTIONS, ANSWERS, & COMMENTS)

	D	P-value
User reputation per question and answer	0.442	$2.2e^{-16}$
User reputation per question and comment	0.355	$2.2e^{-16}$
User reputation per answer and comment	0.087	$2.2e^{-16}$

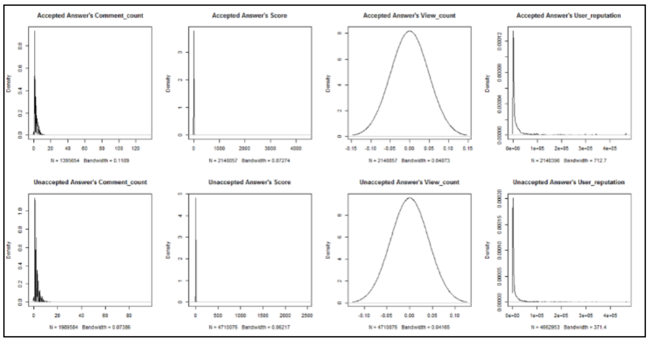
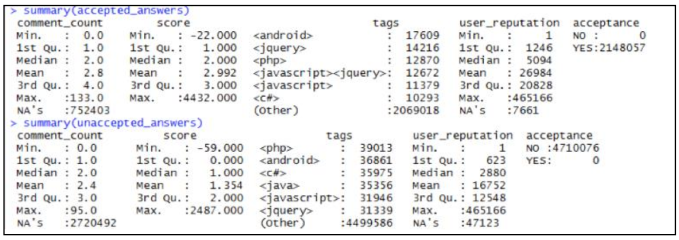


Fig. 4. The distribution of variables of accepted and non-accepted answers



```
> summary(accepted_answers)
 comment_count        score                           tags           user_reputation   acceptance
 Min.    :  0.0    Min.    : -22.000   <android>              :  17609   Min.    :      1   NO :      0
 1st Qu.:  1.0    1st Qu.:    1.000   <jquery>               :  14216   1st Qu.:   1246   YES:2148057
 Median :  2.0    Median :    2.000   <php>                  :  12870   Median :   5094
 Mean   :  2.8    Mean   :    2.992   <javascript><jquery>:  12672   Mean   : 26984
 3rd Qu.:  4.0    3rd Qu.:    3.000   <javascript>           :  11379   3rd Qu.: 20828
 Max.   :133.0    Max.   :4432.000   <c#>                   :  10293   Max.   :465166
 NA's   :752403                        (Other)                :2069018   NA's   :7661
> summary(unaccepted_answers)
 comment_count         score                   tags           user_reputation   acceptance
 Min.    : 0.0      Min.    : -59.000   <php>         :  39013   Min.    :      1   NO :4710076
 1st Qu.: 1.0      1st Qu.:    0.000   <android>     :  36861   1st Qu.:    623   YES:       0
 Median : 2.0      Median :    1.000   <c#>          :  35975   Median :   2880
 Mean   : 2.4      Mean   :    1.354   <java>        :  35356   Mean   : 16752
 3rd Qu.: 3.0      3rd Qu.:    2.000   <javascript>:  31946   3rd Qu.: 12548
 Max.   :95.0      Max.   :2487.000   <jquery>      :  31339   Max.   :465166
 NA's   :2720492                        (Other)       :4499586   NA's   :47123
```

Fig. 5. Statistical summary of all the variables of accepted and non-accepted answers

Finally, we ran the KS test on number of comments, score, number of views and user reputation of accepted answers and unaccepted answers with each other. Some of the inferences that can be derived from this test according to Table 5 are as follows:

- The distribution of views is equally similar among the accepted and unaccepted answers.
- The KS test reports that number of views of accepted answers and unaccepted is from the same distribution and that the largest distance between them would probably be around 0.000000000000007
- The KS test reports that number of comments, user reputation and score seem to come from different distributions. However, the distance between them is not towards extreme.

