Patent Equity | MIT Collaboration Data Analysis

Jasper Xin, Jiarui Fu, Wanyi Chen, Yuxin Liang, Weirun Huang

Abstract

We collected the name, department, work year, and rank information for 2015 MIT faculty members from the MIT collaboration website. We further acquired the publication dates for 25718 conference proceedings and 131307 articles and the filing dates for 3873 patents (excluding patent applications and those published without USPTO numbers) associated with each of the professors. We found no statistically significant (alpha=0.05) gender disparity between the genders for patents, articles and conference proceedings. We found statistically significant disparity in patents for the Biology and Chemistry department, articles for Architecture and Biology, and conference proceedings for Biology, Brain and Cognitive Sciences, and Mathematics. Analysis using disparity index demonstrates that gender disparity is improving for patents and articles with p-values less than an alpha of 0.05 and both genders are equally represented for conference proceedings. Gender analysis based on "ranks" (e.g. Professor, Associate Professor, etc.) remains inconclusive in statistical testing.

Introduction

Our project is "Patent Equity | MIT Collaboration Data Analysis", with client Prof. Jordi Goodman. She is interested in finding whether there is a correlation between the number of conference proceedings and/or articles and the number of patients each year. Besides, she also aims to find out if professors' race, gender, and faculty level will make a difference in their number of articles, conference proceedings, and patents, and if they do, how does the number change over the years. Our client gets her inspiration from a conference where she got access to the MIT website http://collaboration.mit.edu/.

To lead us to obtain a useful conclusion, Prof. Goodman separated the whole project into three questions that we need to answer during the analysis. The first question is the most general one, which asks us to provide if there is a statistically significant disparity in patent applications, patents, papers, and/or conference proceedings caused by races and/or gender. To be more specific, our client acquired the disparity both in each department and overall in the whole school. Afterward, the question asked us to find out if the disparity has changed over the years and how it has changed. By answering these two questions, we can solve the puzzle - the influence of race or gender on professors' number of articles, conference proceedings, and patents as well as the way the numbers change over time. The last question focuses on the effect of the faculty level of the professors. We are not directly finding the level but rather finding the years that each professor stays in MIT. By figuring out if newer professors experience similar gaps to older professors, we can successfully demonstrate the outcome of faculty level by years on the number of conference proceedings and/or articles and the number of patents.

Data Collection / Preprocessing

MIT Collaboration Website

The MIT Collaboration Website is our primary source of information. It provides a list of MIT faculty members for us to work with, several important attributes (e.g. name, department, number of articles), as well as the information for individual patents, articles, and conference proceedings related to a faculty member. However, scraping the website has proven to be difficult, and some of the information (work years, USPTO numbers, publishing dates) requires indirect approaches to obtain.

The MIT collaboration website does not provide a convenient way to access the dataset it uses, and thus we had to utilize two undocumented APIs, the SearchPersons API and Productivity API, which are discovered via Chrome's developer tools, to access most of the information. We set up an automated pipeline for acquiring the data from MIT collaboration's website using Python and its *requests* and *pandas* module.

The SearchPersons API provides information about the names, departments, and the number of patents, articles, and conference proceedings for each of the faculty members. In our tests, we found that the SearchPersons API that the MIT collaboration website uses only returns a maximum of 100 rows of data per query meaning that some faculty members' information might be unavailable to us if they are intended to be from queries that exceed the 100 rows limitations. To get around this limitation, we combined the results (while eliminating duplicates) from three scraping approaches to maximize our dataset. After looking through the source code of the MIT collaboration's website, we found the IDs of the departments and centers. The IDs enabled us to scrape the website twice, once by using the departments as a filter, once by using the centers as a filter. We are able to acquire 1312 professors by combining the datasets obtained from the two approaches compared to 1000~ scraped from each approach alone. Then we applied another scraping approach by using the work years as a filter and acquired an additional 89 professors in the process, summing to a total of 1401 professors. Moreover, by using the work years as a filter, we are able to determine the information of what professors work for MIT in what years. Finally, we applied an extensive approach to ensure we have the fullest dataset available. We performed a search using each of the alphabets as well as all the combinations of two alphabets as the field for the name of the professor. Using this method, we managed to scrape a total of 2015 professors, which we used as our final dataset. We are confident that our dataset scraped from this combined approach will be very close to the actual dataset used by the MIT collaboration website.

The Productivity API provides information for the individual patents, conference proceedings, and articles for a designated faculty member. However, the API returns data in an HTML code format because it is intended to be directly run by the browser to display the data instead of being used as a table. Therefore, we had to decode the raw data returned by the API and use pattern matching (regex) to find the information we needed. In this process, we acquired

the publishing dates for the patents, articles, and conference proceedings as well as the USPTO numbers for the patents.

USPTO Database

Because the MIT collaboration website only returns the publication dates for the patents and the client indicated a preference for the use of filing dates, we used the USPTO numbers scraped from the MIT collaboration website and the USPTO database to acquire the filing dates for the patents that have USPTO numbers. We used the bulk data acquired from the Patent Examination Data System[1], which contains information related to every patent application submitted since 2000. In the process, we are to link 3873 patents (3018 unique) out of 6179 patents found in the dataset scraped from the MIT collaboration website to those found in the USPTO's dataset. We found that many of the original set of 6179 patents are patent applications or have filing dates earlier than 2000 and thus we discarded them from our analysis.

Gender Prediction

We applied an online API called *gender-api*[2] to predict the gender of each professor based on their first names. To demonstrate the reliability of gender-api as our model for assigning gender, we cross-referenced our results with Prof. Goodman results. We found 613 matches out of 692 professors in Prof Goofman's list; unmatched entries are likely due to a typo in the name or misformatting. In the matched dataset, *gender-api* achieved an accuracy of 0.906 in predicting the gender of the professors (script for analysis can be found in gender_analysis.ipynb). A peer-reviewed study[3] showcases a similar finding to ours, with *gender-api* achieving highest accuracy out of all current name-to-gender inference services at a misclassification rate of less than 5%. Therefore, we determined that *gender-api* is a reliable tool that is suitable for our use cases.

