Lab_10

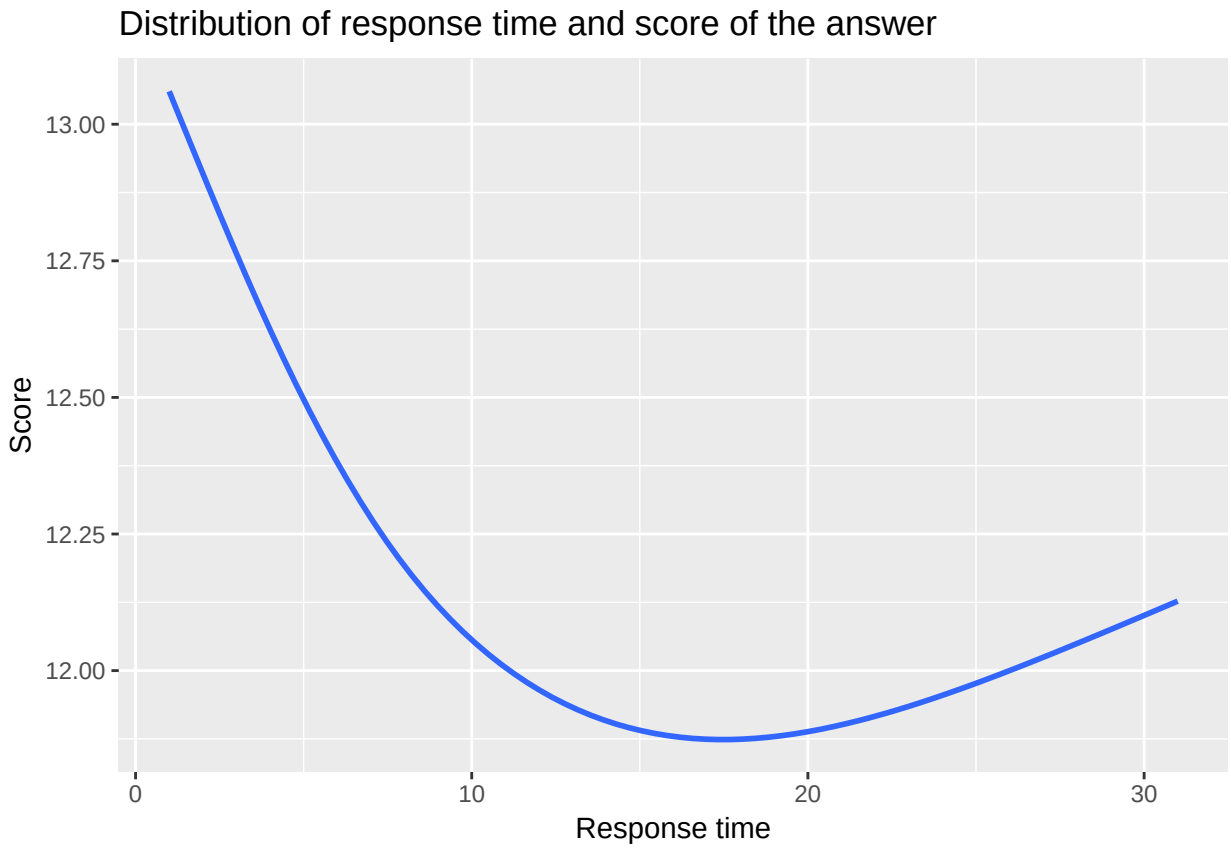
Fantastic Four

Team Question

How does the timeliness of a response affect the score it receives?

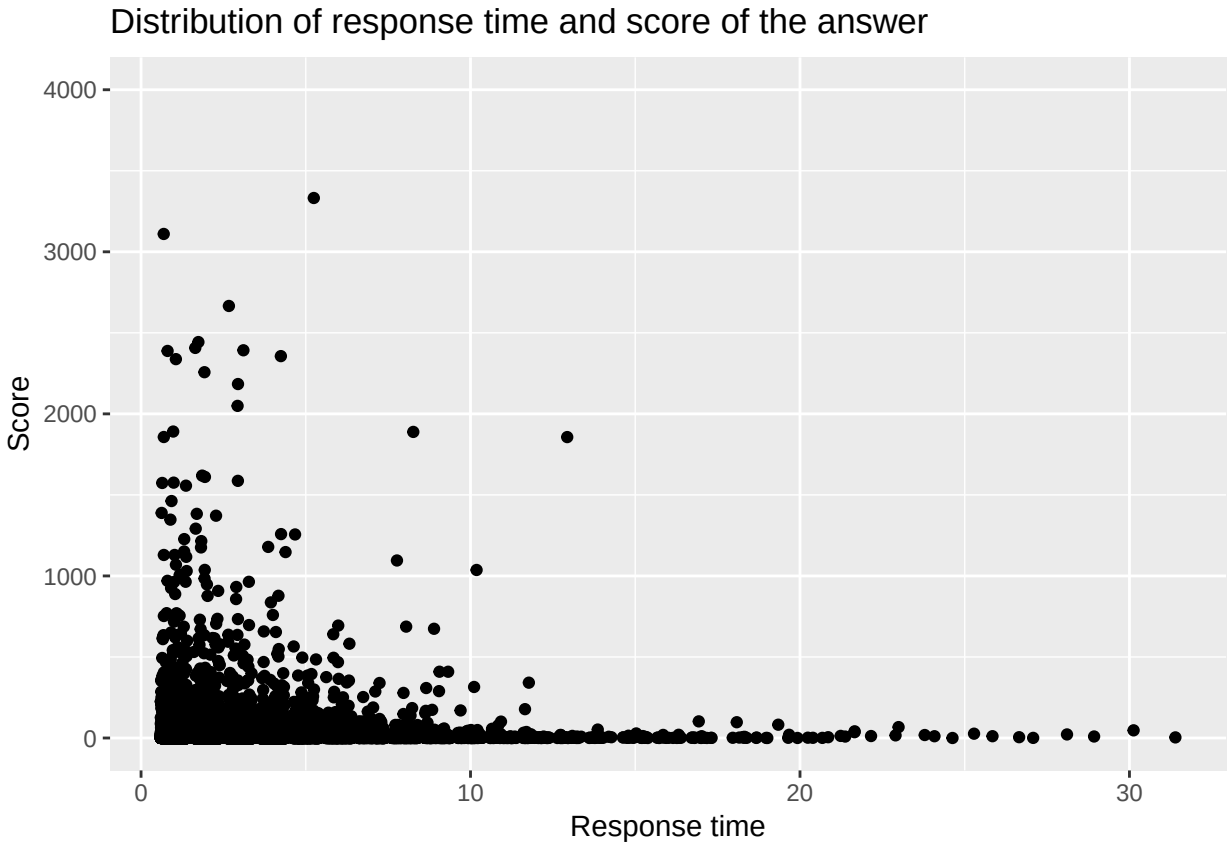
```
answersTime<-answers%>%
  group_by(ParentId)%>%
  arrange(CreationDate)%>%
  mutate(responseTime=row_number())

ggplot(data=answersTime)+ geom_smooth(mapping=aes(x=responseTime, y=Score),se=FALSE)+ggtitle("Distributi
```

```
## `geom_smooth()` using method = 'gam'
```




```
ggplot(data=answersTime)+ geom_jitter(mapping=aes(x=responseTime, y=Score))+ggtitle("Distribution of res
```

```
## Warning: Removed 4167 rows containing missing values (geom_point).
```