G. *Analysis Methods for RQ_6*

Recall that RQ_6 aims to find the correlation and the relationship between the user reputation and answer's score for accepted answers. Hence, we performed the three types of similarity tests as in RQ_4 and RQ_5 between the reputation and scores of accepted answers and unaccepted answers separately. Following this, we did a correlation test between user reputation and acceptance field of the main answers data frame. We plotted a grid of the density distributions of score and user reputation of the accepted and unaccepted answers separately, to visually look for similarities as in Fig.6. Then, we collected statistics of user reputation and score of accepted answers, followed by the same for unaccepted answers as in Table 6. We found that the user reputation and score of accepted answers tends towards the higher side of their range. In addition, we ran the KS test between user reputation and score of accepted answers, followed by the same for unaccepted answers. By analyzing the results as shown in Table 7, we drew some findings:

- The KS Test reports that the user reputation of accepted and unaccepted answers does not belong to the same distribution.
- The score of the two types of answers also is returned to be belonging to different distributions with a distance between them.

TABLE V. THE KS-TEST VALUES OF ACCEPTED AND NON-ACCEPTED ANSWERS VARIABLES

	D	P-value
Number of comments	0.079	$2.2e^{-16}$
Score	0.282	$2.2e^{-16}$
Number of views	$7.26e^{-15}$	1
User reputation	0.105	$2.2e^{-16}$

TABLE VI. STATISTICAL SUMMARY OF SCORE AND USER REPUTATION OF ACCEPTED AND NON-ACCEPTED ANSWERS

Attribute	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
User reputation/accepted answer	1	1246	5049	26980	20830	465200
User reputation/non-accepted answer	1	623	2880	16750	12550	462500
Score/accepted answer	-22	1	2	2.99	3	4432
Score/non-accepted answer	-59	0	1	1.35	2	2487

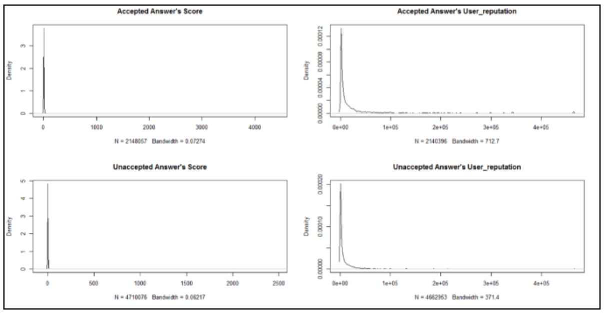


Fig. 6. The density distributions of score and user reputation of accepted and non-accepted answers

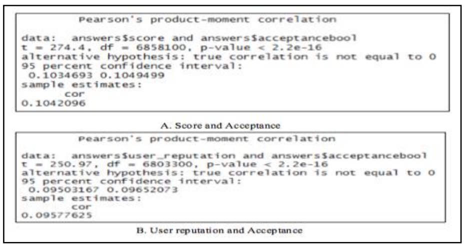
Then, we performed a correlation test between user reputation and score of answers in order to check the probability of the existence of the correlation between the two. For this we used the function (cor.test) in R. From the given results in Fig. 7, we found that there is a very less probability of existence correlation. Hence, it can be said that there are less chances that a correlation between user's reputation and acceptance, and score and acceptance. However, note that this does not mean that there is no correlation between the two.

H. Analysis Methods for RQ$_7$

We first aggregated the data by tags and repeated the SQL queries. We also normalized the tags by removing unnecessary symbols such as "<" and ">" which surrounded every tag. For each tag in questions, answers, and comments we counted its occurrences in every data set, i.e., counting the frequency of each label. We started by looking for tags that tend to appear more frequently in the three data sets, which are questions, answers, and comments. To be more specific, we were interesting to find the top ten tags in the three data sets. According to **Error! Reference source not found.**Fig.8, C# tag is the one which has the highest frequency among all the data sets. The top ten tags are similar in the three data sets, but they tend to have different ranks. In questions, the iPhone tag has the seventh rank where it does not appear in answers and comments tags list. Moreover, Python label has the ninth and the eighth ranks in answers and comments respectively where it is not listed in questions. In comments, the HTML tag has the tenth rank where we could not find it in the tags list of questions and answers.