There are unknown values returned by *gender-api*, which is likely caused by the lack of data for those names in *gender-api*'s database. For the unknown values, we performed a manual search for the faculty members on Google to determine their genders. *gender-api* also reports the accuracy for their predictions. As an extra effort to achieve higher quality for our data, for those predictions that have accuracies less than 97%, we performed a manual search as well.

[1] https://ped.uspto.gov/peds/#!/#%2FapiDocumentation
[2] https://gender-api.com/
[3] https://peerj.com/articles/cs-156.pdf

Preliminary Analysis

Using the methods mentioned above, we obtained a sample size of 2015, 1544 of which are male and 471 female. To have a better understanding of our data, we performed analyses on several aspects of the dataset. A few of the first fields we looked at were the total and average count of patents, articles, and conference proceedings for male and female faculty members and the number of male and female faculty members per department, as shown in the plots below.

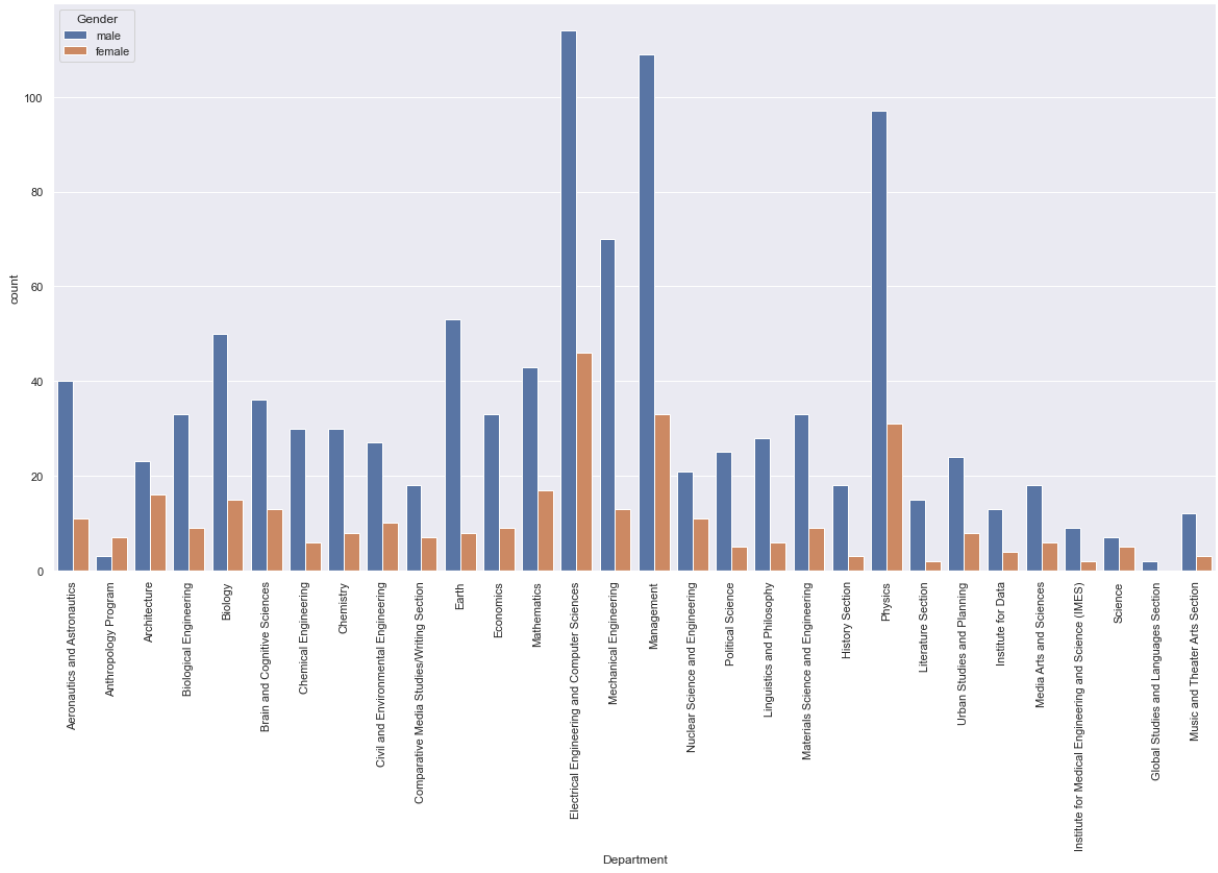

Figure 1. Number of male and female faculty members in each department

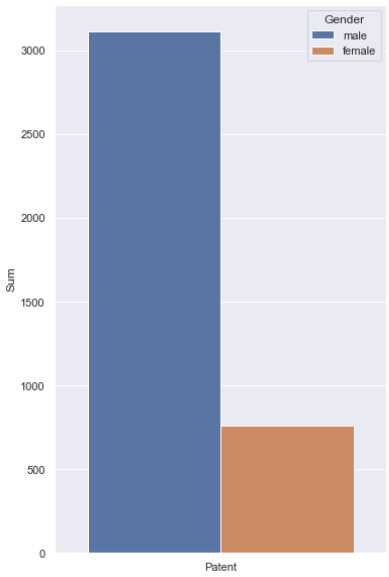

Figure 2.1. Sum of patent count for each gender

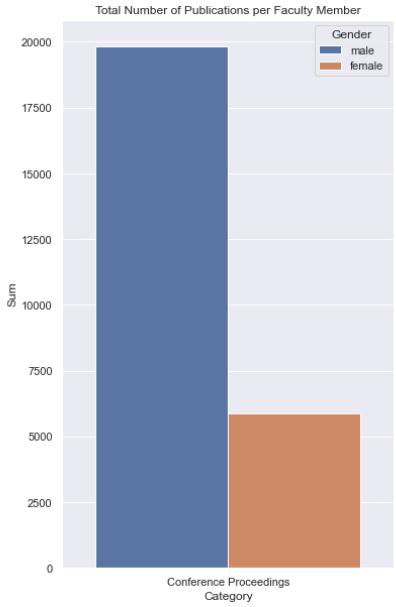

Figure 3.1. The Sum of conference proceedings count for each gender

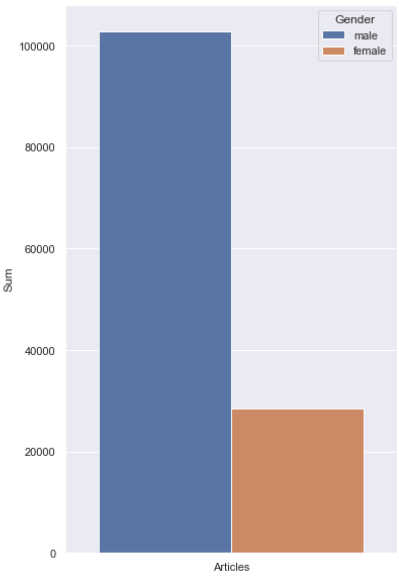

Figure 4.1. Sum of article count for each gender

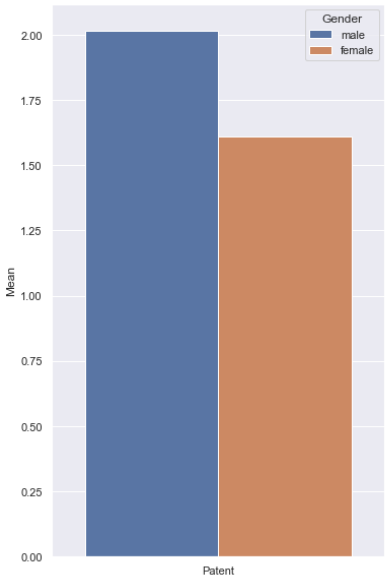

Figure 2.2. Average of patent count for each gender

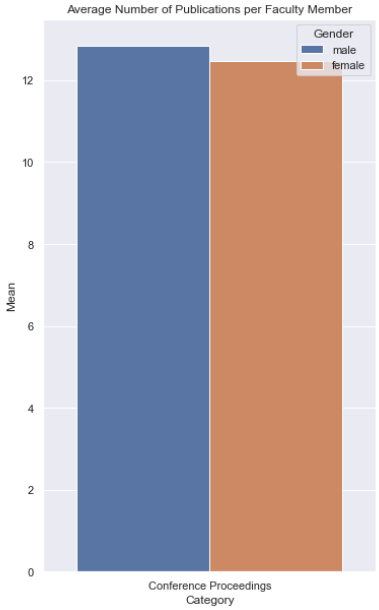

Figure 3.2. The average of conference proceedings count for each gender

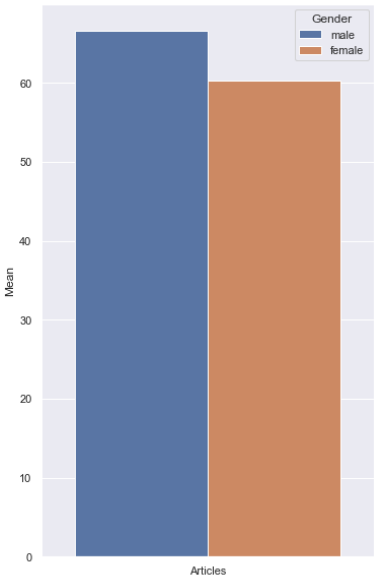

Figure 4.2. Average of article count for each gender

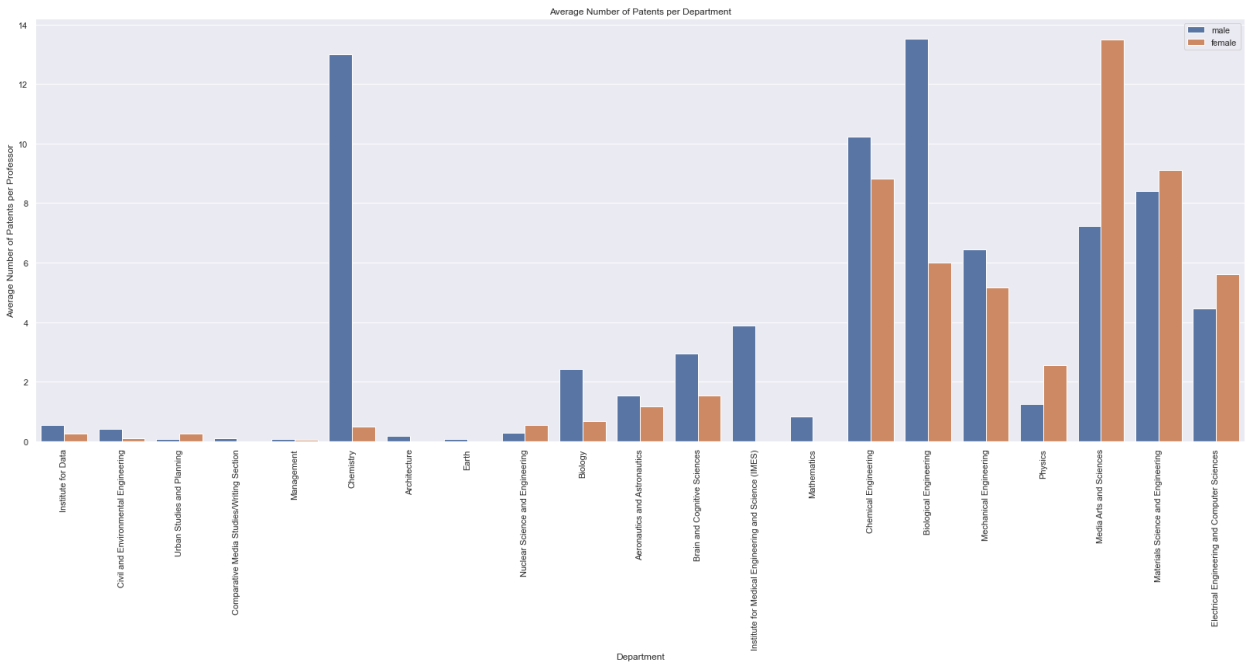

Figure 5. The bar graph shows the average number of patents per department. Blue represents female professors and orange represents male professors.

