insights_writeup

Benji Lu

December 13, 2015

Topic Insights

Here, we highlight some interesting findings that we discovered through our topic analysis, beginning with topic trends.

Two prominent movements in the recent history of the Claremont Colleges are the divestment campaign and the push for the unionization of the Pomona's dining hall workers.[1][2][3][4][5] Using our model's visualizations of topics over time, we found the following trend of *TSL*'s coverage of the two topics:

```
# INSERT TOPIC GRAPH HERE
```

The graph shows that *TSL* coverage of the two topics was robust from the spring of 2009 up to and including the spring of 2013. In particular, it reached its height between the fall of 2011 and the spring of 2013. Some interesting article headlines include "NLRB Will Investigate Labor Practice Charges at Pomona Dining Halls" at the peak in December 2011, "WFJ Votes to Unionize" at the very end of May 2013, "Pomona Opts Not to Divest" in October 2013, and "Pitzer College to Divest From Fossil Fuel Funds" around April of 2014.

Unfortunately, our model wasn't able to separate the two topics, but the results are nonetheless fascinating. From our own perspective as juniors at Pomona College, the visualization is especially illuminating. We came to Pomona in the fall of 2013, right at the tail end of the peaks of the dining hall and divestment movements, as the visualization accurately shows. The semester before we arrived, the dining hall workers voted to unionize; the semester we arrived, President David Oxtoby announced Pomona's decision not to divest from fossil fuels; and the semester after we arrived, Pitzer College announced that it would divest from fossil fuels. Shortly after that, the discussions about the two topics died down, both in *TSL* and in general campus discourse. Regardless of one's opinions on these issues, they were exciting events in the colleges' history, and while we were able to witness the ultimate outcomes, we had virutally no knowledge of the related events and discussions that preceded the outcomes. With the visualizations our model provides, we're now better able to access key institutional memory in the form of *TSL* coverage of the topics in order to understand these outcomes in their broader context that spans several years.

Another interesting topic that has been trending recently in *TSL* coverage is sexual assault. Here's the graph that our model produced:

```
# INSERT GRAPH HERE
```

The trend is quite dramatic: Sexual assault was rarely covered in *TSL* before 2012, but it quickly gained prominence starting in 2013 and appears to have peaked in 2014. This is pretty consistent with national coverage of college sexual assault.[6] The White House published its first report on college sexual assault in April 2014, Columbia University student Emma Sulkowicz began carrying a mattress around campus beginning in September 2014, California became the first state to institute an affirmative consent law in September 2014, and Title IX complaints filed against colleges increased from 11 in the 2009-2010 fiscal year to over 37 in the 2013-2014 fiscal year.[7]

[1] http://www.pomona.edu/news/2013/09/25-divestment-decision
[2] http://www.dailykos.com/story/2014/5/4/1296249/Rethink-Divestment-5-1-2014-Pomona-College
[3] http://www.latimes.com/opinion/op-ed/la-oe-morrison-gould-20141021-column.html
[4] http://www.nytimes.com/2012/02/02/us/after-workers-are-fired-an-immigration-debate-roils-california-campus.html?_r=0
[5] http://articles.latimes.com/2013/may/01/local/la-me-ln-college-workers-20130501
[6] http://america.aljazeera.com/watch/shows/america-tonight/articles/2014/11/10/timeline-collegesexualassault.html
[7] http://www.nytimes.com/2014/05/04/us/fight-against-sex-crimes-holds-colleges-to-account.html

From some of the headlines in the visualization, we can see that the Claremont Colleges have undergone a multitude of changes in response to sexual assault beginning in 2013, including changes in party culture, the creation of task forces dedicated to sexual assault, the implementation of bystander training programs like Teal Dot, and the creation of full-time Title IX coordinator positions.

The final topic trend that we found particularly insightful is that related to movements centered around inequality. Here's the trend that our model produced:

```
# INSERT GRAPH HERE
```

Based on the graph, it seems that events and discussions centered around inequality (our model appears to overlap racial, sexual, and economic inequality to some degree) have been occurring pretty consistently over time at the Claremont Colleges. We can see that in December 2012, the Occupy Movement gained some coverage by *TSL*. More recently, in December 2014 students reacted to events in Ferguson by marching through the colleges. The most obvious observation, though, is that the topic really blew up in November 2015, when students protested recent events related to race centered at Claremont McKenna College. [ELABORATE MORE HERE]

Another neat feature of our model is its analysis and visualization of topic coverage by writer. Here's an example:

```
# INSERT GRAPH OF DIANE LEE HERE
```

The graph shows the topics the writer has written the most about. In this case, we can see that Diane Lee wrote a lot of pieces on administrative actions and decisions at the Claremont Colleges. [maybe elaborate on her positions through the years]

Here's another example:

```
# INSERT GRAPH OF SANA KADRI
```

It seems like Sana Kadri wrote a lot about internationalism and race for *TSL*. Our team for this project includes the opinions editor for *TSL* this semester, who can confirm that Sana, who was an opinions columnist, was very passionate about those topics.