TABLE VII. THE KS-TEST VALUES OF USER REPUTATION AND SCORE OF ACCEPTED AND UNACCEPTED ANSWERS

	D	P-value
User reputation	0.105	2.2e^{-16}
Score	0.282	2.2e^{-16}



```
          Pearson's product-moment correlation

data:  answers$score and answers$acceptancebool
t = 274.4, df = 6858100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1034693 0.1049499
sample estimates:
      cor
0.1042096
```

A. Score and Acceptance

```
          Pearson's product-moment correlation

data:  answers$user_reputation and answers$acceptancebool
t = 250.97, df = 6803300, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09503167 0.09652073
sample estimates:
       cor
0.09577625
```

B. User reputation and Acceptance

Fig. 7. Correlation test results of (A) score and answer acceptance, and (B) user reputation and answer acceptance

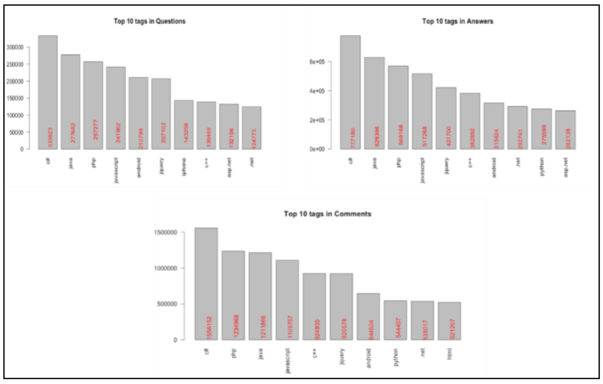
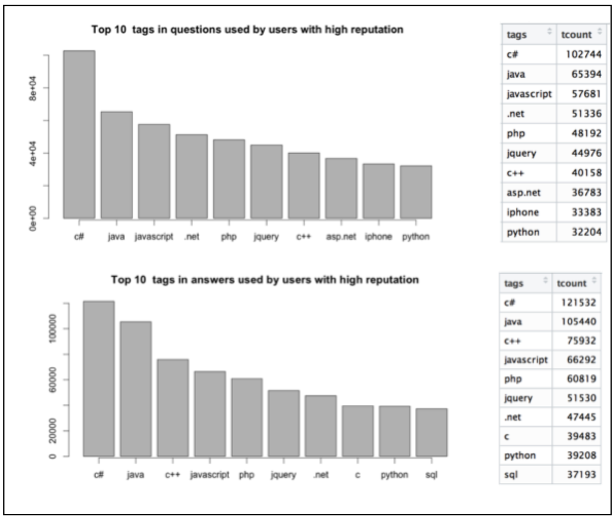


Fig. 8. The top ten tags in the three data sets

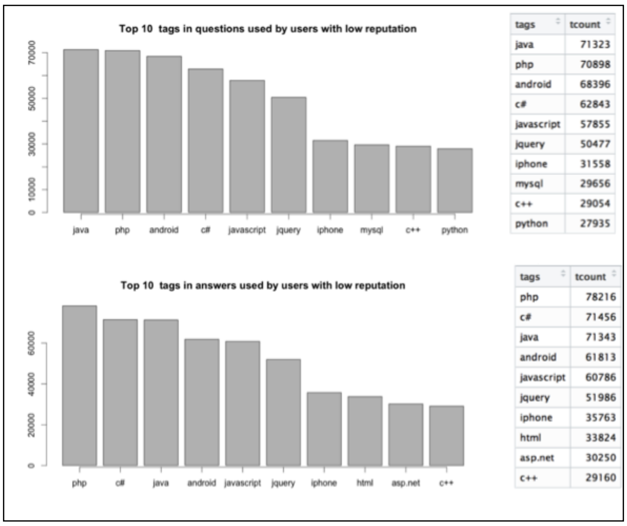
Then, we got a closer look at the top ten tags for the top twenty-five high reputation users in questions and answers. Figure 9 shows that both data sets has C# as the tag with the highest rank which means that users with high reputation tend to tag their questions with this label or they tend to use it more in their answers just because its either their preferable programming language or simply because it's the most common language among most of the developers. Again, the two data sets have the same tags with either identical or different ranks. We could see also that some tags appeared in one data set and not in the other like the iPhone tag in questions and C and SQL in answers.