This graph presents a distribution that shows that the more time it takes for a response to be posted, the lower its score will be. However, when looking at the correlation between score and answer time:

```
cor(answersTime$Score,answersTime$responseTime)
```

```
## [1] -0.006214747
```

We find out that there is no significant correlation between these two variables.

Conclusion

The most impactful factors of the score of a post are the amount of time until the response, as well as the length of the post. While there are other elements which affect the post score, such as key words, hyperlinks and post structure, the most significant are the response length and the time to respond. It was determined that the longer the response, the higher the resulting score, and that the quicker a response was posted the higher score it recieved.

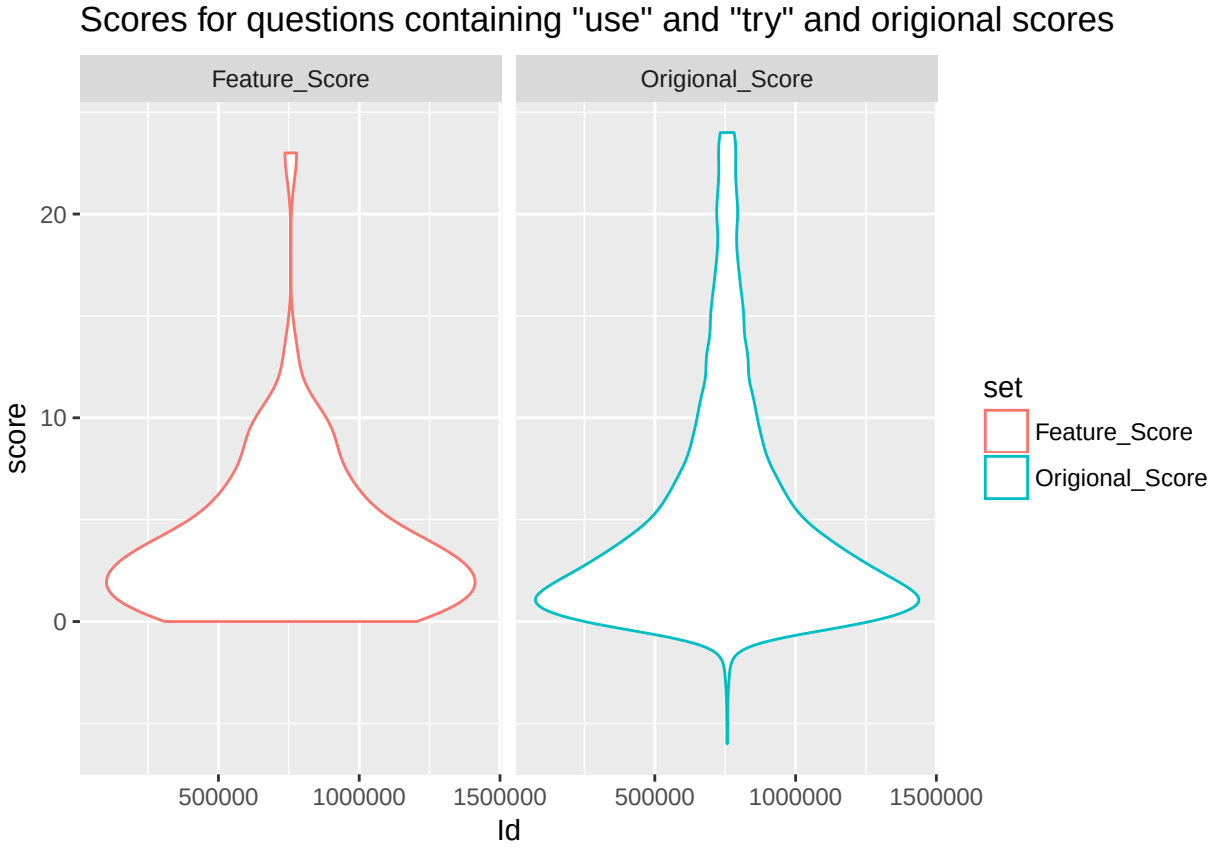
Individual Findings

What features affect the scores of the questions and answers and what is their impact?

