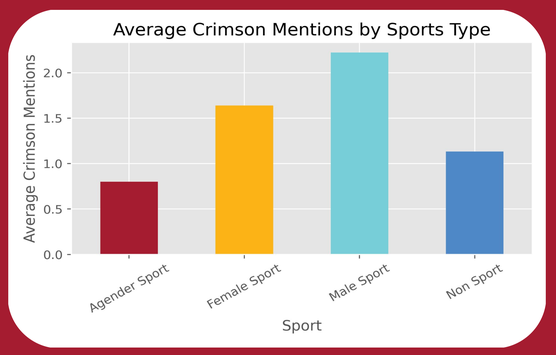
Who Gets Mentioned in the Crimson? By Carmen Chan ’22

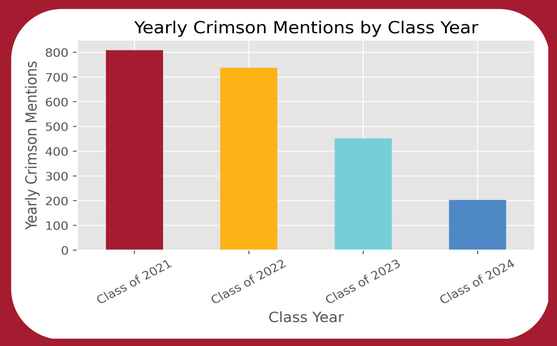
Data Science Major Capstone: Demographics of the most discussed Harvard students, according to Crimson mentions

Introduction

We started with a dataset containing the names of every current undergraduate mentioned in The Crimson between 2016 and 2020, extracted using named entity recognition. We then scraped the Harvard College Facebook to find the names, years, and House affiliations of all current students. We also scraped the Harvard athletic teams’ rosters to associate student athletes with their respective gendered sports teams.

We used Python libraries like googlesearch, matplotlib, and pandas, alongside R for regressions. We collected data from public Facebook and Harvard websites, and we obtained approval from Harvard Open Data Project to present these summarized results.



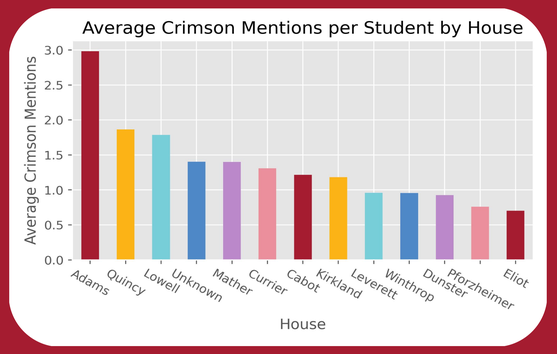



```
Call:
lm(formula = log(1 + numMentions) ~ year + sport + year:sport,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3825 -0.2944 -0.1633 -0.0322  5.0468

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.467078   0.329482   4.453 8.66e-06 ***
year                     -0.395637   0.117307  -3.373  0.00075 ***
sportFemale Sport         0.158441   0.344661   0.460  0.64575
sportMale Sport           0.358999   0.341048   1.053  0.29256
sportNon Sport           -0.910498   0.330194  -2.757  0.00585 **
year:sportFemale Sport   -0.007195   0.122586  -0.059  0.95320
year:sportMale Sport     -0.047943   0.121250  -0.395  0.69256
year:sportNon Sport       0.264543   0.117566   2.250  0.02448 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6015 on 5174 degrees of freedom
  (128 observations deleted due to missingness)
Multiple R-squared:  0.1404,    Adjusted R-squared:  0.1393
F-statistic: 120.8 on 7 and 5174 DF,  p-value: < 2.2e-16
```



Predicting Crimson Mentions

We built an OLS model to predict a student’s number of Crimson mentions given various characteristics of that student. The distribution of number of mentions was very long-tailed, so we performed a log transformation. We used House and Sports Type as categorical features, and we used Class Year as a numerical feature (1–4).

Using a first-order regression, our model had R2→0.119, adj R2→0.113, AIC→9628, and BIC→9858. Our most statistically significant features were Year and Intercept (Adams House and Agender Sport).

We ran a full second-order regression, with R2→0.1523, adj R2→0.1345, AIC→9576, and BIC→10290. The Year:Non Sport interaction had a particularly large positive coefficient and was statistically significant at the 5% level. However, this model had many unnecessary interaction terms, and the low adj R2 was suggestive of overfitting.

So, we ran a stepwise regression to be more selective in including interaction terms. Both AIC and BIC criterion led to the same model: Year, Sports Type, and the Year:Sports Type interaction. This model had R2→0.1404, adj R2→0.1393, AIC→9448, BIC→9507. Of particular note were the Year -0.396 and Year:Non Sport 0.265 coefs. These coefs together imply that athletes improved in Crimson mentions as they aged at a faster rate than did non-athletes. Also, the coef for Male Sport exceeded the coefficient for Female Sport by 0.2. Surprisingly, House was left out of the final model entirely.

Crimson Mentions vs. Sports

Basketball players were the most mentioned in The Crimson, with 249 mentions, followed by ice hockey players at 182, football players at 157, and track and field athletes at 134. Meanwhile, skiers, heavyweight rowers, and cross country runners were all rarely discussed, with 5, 3, and 0 total Crimson mentions, respectively.

These seem consistent with the general coverage of sports outside of Harvard, with large team sports generally being more discussed than individual sports. However, there are definitely exceptions, like baseball players’ lower mentions. Agender sports include only the coed skiing and sailing teams, which were also infrequently mentioned.

On average, every athlete is mentioned approximately 1.94 times in The Crimson, whereas a non athlete is only mentioned around 1.13 times. Assuming a null hypothesis of equal mentioning probabilities, we calculated a z-score of 17.46. So student athletes are indeed more likely to be mentioned in The Crimson as compared to their non-athlete counterparts; we suspect this is due to the Sports section.

Crimson Mentions vs. Class Year

Our data was taken over a four-year timespan, which would be a full cycle from freshman to senior year for the Class of 2021, assuming a four-year college experience. Controlling for number of years spent at Harvard, the Class of 2021 is still the most mentioned class year. The Class of 2024 still has relatively low mentions, which we think could be attributed to how new and unconnected incoming freshmen are.

It is also possible that students are mentioned increasingly in The Crimson as they go through their 4 years at Harvard because they take on more leadership positions and become more involved in campus matters, making them invaluable to Crimson articles. Whatever the case, it seems a safe bet to say that the Class of 2021, the seniors at the time of this data, are the most Crimson-mentioned class year.

Conclusion

The limitations of this research included 1. Inability of named entity recognition to tell the author apart from the article’s subject, 2. Incompleteness of our scraped athletic rosters and Facebook datasets, 3. Human errors such as misspelled names, and 4. Non-optimal model selection due to the greedy approach to add second-order interactions.

We found the most likely demographics to be mentioned in The Crimson were: the Class of 2021, Adams House residents, and male student athletes. It is important that we present only the summarized results here to respect the privacy of the students in this study.