For all female professors, there are approximately 750 patents, while male professors have more than 3000 patents (Figure 2.1). The average patent count for males and females from Figure 2.2 shows that the average for male professors is roughly 125% the amount for females, further suggesting a disparity. We also grouped the number of patents by departments to investigate gender representation on the average patent count in each department (Figure 5). The result shows that several of the departments see female professors having an average number of patents matching or even higher than their male peers, notably Physics, Media Arts & Science, Nuclear Science and Electrical Engineering & Computer Science department, which has female professors acquiring on average several times more patents than the male professors respectively. However, the large disparity occurs in most departments, with a number of departments having no patents for female professors.

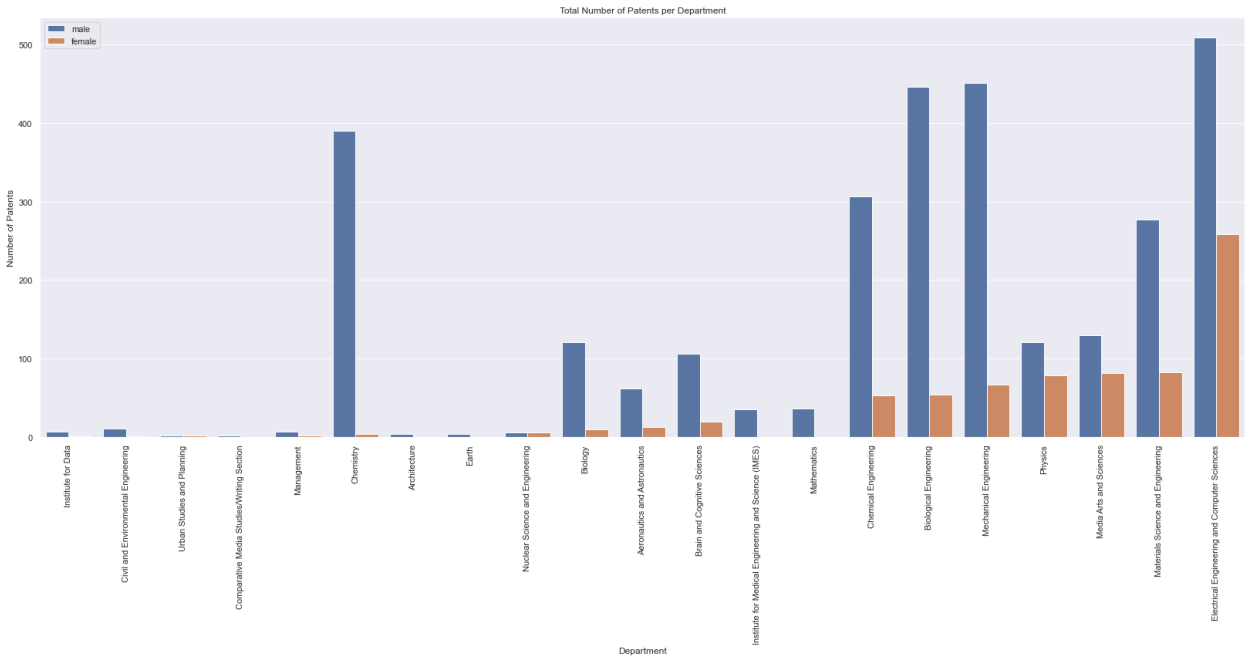

Figure 6. The bar graph shows the total number of patents per department. Blue represents female professors and orange represents male professors.

Afterward, we plotted the total number of patents filed for each year to see if the disparity is changing over time as figures 7. As shown in the graphs, we can see that the number of patents that belong to a female inventor increased steadily from 2000 to 2010 and remained roughly constant through 2011 to 2018.