Lindsay Gettel

```
Questions<-questions%>%
  mutate(questionLength=nchar(Body, type="chars")+ nchar(Title, type="chars"))
hasUse<-filter(questions, str_detect(questions$Title, "\\suse\\s"),
               str_detect(questions$Body, "try"), Score<100)%>%
  mutate(questionLength=nchar(Body, type="chars")+ nchar(Title, type="chars"))
newQ<-left_join(Questions, hasUse, by="Title")
newQ%>%
  rename(Feature_Score="Score.y", Origional_Score="Score.x", Id="Id.x")%>%
  filter(Origional_Score<25)%>%
  gather(`Feature_Score`,`Origional_Score`, key="set", value="score" )%>%
  ggplot()+ geom_violin(mapping=aes(x=Id, y=score, color=set))+facet_grid(~set)+ ggtitle("Scores for qu
```

```
## Warning: Removed 8421 rows containing non-finite values (stat_ydensity).
```

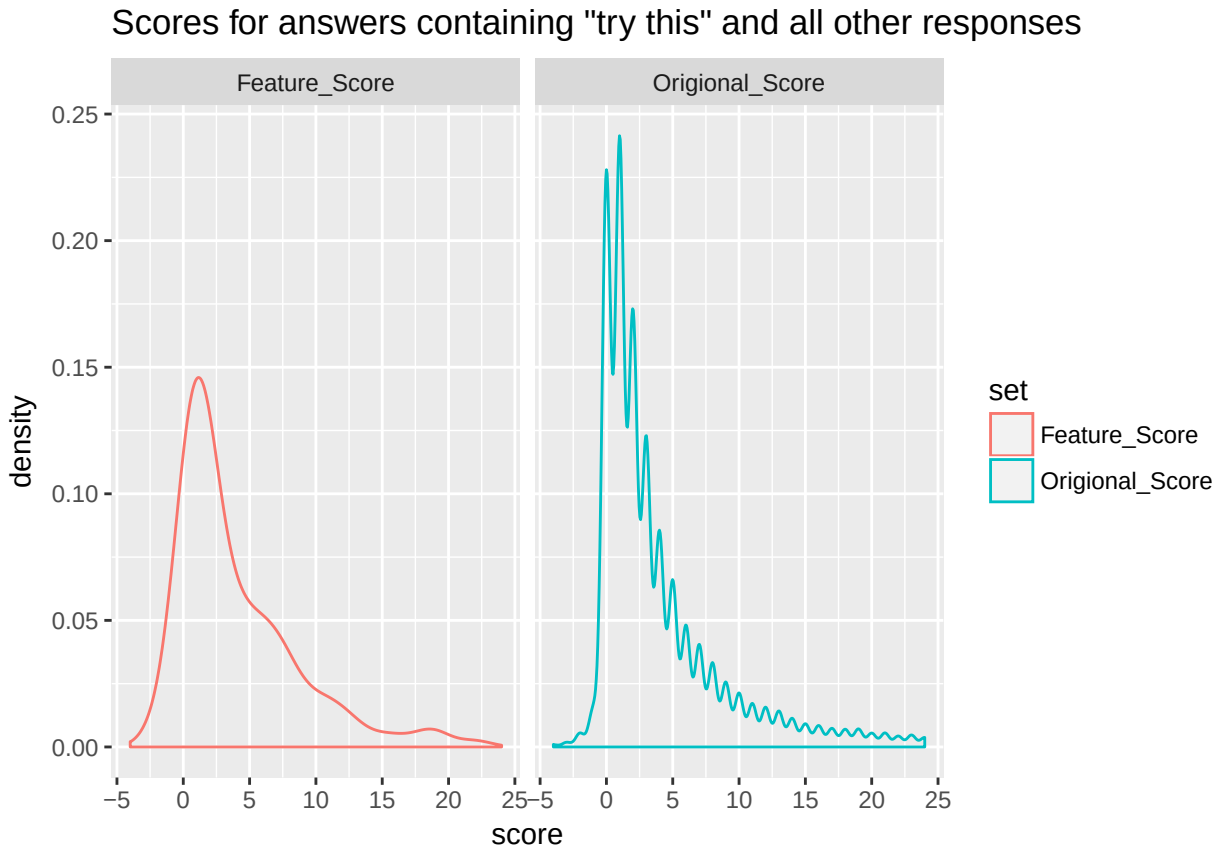


To analyze the questions, the focus feature was titles containing the word “use”, and the body containing variants of “try”. Connotation of use implies more logistical and processing related questions. This implementation could lead to a more specific question, which is easy to understand and interpret, resulting in a better score and reception of the question. The graph above demonstrates that questions with the given key words have a slightly higher average score than the average of the data set, there are also more scores ranked higher. The average for the questions with the applied feature have an average score of about 3, while the average score of all the questions is around 2, as seen in the graph there is also more scores in the range 3-10 for the feature questions.

```
hasTry<-filter(answers, str_detect(answers$Body, "\\stry\\sthis"))%>%
  mutate(responseLength=nchar(Body, type = "chars"))
newA<-left_join(answers, hasTry, by="Id")
newA%>%
  rename(Feature_Score="Score.y", Origional_Score="Score.x")%>%
```

```
  filter(Origional_Score<25, Origional_Score>-5)%>%
  gather(`Feature_Score`,`Origional_Score`, key="set", value="score" )%>%
  ggplot()+ geom_density(mapping=aes(x=score, color=set))+ facet_wrap(~set)+ ggtitle("Scores for answers
```