Finally, we can examine the topic content of articles written by the editorial board of *TSL*. Each issue, the editorial board, which consists of the editor-in-chief and the two managing editors, writes a brief editorial that is published in the opinions section.[8] Often, it reflects on another article printed in the same issue. Here's the topic content analysis of the editorial board's pieces:

```
# INSERT GRAPH OF EDITORIAL BOARD
```

The chart indicates that the editorial board seems to focus primarily on on-campus issues related to inequality. This would make sense, since events related to such issues often occur and are covered by *TSL*. Also, these issues often generate a lot of discussion and, at times, controversy on campus, making them relevant and appealing topics of conversation for the editorial board to discuss.

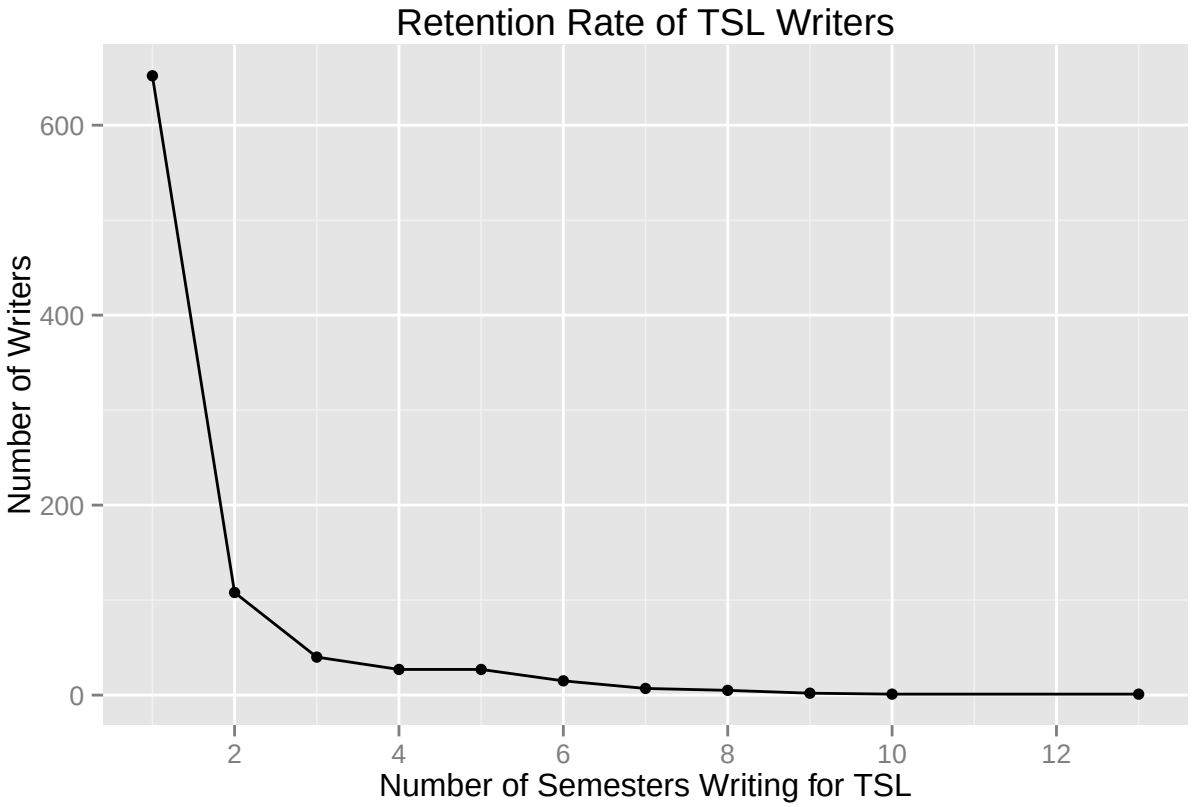
Miscellaneous Insights

There's a lot to be learned from the data beyond the topic content analysis. For example, with a little bit of wrangling and visualization, we were able to arrive at some interesting findings that might be of some value to the staff at *TSL*. For example, it's no secret that the newspaper struggles with retaining staff writers.

[8] It shoud be noted that, since *TSL* staff changes every semester, the people who are part of editorial board are not necessarily the same from one semester to the next.

But just how bad is it? By tracking the date of each writer's first published article and the writer's last published article, we were able to get a sense of how long they wrote for the paper. Plotting the data, we got the following graph:

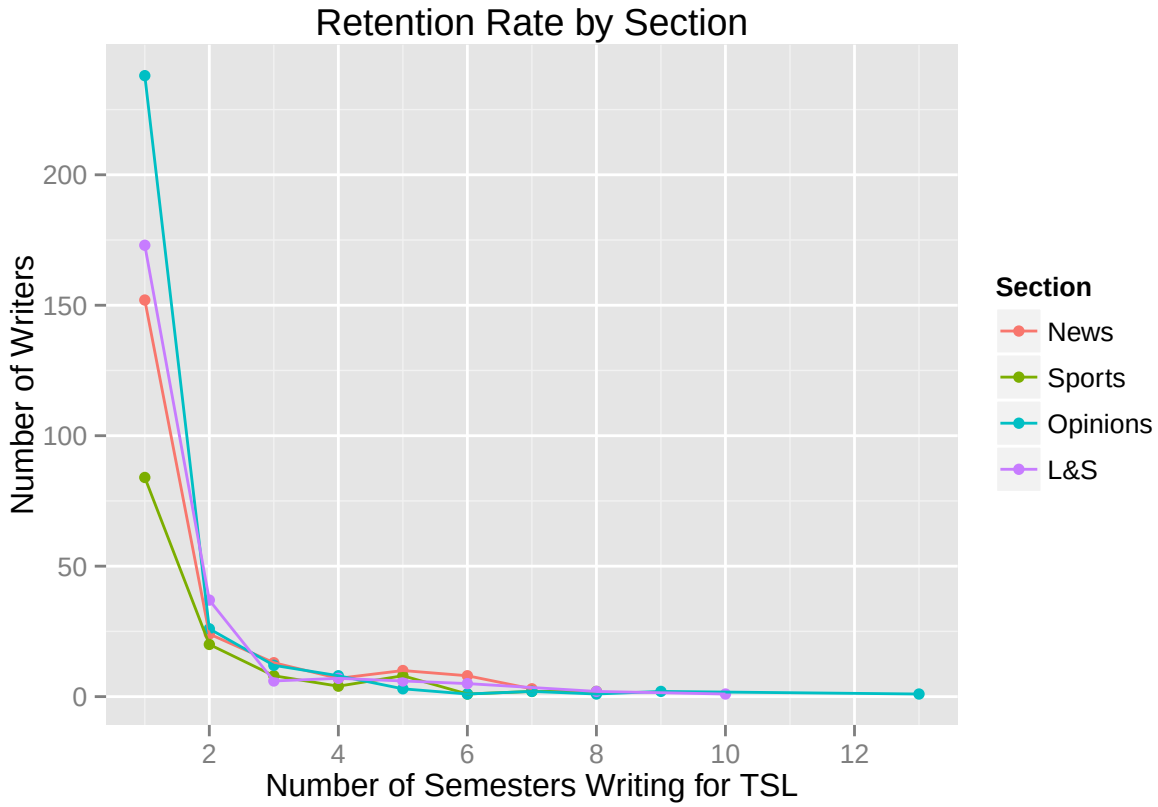
```
retention %>%
  group_by(semestersTotal) %>%
  summarise(n = n()) %>%
  ggplot(aes(x = semestersTotal)) + geom_line(aes(y=n)) + geom_point(aes(y=n)) +
  xlab('Number of Semesters Writing for TSL') + ylab('Number of Writers') +
  ggtitle('Retention Rate of TSL Writers') +
  scale_x_continuous(limits=c(1,13),breaks=seq(0,13,2))
```



It seems like *TSL*'s reputation for shedding writers isn't exaggerated. Out of the pool of writers over the past five years, 652 writers stay for one semester or less; 108 writers stay for two semesters; 40 stay for three; and 85 stay for four or more semesters. There are a few caveats, though. First, since the database only includes articles published within the last five years, it's possible that some writers who were seniors when the data first began being collected had been writing for *TSL* for many semesters but are represented as only writing for one semester because their earlier articles aren't included in the database. This might lead to an overestimate of the number of writers who only stay for one or two semesters. Second, the method we applied above to calculate retention reports the difference between the first semester they began writing for *TSL* and the last semester they wrote for *TSL*. Some writers, however, take a semester or more off from writing for *TSL* to study abroad or simply pursue other interests before returning to write again. In these cases, the number of semesters they spent writing for *TSL* are overestimated. These cases probably consistute only a small minority of the data points, however, so the overall trend still holds.

We can easily track the retention rate by section as well:

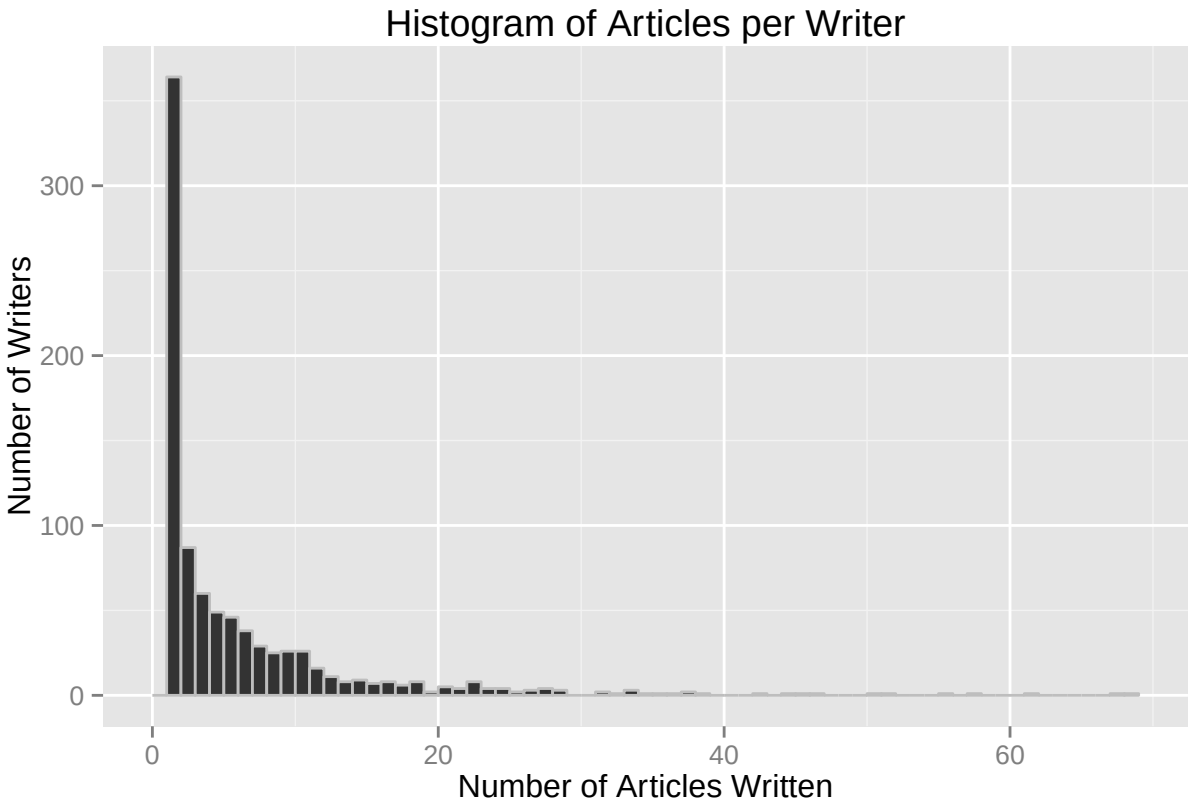
```
retention %>%
  filter(section_id != 5) %>%
  group_by(semestersTotal, section_id) %>%
  summarise(n = n()) %>%
  ggplot(aes(x = semestersTotal, y=n, color=factor(section_id))) + geom_point() +
  geom_line() + scale_color_discrete(name="Section", labels =
  c('News','Sports','Opinions','L&S')) + xlab('Number of Semesters Writing for TSL') +
  ylab('Number of Writers') + ggtitle('Retention Rate by Section') +
  scale_x_continuous(limits=c(1,13),breaks=seq(0,13,2))
```



It seems like the opinions section suffers the steepest drop in returning writers. A lot of this may be attributable to the fact that anywhere from one-fourth to one-half of the articles in the opinions section each week are written by guest writers, who typically write only for a single, specific event or issue. Unfortunately, we can't filter those writers out given the data that we have.

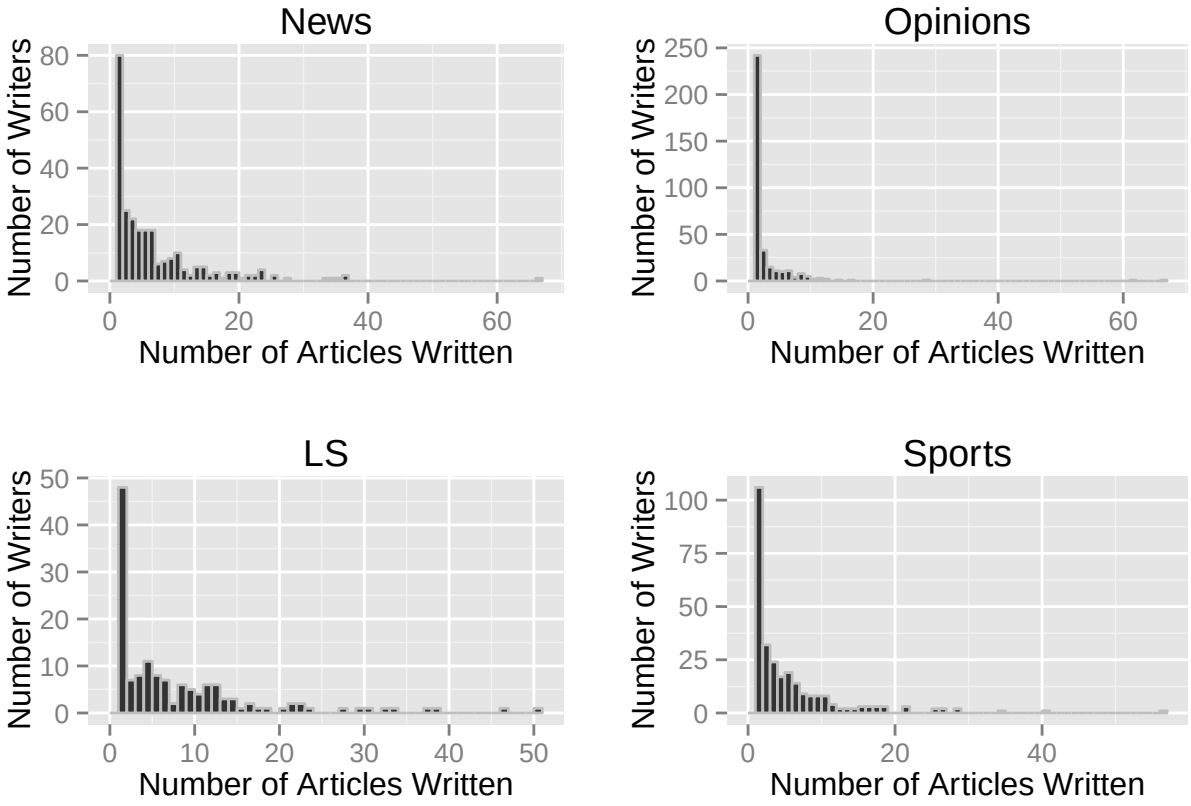
Another area of interest might be how prolific writers for *TSL* are–that is, how many articles each writer writes. It seems pretty reasonable to expect that trend to resemble the retention rates shown above. After all, a person can only write so many articles if they're only a writer for one semester. And indeed, we can see that this is the case:

```
articles_per_writer %>%
  ggplot(aes(x = total)) + geom_histogram(binwidth = 1, col=I('gray')) +
  ggtitle('Histogram of Articles per Writer') +
  xlab('Number of Articles Written') + ylab('Number of Writers')
```