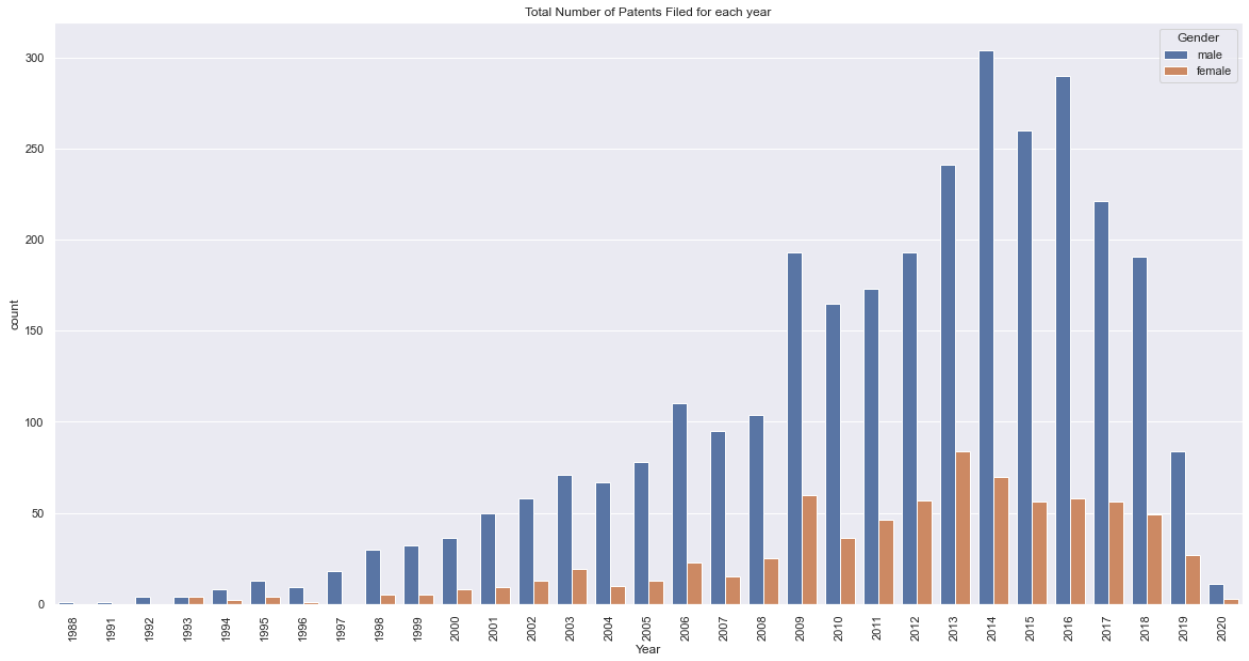

Figure 7. The plot shows the total number of patents for each gender in each year. Blue represents male professors and orange represents female professors.

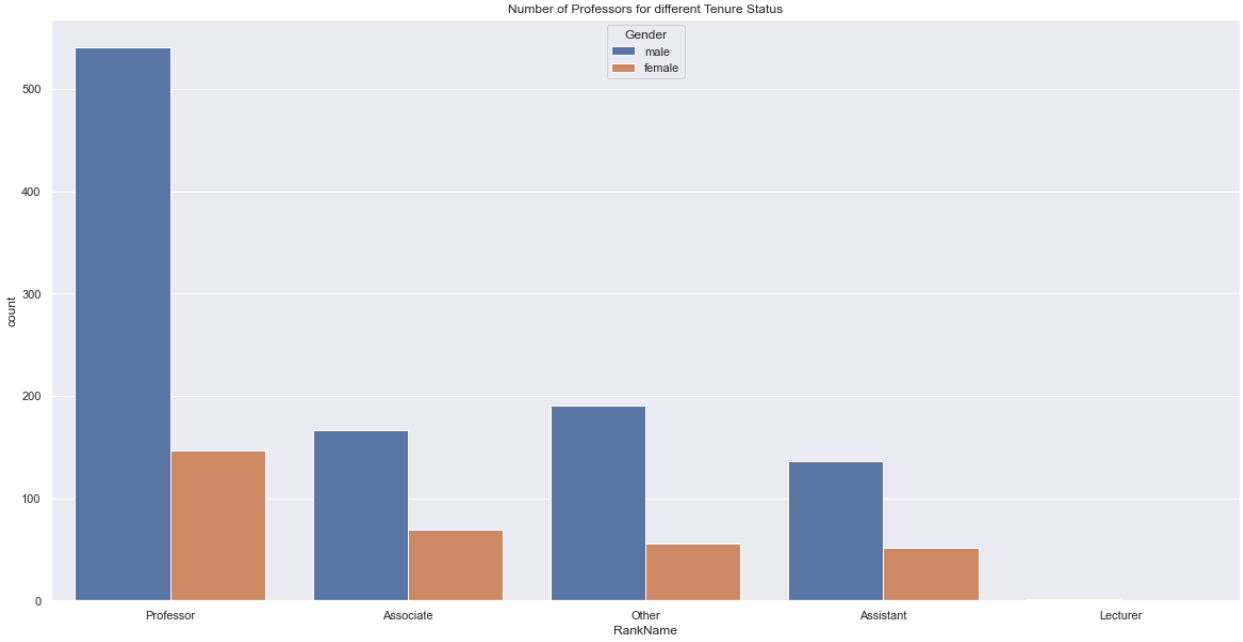

Figure 8. The plot shows the distribution of faculty members based on tenure status

In our next stage of analysis, we aimed to compare the average in patents, articles, and conference proceedings between newer & older and male & female professors, to get a brief idea of how this disparity between gender affects different people differently. Although we were able to obtain the work year information for most faculty members, the result of the tests when we partitioned the dataset by year was not ideal. An alternative method of comparing older and newer faculty members was suggested, which is to partition the dataset based on faculty ranking. From the MIT Collaboration website, we were able to scrape faculty members with rankings of professors, associate professors, assistant professors, lecturers, and others (Figure 8).

Figures 9, 10 & 11 display the total patents, total conference proceedings, and total article count for 4 of the 5 different ranks. This is due to the fact that there was only 1 lecturer found and that he/she had 0 patents, conference proceedings, or articles. From the graph, we can see that across the board professors hold the majority of patents, conference proceedings, and articles. However, this large difference may disappear once we compare the average values in later sections of the report.