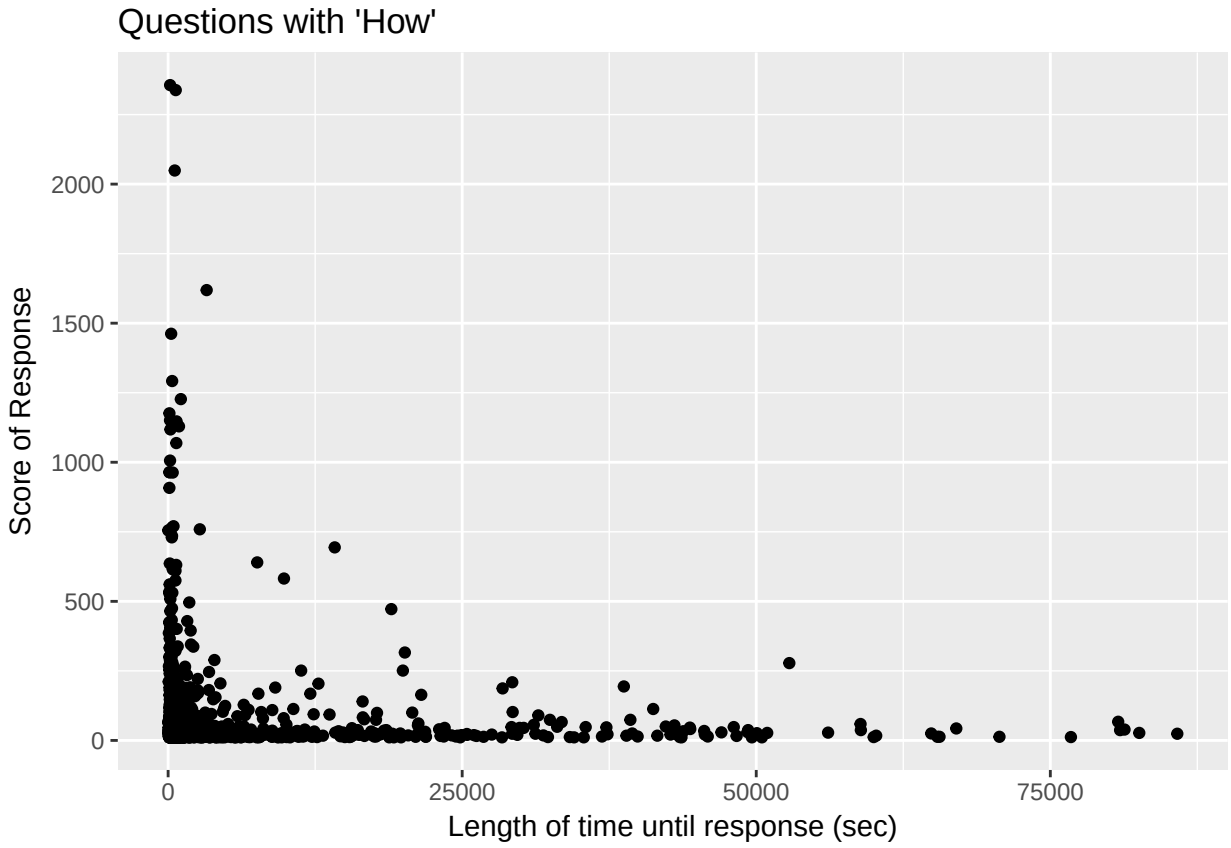
```
## Warning: Removed 36706 rows containing non-finite values (stat_density).
```



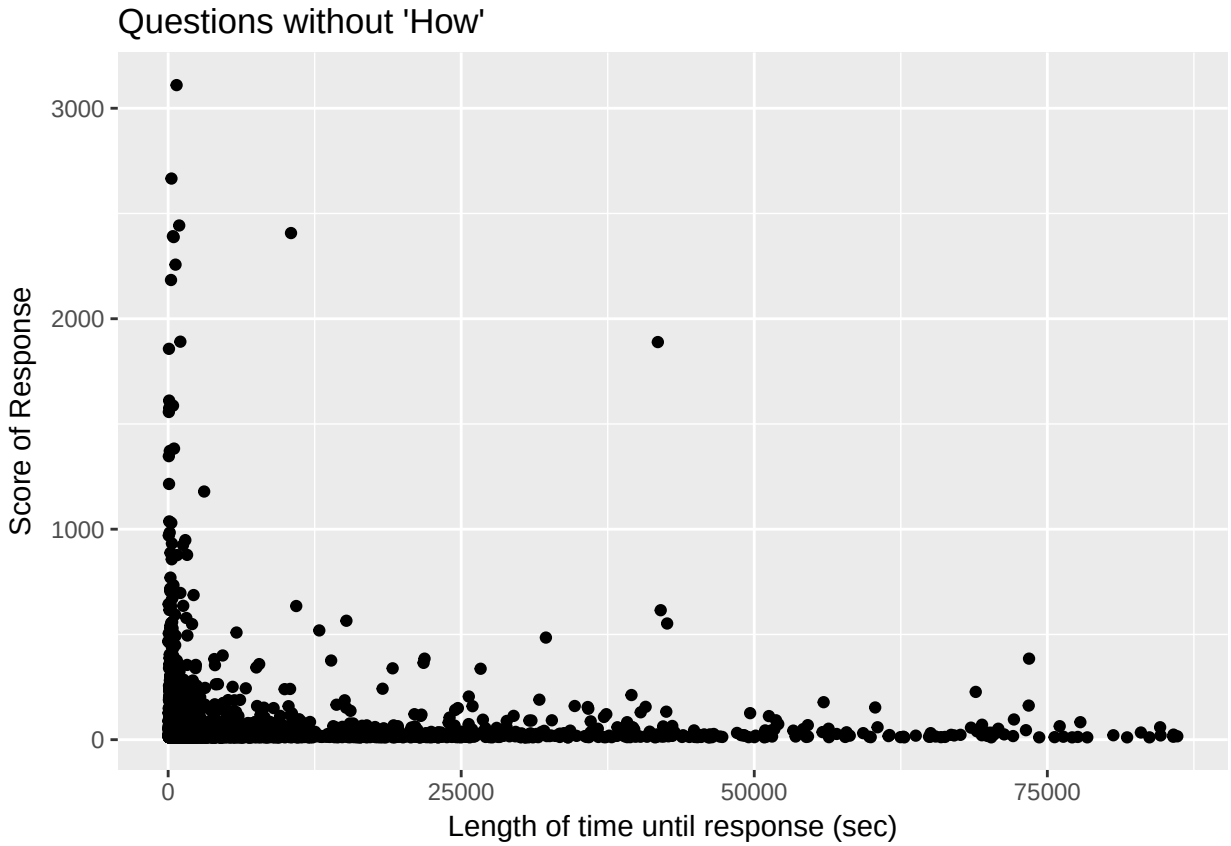
The feature applied to the answers were the key words “try this”, which is significant in indicating a direct way to troubleshoot and answer the question. It is also possible that the key words could be followed by multiple solutions and examples, this would be an efficient and helpful answer to a given question and should have a higher score. The above graph illustrates that the average score for the featured answers is fairly close to the overall average, at about 2. The two representations have similar distributions, just with slightly varying frequencies, which is logical because the featured answers are a subset and contains less responses so it follows that the frequencies would be less. Therefore, there does not appear to be any correlation between the key words “try this” and the ranking that the answer received.