We can also break it down by section:

```
sections = c('News','LS','Opinions','Sports')
for (num in 1:4) {
  section <- sections[num]
  assign(paste(section, 'section_articles_per_writer', sep='_'), articles %>%
    filter(section_id == num) %>%
    group_by(profile_id) %>%
    summarise(total = n()) %>%
    ggplot(aes(x = total)) + geom_histogram(binwidth = 1, col=I('gray')) +
    ggtitle(section) + xlab('Number of Articles Written') + ylab('Number of Writers'))
}
grid.arrange(News_section_articles_per_writer,Opinions_section_articles_per_writer,
             LS_section_articles_per_writer,Sports_section_articles_per_writer, ncol=2)
```



Finally, we were able to get a sense of how popular articles published this semester have been by looking at the number of times each article was visited. Because data on the number of visitors to each article only began being collected in the fall semester of 2015, we had to limit our analysis to articles in that time period. Below are the top 10 most-viewed articles from the fall:

```
##                                                                                                 title
## 1                                                Who Is the Happiest at the "Happiest College in America"?
## 2                                  This Year's Pomona Essay Questions Discourage Underrepresented Applicants
## 3  CMC Students of Color Protest for Institutional Support, Call for Dean of Students to Resign
## 4                                              CMC’s Black Students See Low Graduation Rates, Lack of Support
## 5                                                       Amid Calls for Resignation, Dean Spellman Steps Down
## 6                                                                                  Why I Left (Pomona College)
## 7                                                                Pomona College Receives Title IX Complaint
## 8                                                                 Break the Mold: Why I Won’t Donate to CMC
## 9                                                  Pomona Taco Crawl: The Must-See Taquerias in Pomona
## 10                                            Even in Foam and Plastic, Gun  Violence Does Not Belong at 5Cs
##    published_date             author_name clicks
## 1      2015-10-23        Lisette Espinosa  22012
## 2      2015-11-06            Chuck Herman   8602
## 3      2015-11-11          Kevin Tidmarsh   8251
## 4      2015-10-09          Sam McLaughlin   4162
## 5      2015-11-12          Kevin Tidmarsh   3862
## 6      2015-11-06 Conner Bouchard-Roberts   3669
## 7      2015-10-09          Kevin Tidmarsh   3577
## 8      2015-11-11       Aseem Chipalkatti   2467
## 9      2015-09-18        Joaquin Banuelos   2187
## 10     2015-10-09          Benjamin Cohen   2129
```

For those who are familiar with recent events on campus this semester, the list makes sense. For example,

Lisette Espinosa emailed her opinions piece to Claremont McKenna College's former Dean of Student Mary Spellman, who gave a controversial reply that, some would argue, led to her resignation. Indeed, it seems like a lot of the top articles are related to issues of race, sexual assault, campus climate, and college image.

Similarly, we can identify the most-viewed writers on average:

```
## Source: local data frame [10 x 3]
##
##                author_name average_views articles_published
##                      (fctr)         (dbl)              (int)
## 1          Lisette Espinosa     22012.000                  1
## 2              Chuck Herman      8602.000                  1
## 3  Conner Bouchard-Roberts      3669.000                  1
## 4            Kevin Tidmarsh      2637.714                  7
## 5         Aseem Chipalkatti      2467.000                  1
## 6            Benjamin Cohen      2129.000                  1
## 7             Adin Bonapart      1648.000                  1
## 8          Joaquin Banuelos      1270.500                  2
## 9          Carol Ann  Routh      1174.000                  1
## 10             Tom Schumann      1167.000                  1
```