tags	tcount
c#	102744
java	65394
javascript	57681
.net	51336
php	48192
jquery	44976
c++	40158
asp.net	36783
iphone	33383
python	32204

tags	tcount
c#	121532
java	105440
c++	75932
javascript	66292
php	60819
jquery	51530
.net	47445
c	39483
python	39208
sql	37193

Fig. 9. Top ten tags for the top twenty-five high reputation users in questions and answers



Top 10 tags in questions used by users with low reputation

tags	tcount
java	71323
php	70898
android	68396
c#	62843
javascript	57855
jquery	50477
iphone	31558
mysql	29656
c++	29054
python	27935

Top 10 tags in answers used by users with low reputation

tags	tcount
php	78216
c#	71456
java	71343
android	61813
javascript	60786
jquery	51986
iphone	35763
html	33824
asp.net	30250
c++	29160

Fig. 10. Top ten tags for the top twenty-five low reputation users in questions and answers

Finally, we analyzed the top ten tags that were used more frequently by the top twenty-five users with low reputations as in Fig.10. It seems that JAVA was tagged by most of the low reputation users in questions, where PHP was tagged by most of them in answers. In general, both data sets have some common tags but with different ranks. However, some tags like Python appears in questions but not in answers, and ASP.NET and HTML tags were found in answers and not in questions.

III. Related Works

Stack Overflow website has been studied from different perspectives. In this section, we review number of different projects that applied statistical analysis techniques on Stack Overflow data to obtain results for different purposes.

One of the studies is “A Discriminative Model Approach for Suggesting Tags Automatically for Stack Overflow Questions”. The main idea in [2] is helping the questioner in tagging his question by automating tag suggestions in Stack Overflow website. They used discriminative model on millions of SO questions. The focus of the analysis in this study is on analyzing the text content. Similarly in [3], the researchers’ goal is to analyze the non-functional requirement to investigate the main topic of the discussion in Stack Overflow website. Again here they are concentrating on the content of the text. They used latent Dirichlet allocation (LDA) to do the analysis.

In [4], the main idea is to analyze the Stack Overflow processes and reasoning the results. By studying the question and its answers as pairs with a focus on the voting process. They have two main tasks: predicting long-term value of a question and predicting whether a question has been sufficiently answered. However, the authors in [5] investigate how gender effects online participation, in order to encourage women participation. They did the analysis on gender resolution to explore the mannerism and timeliness of women engagement.

All the projects mentioned above are different from our project, since we studied more than one attribute on Stack Overflow attributes to get a statistical result rather than modeling a new technique.

IV. Discussion and Future works

Due to the limited time to finish this research, we could not dive deep in the data sets to acquire more analysis. We are planning to relate the distributions of the data sets and their variables to the multiplicative impact of several physical factors. Moreover, we only analyzed the tags with the top high or low reputation users, but we believe that there could exist a relationship with another variable or attribute as an example the score of a question or an answer. Knowing if a tag is new or an old one and it is mentioned in the appropriate question or answer could provide more insight to the analysis.

V. Conclusion

In this research, we analyzed the behavior of Stack Overflow statistics. Hence, we performed statistical methodologies to describe measures like score of questions, number of answers and comments, views, tags and users’ ratings.

Acknowledgment

We would like to thank Prof. Swapna Gokhale for her insightful comments and support during the entire duration of this research.

References

[1]. A. Bacchelli, "Mining Challenge 2013: Stack Overflow," in MSR'13.

[2]. A. Saha, R. Saha and K. Schneider, “A Discriminative Model Approach for Suggesting Tags Automatically for Stack Overflow Questions”. 2013.

[3]. Zou, L. Xu, W. Guo, M. Yan, D. Yang and X. Zhang, “Which Non-functional Requirements do Developers Focus on?”, in 12th Working Conference on Mining Software Repositories, 2015.

[4]. A. Anderson, D. Huttenlocher, J. Kleinberg and J. Leskovec, “Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow”, In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012.

[5]. B. Vasilescu, A. Capiluppi and A. Serebrenik, "Gender, Representation and Online Participation: A Quantitative Study", 2013.