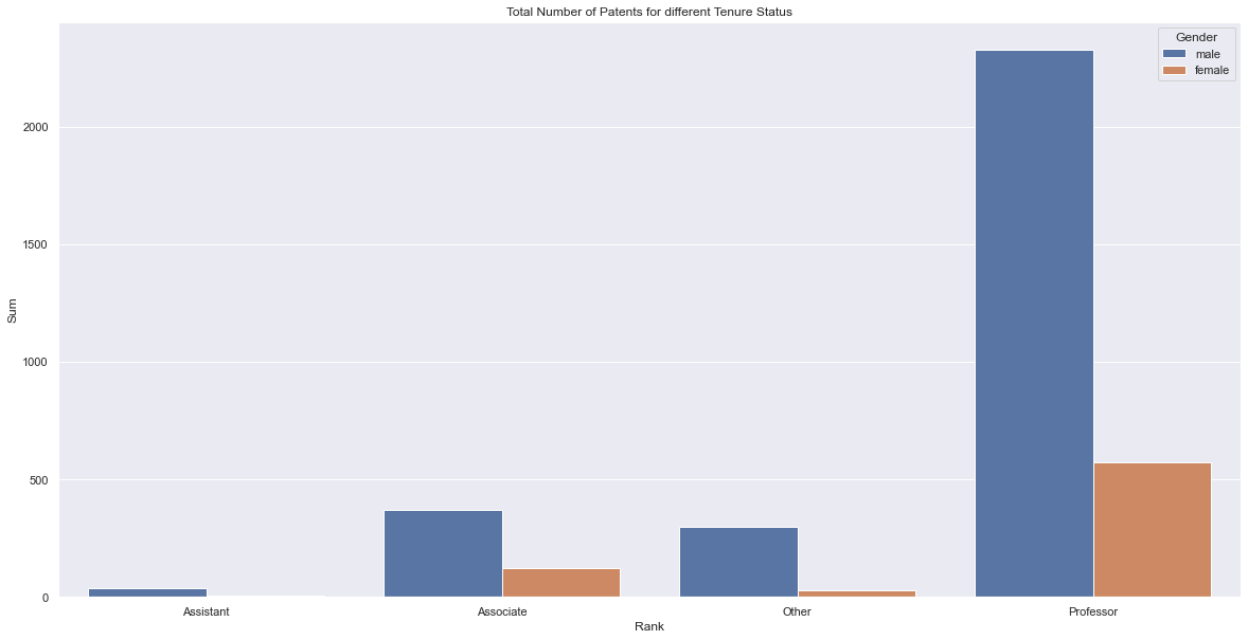

Figure 9. Bar graph of average patents between newer professors and older professors

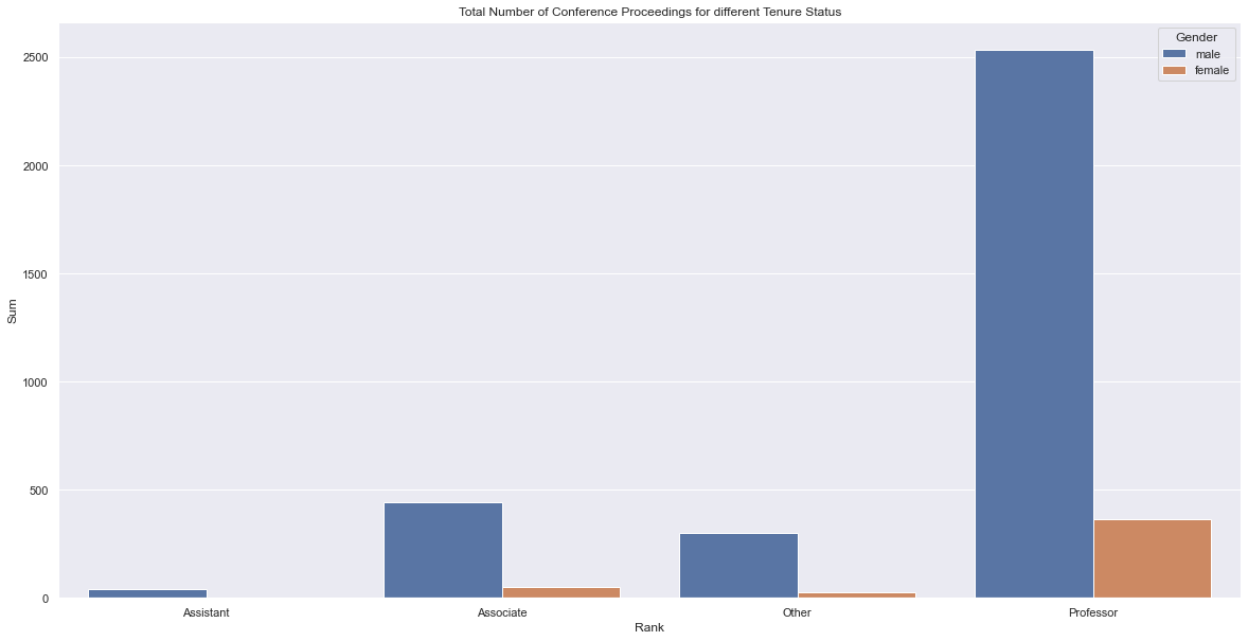

Figure 10. Bar graph of average conference proceedings between newer professors and older professors