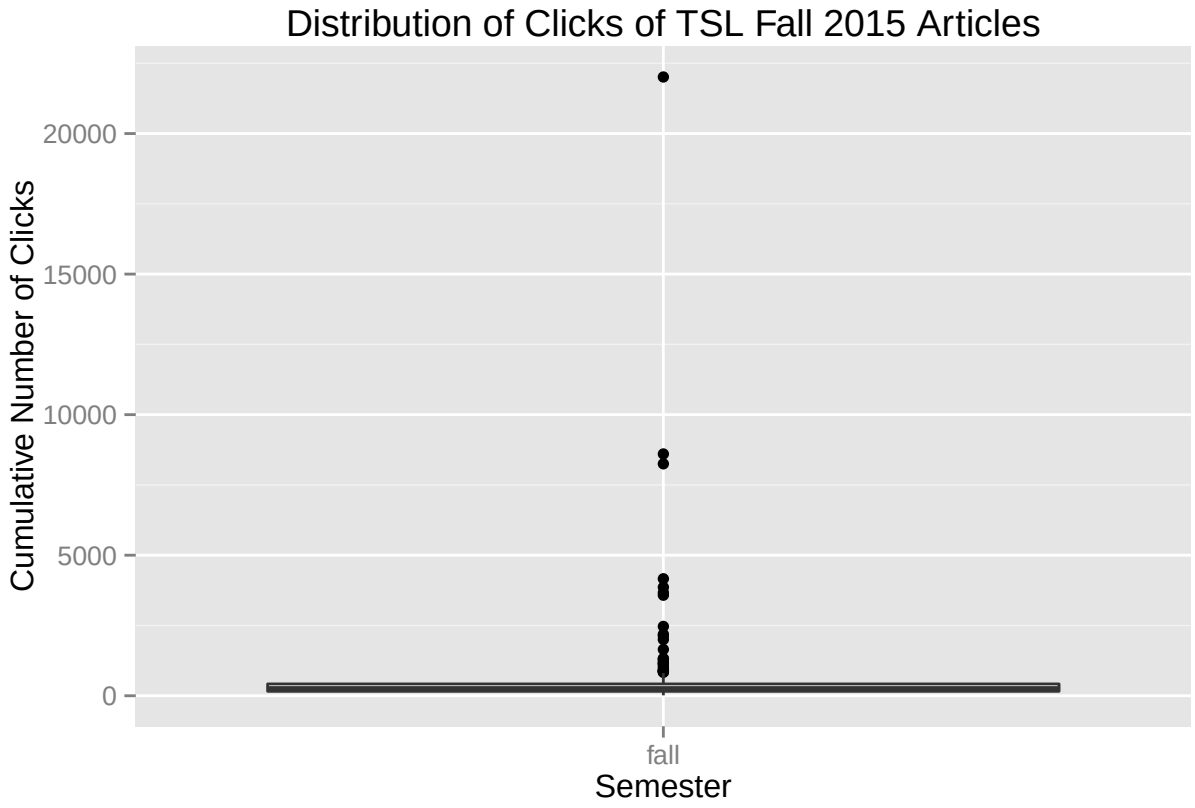
A lot of these are one-hit writers, which makes sense. The opinions section regularly invites guest writers to publish a piece in the paper. Typically, when this happens, the guest writer writes about a particularly controversial or timely subject; that is what motivates them to write, after all. As such, their pieces often get a lot of views. On the other hand, the other sections' pieces are written by regular staff writers, who have to cover the mundane along with the occasional big events each semester. Since the guest writers typically only write once about a hot-button issue, it makes sense that many of the most-viewed writers have only written once for the newspaper.

Nonetheless, we might be interested in seeing which regular contributors to the paper's content get the most views:

```
## Source: local data frame [10 x 3]
##
##            author_name average_views articles_published
##                  (fctr)         (dbl)              (int)
## 1       Kevin Tidmarsh     2637.7143                  7
## 2       Sam McLaughlin     1104.1667                  6
## 3           Sana Kadri      608.7500                  4
## 4     Editorial Board       603.7000                 10
## 5   William Schumacher      587.2500                  4
## 6        Elizabeth Lee      471.8333                  6
## 7           Sean Ogami      442.5556                  9
## 8       Harini Salgado      432.0000                  5
## 9            Diane Lee      429.8000                  5
## 10        Natalie Quek      396.5000                  4
```