Lexie Marinelli

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```




```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

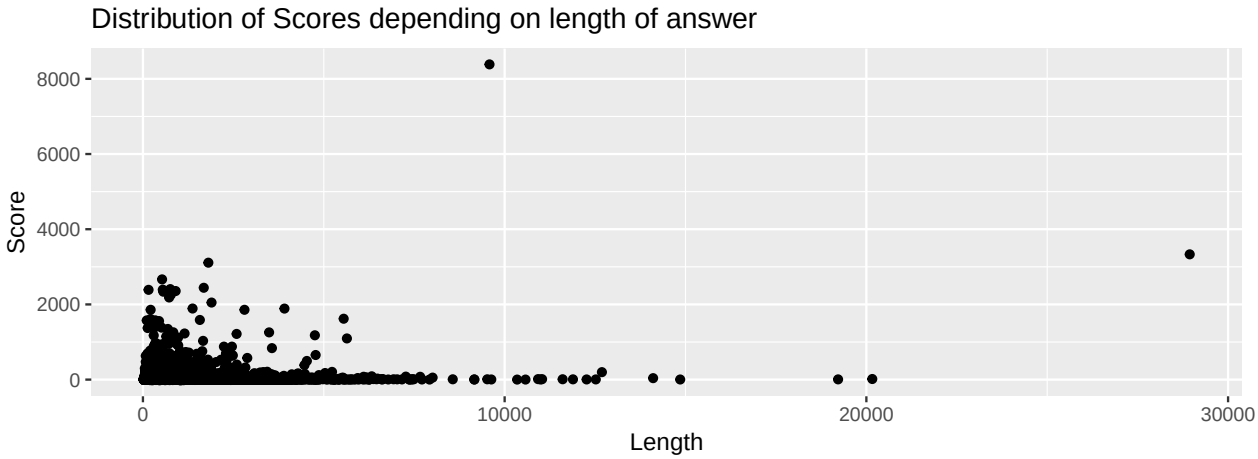


The two plots show us that there is no apparent relationship between questions that contain 'How' compared to questions that do not. The plots also show us that there is a direct relationship between the timeliness of the response and the score of the answer.

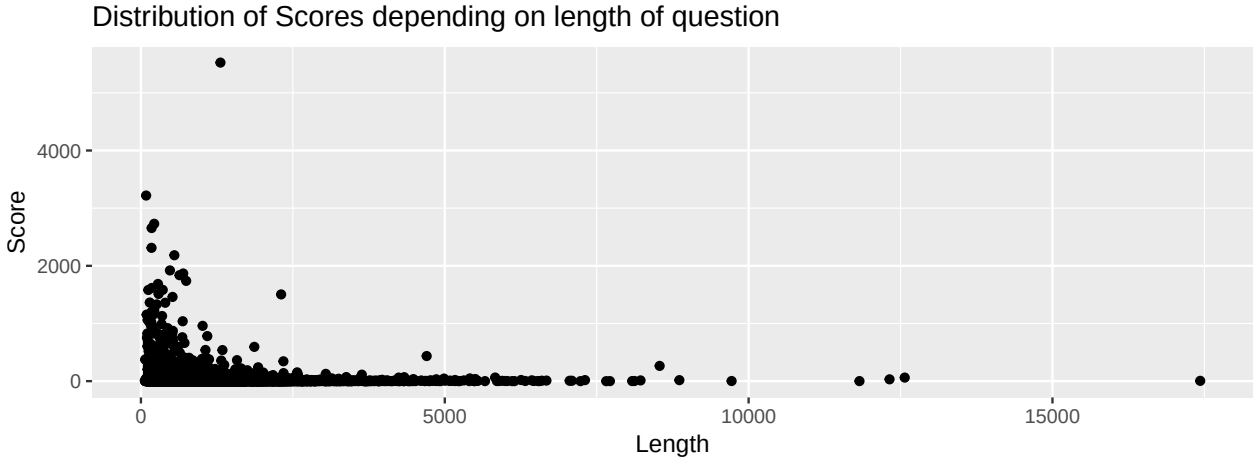
1. The feature I focused my search around was the timeliness of the response compared to the score of the answer separated by questions containing and not containing the word 'How'. I chose this because I was thinking that people who asked direct questions with a common question word would get a better response and if the difference in time the answer was posted, over the length of one day, had any effect on the score of the answer.