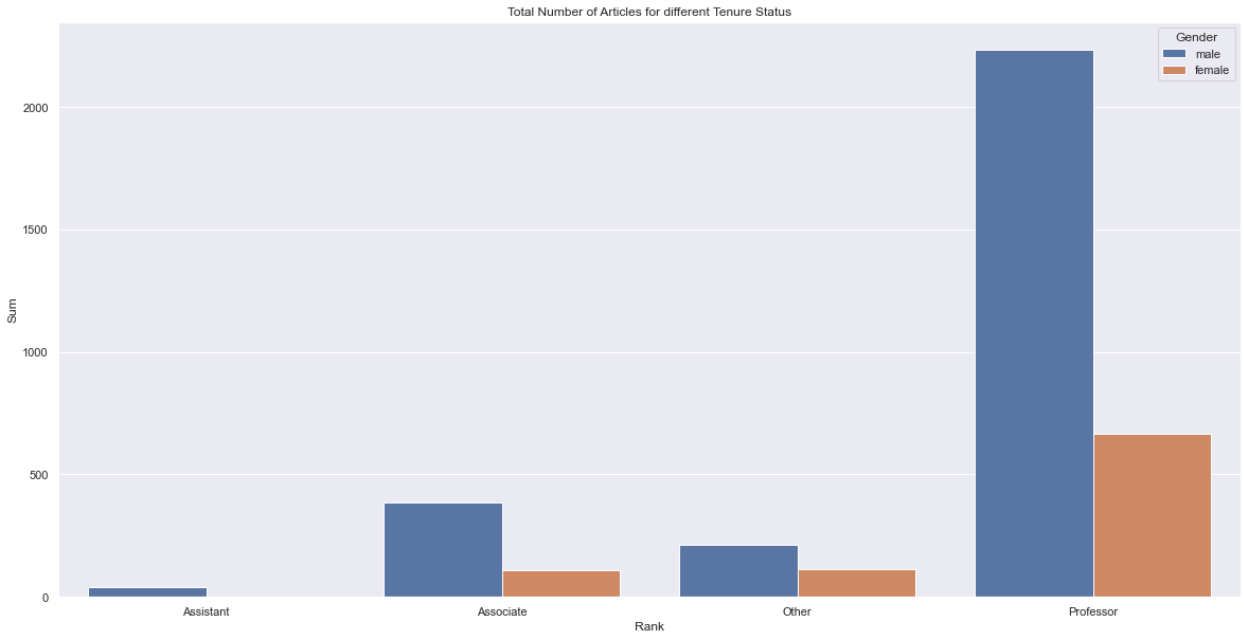

Figure 11. Bar graph of average article between newer professors and older professors

Key Question 1

"Is there a statistically significant disparity by gender in patents, papers, and/or conference proceedings relative to the representation in each department or overall at MIT?"

To answer this question, we coded programs that could automatically perform the Student's T-test between the average value for males and females on the field of interest (patent, article, or conference proceeding count). If the variance between the two groups was significantly different the code would perform the Welch's T-test instead for a more accurate test result. The null hypothesis is that there is no disparity between male and female faculty and the alternative is that there is a statistically significant disparity. The confidence level used is 95%, but this can be easily changed by changing the corresponding alpha value within the code. The result for MIT overall is as follows:

- For Patents, we failed to reject the null hypothesis at alpha = 0.05. This means that we don't have enough evidence to prove that there is a significant disparity between male and female faculty.
- For Articles, we failed to reject the null hypothesis at alpha = 0.05. This means that we don't have enough evidence to prove that there is a significant disparity between male and female faculty.
- For Conference Proceedings, we failed to reject the null hypothesis at alpha = 0.05. This means that we don't have enough evidence to prove that there is a significant disparity between male and female faculty.

We also partitioned the faculties based on department and performed a similar test for each department. For departments that have 0 male or 0 female faculty, the results are insignificant therefore hidden, though a message will be displayed stating the issue. Figure 12 is a sample of the quick summary for each department, for the full result or interpreted result, please see the final deliverable submitted.

```
Legend: Male = mean of male is significantly greater than mean of female
        Female = mean of female is significantly greater than mean of male
        Invalid = not enough data to perform test
        No Disparity = no disparity found
```

Department	Patent	Article	Conference Proceedings
Aeronautics and Astronautics	No Disparity	No Disparity	No Disparity
Anthropology Program	No Disparity	No Disparity	No Disparity
Architecture	No Disparity	Male	No Disparity
Biological Engineering	No Disparity	No Disparity	No Disparity
Biology	Male	Male	Male
Brain and Cognitive Sciences	No Disparity	No Disparity	Male
Chemical Engineering	No Disparity	No Disparity	No Disparity
Chemistry	Male	No Disparity	No Disparity
Civil and Environmental Engineering	No Disparity	No Disparity	No Disparity
Comparative Media Studies/Writing Section	No Disparity	No Disparity	No Disparity
Earth	No Disparity	No Disparity	No Disparity
Economics	No Disparity	No Disparity	No Disparity
Mathematics	No Disparity	No Disparity	Male
Electrical Engineering and Computer Sciences	No Disparity	No Disparity	No Disparity
Mechanical Engineering	No Disparity	No Disparity	No Disparity
Management	No Disparity	No Disparity	No Disparity
Nuclear Science and Engineering	No Disparity	No Disparity	No Disparity
Political Science	No Disparity	No Disparity	No Disparity
Linguistics and Philosophy	No Disparity	No Disparity	No Disparity
Materials Science and Engineering	No Disparity	No Disparity	No Disparity

Figure 12. Sample of the disparity test result for each department

Key Question 2

"Has this disparity changed over time and, if so, in what way?"

To answer this question, we looked for trends in how the Disparity Index has changed over time for each gender considering patents, articles, and conference proceedings and applied statistical methods to test for the statistical significance of the trends. The Disparity Index is calculated as $D = P_i/P_j$, where P_i is the percent of patents for male/female for a given year, and P_j is the percent of male/female working professors in MIT for a given year. To test the statistical significance of the change in the Disparity Index over time, we test the null hypothesis for males and females respectively. We set the null hypothesis to be that the slope of the linear regression is 0, while the alternative hypothesis is that the slope is not 0.

In Figure 13, we plotted the Disparity Index of male and female professors from 2004 to 2020 for patents. For male faculties, the p-value was 0.0483 which meant that the slope is significant. However, the value of the slope was roughly -0.0044, a very small value suggesting a slow decrease. When it comes to female faculties, the p-value becomes 0.0381 and the slope 0.0158, since the p-value is smaller than alpha, we reject the null hypothesis. Together with the results of the tests and the trend line, we conclude that the disparity index of the two genders is converging to 1.0 in recent years and the disparity is decreasing for male and female inventors.