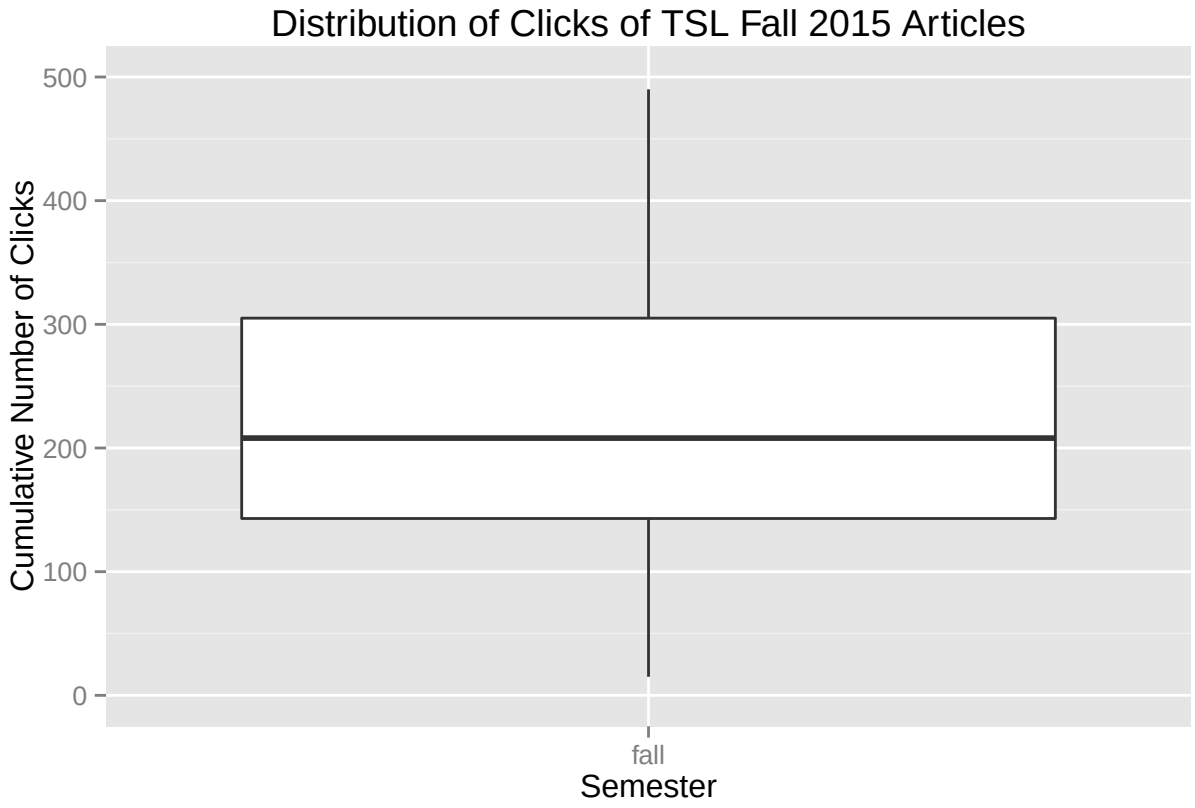
We now move from examining individual view counts to examining the aggregate data. Here's a boxplot of article views for articles published in fall 2015:

```
# plot distribution of views in fall 2015
clicks %>%
  ggplot(aes(x = semester, y = clicks)) + geom_boxplot() + xlab('Semester') +
  ylab('Cumulative Number of Clicks') +
  ggtitle('Distribution of Clicks of TSL Fall 2015 Articles')
```



We can see that there are some extreme outliers—the articles from the most-viewed list above—that make it difficult to look at the rest of the observations, so we'll focus our view on the bulk of the data:

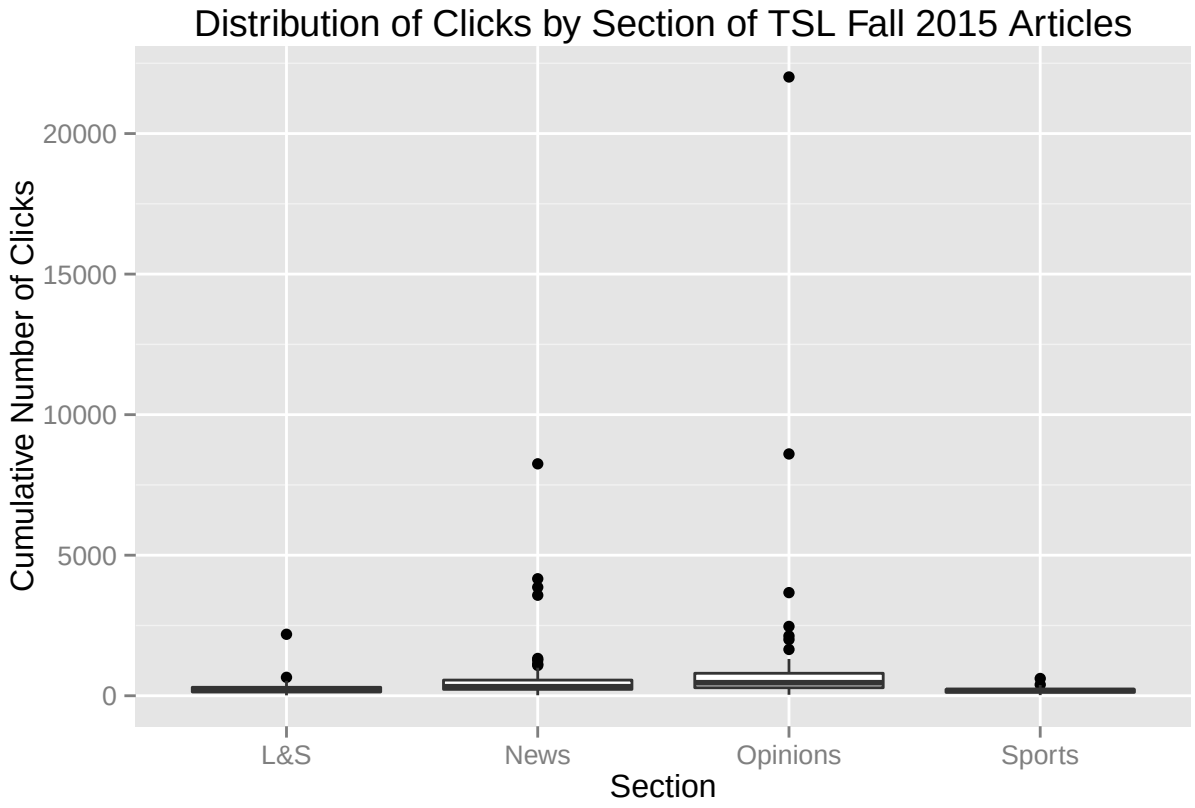
```
# exclude outliers from view
clicks %>%
  ggplot(aes(x = semester, y = clicks)) + geom_boxplot() +
  scale_y_continuous(limits = c(0,500)) + xlab('Semester') +
  ylab('Cumulative Number of Clicks') +
  ggtitle('Distribution of Clicks of TSL Fall 2015 Articles')
```



The median article published this semester, arranged by number of views, was viewed just over 200 times. The upper quartile of articles are viewed at least 300 times each, while the lower quartile are viewed at most about 145 times each.