2. Time has an inverse relationship with the score of an answer. Also the plot containing the question word 'How' did not vary that much from the plot not containing the word.

Scott Baker

The feature I chose to look at is the relationship between the length of the question/response and the score it receives. First, look at the distribution of scores based on answer length:



From the plot, it is seen that the distribution favors higher scores for shorter lengths of responses. For questions, the following is found:



This also shows that there are higher scores for questions that are shorter in length. Now, examine the correlation values for these variables for questions and answers:

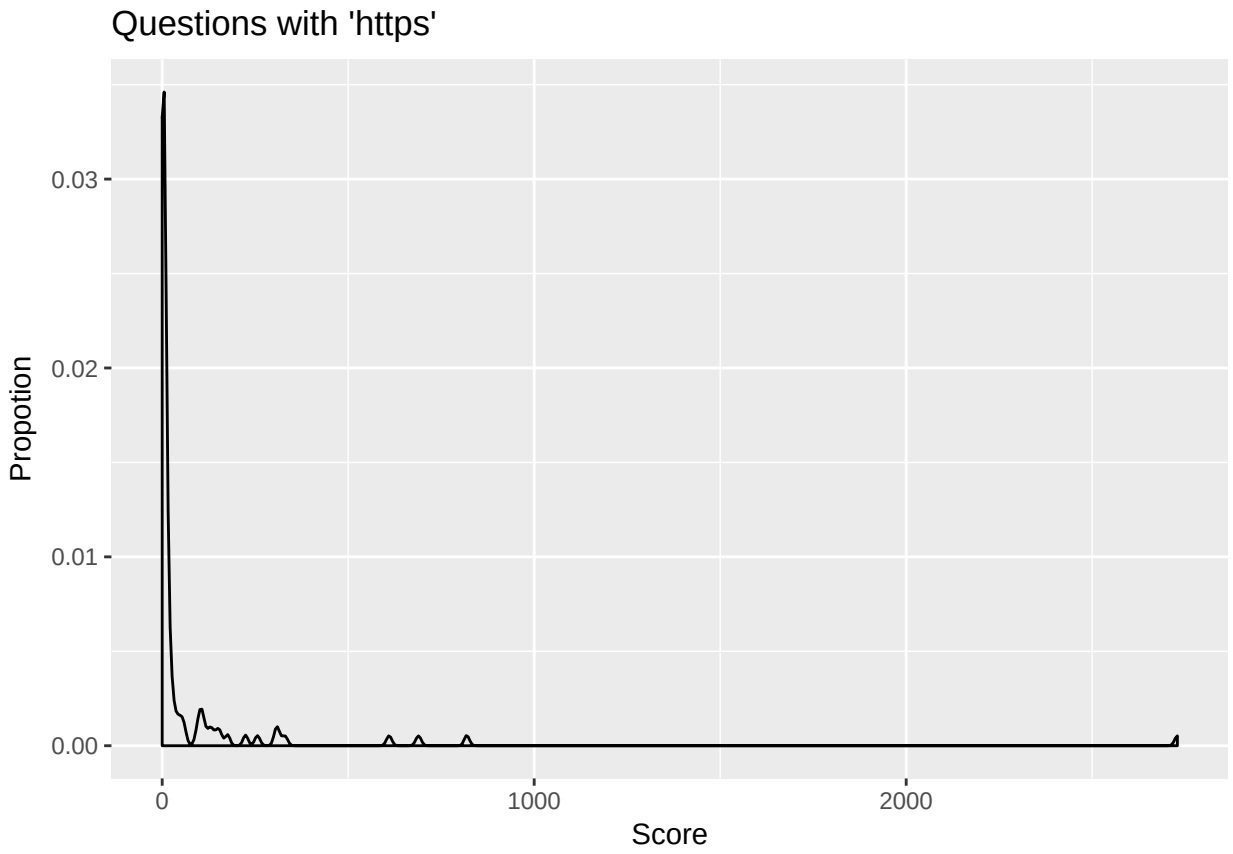
```
## [1] 0.1275499
```

```
## [1] -0.07070342
```

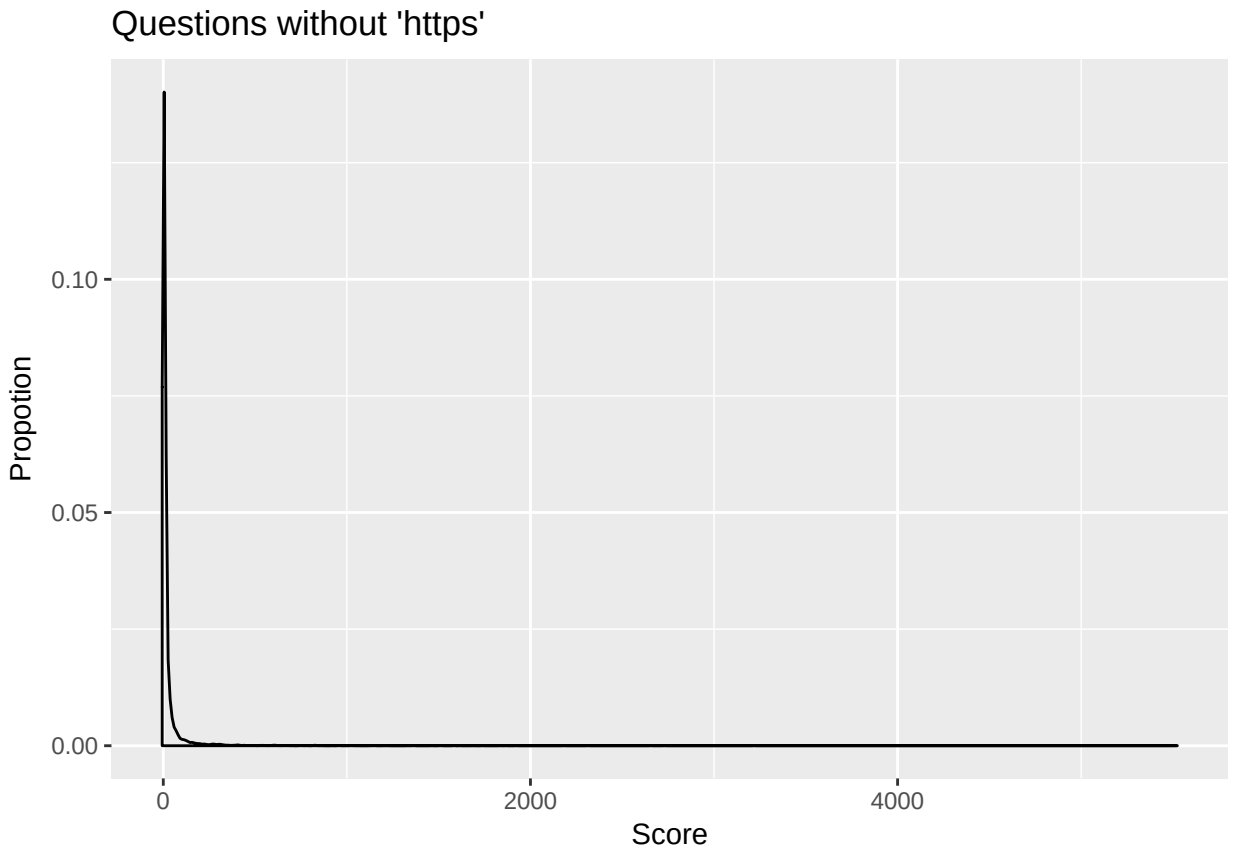
From this, it is seen that there is a small positive correlation (0.1275) between length of answer and the answer's score–which means in general, answers that are slightly longer get better scores. Second, there is a small and insignificant negative correlation (-0.0707) between question length and the score it receives. This means that there is not a relationship between the question length and the score. These two things are significant because it means that people read and vote on the question regardless of the length is has and that has little to do with the score it gets. Also, it means that, more-or-less, longer and more detailed answers will get better scores, since they are more in-depth.

Zhenlong Li

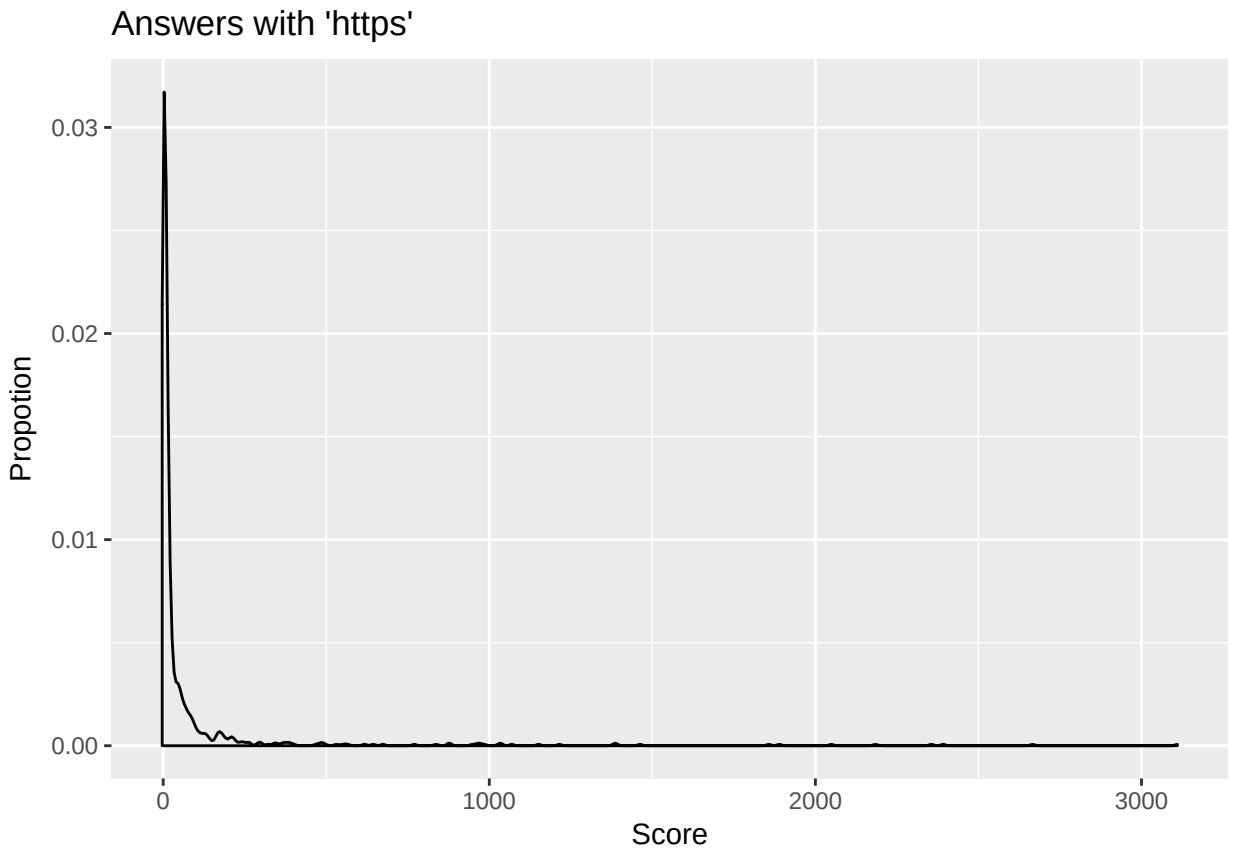
```
qdata = select(questions, Body, Score)
question_https = filter(qdata, str_detect(qdata$Body,"https"))
question_withouthttps = filter(qdata, str_detect(qdata$Body,"[^https]"))
ggplot(data = question_https)+ geom_density(aes(Score))+ggtitle("Questions with 'https'") + xlab("Score
```



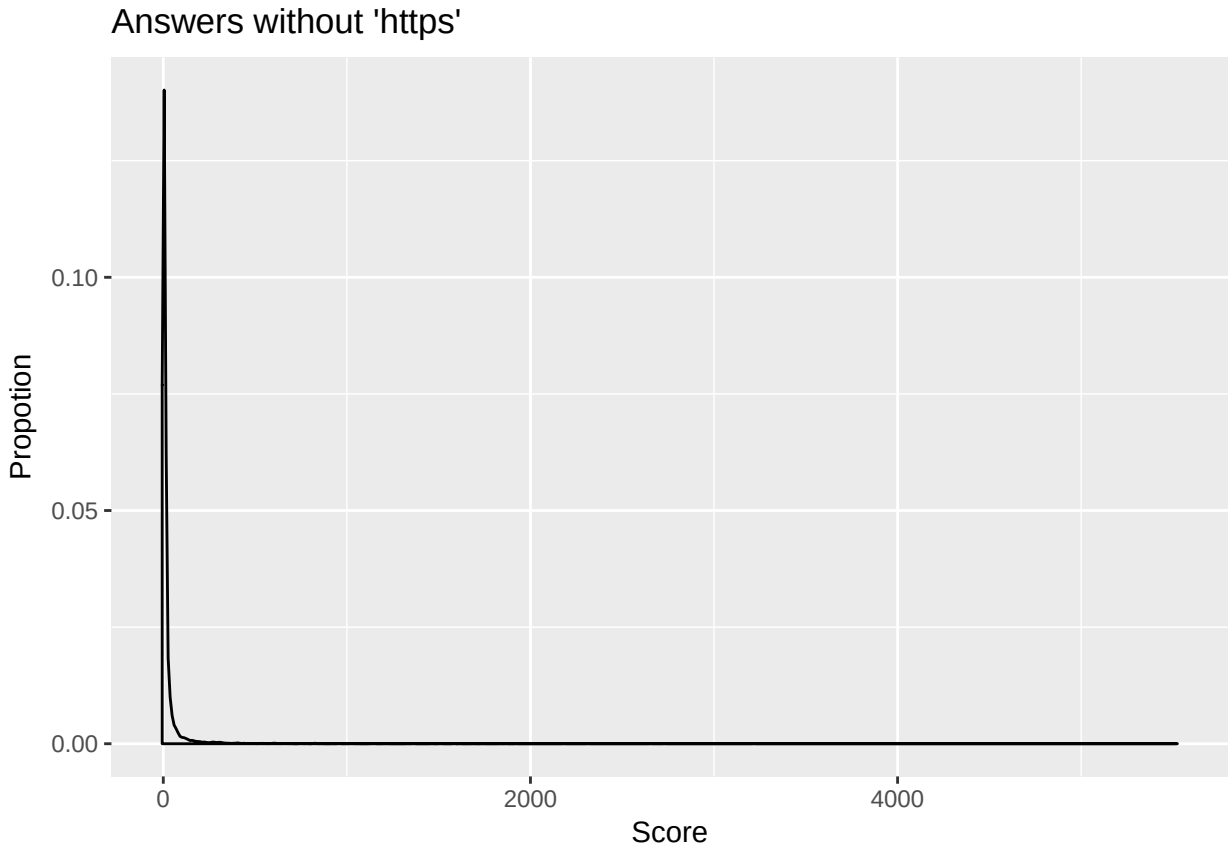

```
ggplot(data = question_withouthttps)+ geom_density(aes(Score))+ggtitle("Questions without 'https'") + x
```




```
adata = select(answers, Body, Score)
answers_https = filter(adata, str_detect(adata$Body,"https"))
answers_withouthttps = filter(qdata, str_detect(qdata$Body,"[^https]"))
ggplot(data = answers_https)+ geom_density(aes(Score))+ggtitle("Answers with 'https'") + xlab("Score") +
```




```
ggplot(data = answers_withouthttps)+ geom_density(aes(Score))+ggtitle("Answers without 'https'") + xlab
```



From the plots, we can see that inserting hyperlink which is included either in questions or answers will somehow help the score but the effect of adding hyperlink on the score is not so strong.

Contributions

- Lindsay: Choose different features and plotted them with the scores as well as the the original set of scores for comparison. Used mutating joins as well as other functions from dplr, ggplot, strings and a few regular expressions.
- Lexie: Created individual plots with different features and analysed the relationship between the features.
- Li: Created four individual plots to determine whether a hyperlink in questions or answers will help the score and knit/submit this lab to OSF.
- Scott: Managed the git and organized files. Also looked at the relationship between length of question/response and the score it received. Was available to knit and submit remotely if needed.