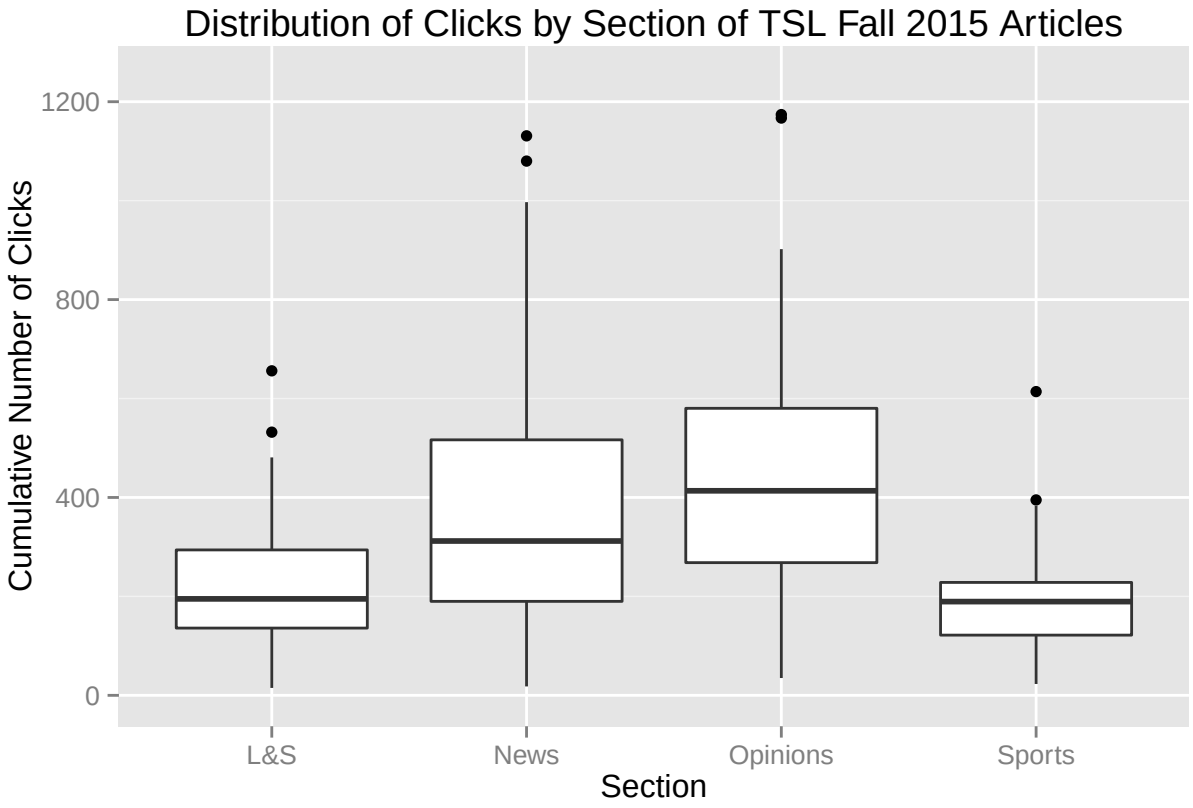
We can break the data down by section too:

```
# do the same as above but by section
clicks %>%
  filter(section_id != 5) %>%
  mutate(section = ifelse(section_id==1, 'News', ifelse(section_id==2, 'Sports',
  ifelse(section_id==3, 'Opinions', 'L&S')))) %>%
  ggplot(aes(x = factor(section), y = clicks)) + geom_boxplot() +
  xlab('Section') + ylab('Cumulative Number of Clicks') +
  ggtitle('Distribution of Clicks by Section of TSL Fall 2015 Articles')
```



Focusing on the bulk of the observations:

```
# exclude outliers from view
clicks %>%
  filter(section_id != 5) %>%
  mutate(section = ifelse(section_id==1, 'News', ifelse(section_id==2, 'Sports',
  ifelse(section_id==3, 'Opinions', 'L&S')))) %>%
  ggplot(aes(x = factor(section), y = clicks)) + geom_boxplot() +
  scale_y_continuous(limits = c(0,1250)) + xlab('Section') +
  ylab('Cumulative Number of Clicks') +
  ggtitle('Distribution of Clicks by Section of TSL Fall 2015 Articles')
```



It seems like opinions pieces are viewed the most, though news pieces are not far behind. Articles in the sports and life & style section, though, are not viewed as much; the upper quartile for sports articles is below the lower quartile for opinions pieces, and the upper quartile for life & style articles is not much higher. This makes sense given the atmosphere at the Claremont Colleges. Students generally seem less interested in topics covered in sports and life and style pieces than they are in topics covered in news and opinions pieces.