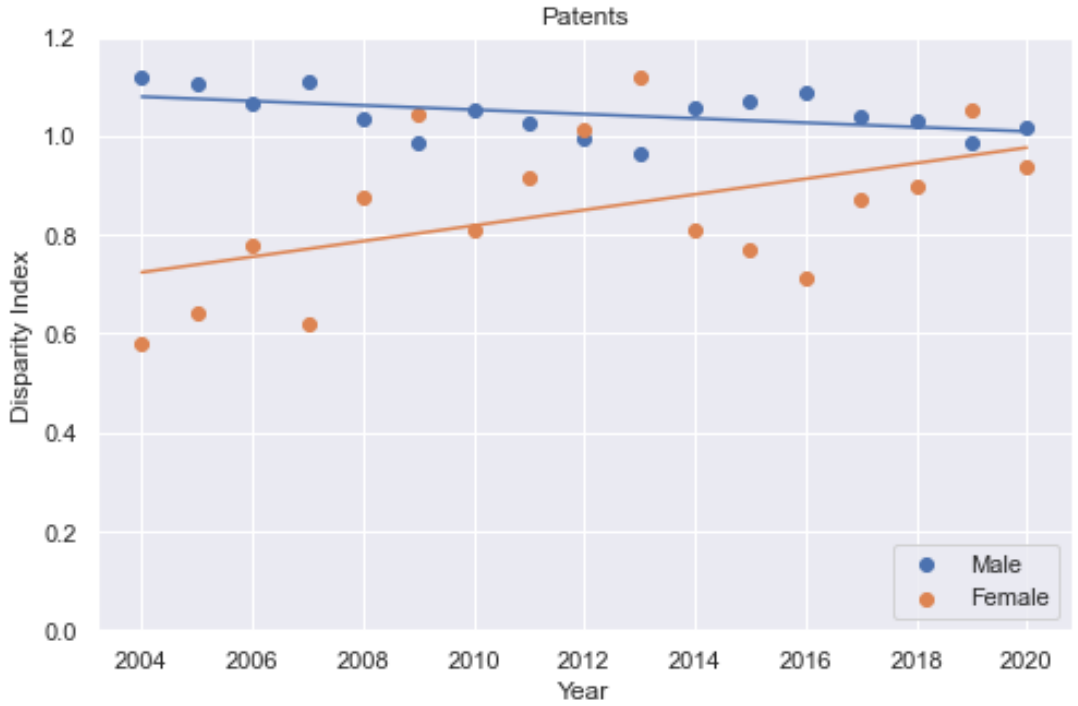

Figure 13. The graph shows the Disparity Index of male and female professors in patents from 2004 to 2020 and the trend lines for each gender.

We plotted how the articles' Disparity Index has varied over time for male and female faculty members in Figure 14. After performing linear regression on the data, we found that the p-value for both male and female faculty members is very close to 0, which means that we reject the null hypothesis for both tests. For male faculty members, the slope for the disparity index is -0.0029 which suggests a slow decrease. For female faculty members, the slope is 0.0102, suggesting a gradual increase. Again we see that the disparity index for both genders to been converging at 1.0 in recent years, signifying that the disparity between gender has been decreasing through time.

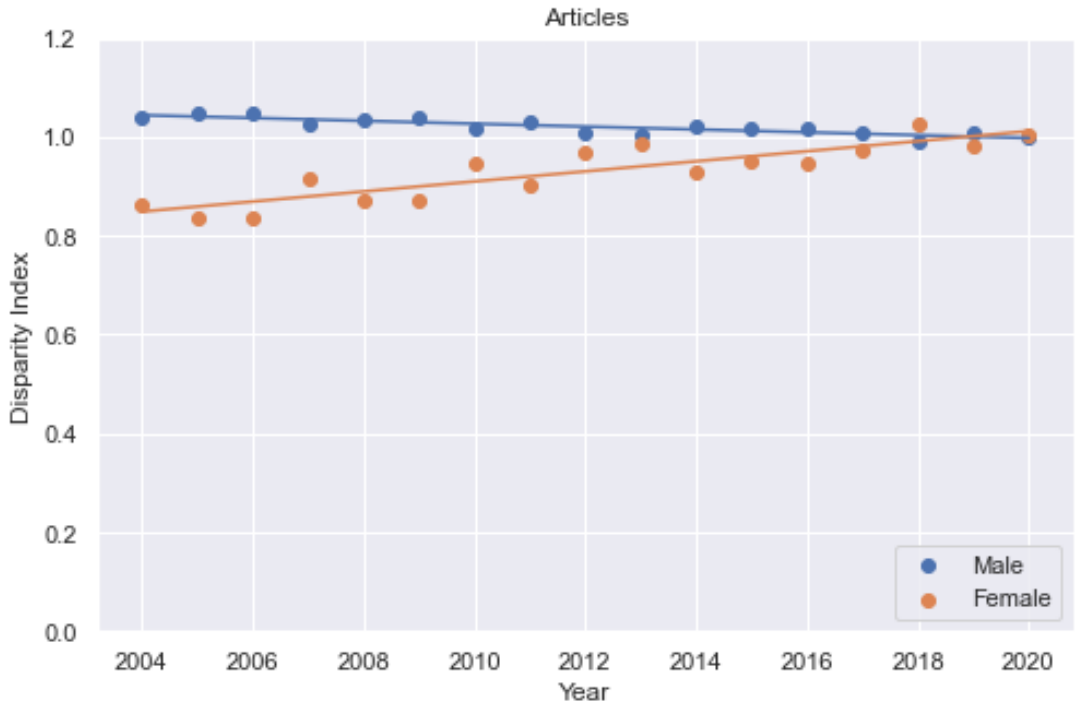

Figure 14. The graph shows the Disparity Index of male and female professors in articles from 2004 to 2020 and the trend line for each gender

Similarly, for conference proceedings, we plotted how the Disparity Index changes over time as well as the Linear Regression results for male and female faculty members (Figure 15) and tested the null hypothesis. The p-value for males is 0.3036 and for females is 0.2889 which means that we failed to reject the null hypothesis for both genders and that the disparity index for both genders has remained relatively constant. If we look at Figure 15 we will find that the

representation of both genders in terms of conference proceedings has been relatively balanced throughout the past two decades, floating around 1.0, which supports the conclusion of the test. Though there is some variance for the indices, especially between 2007 - 2014, to the eye there is no evidence of a clear trend.

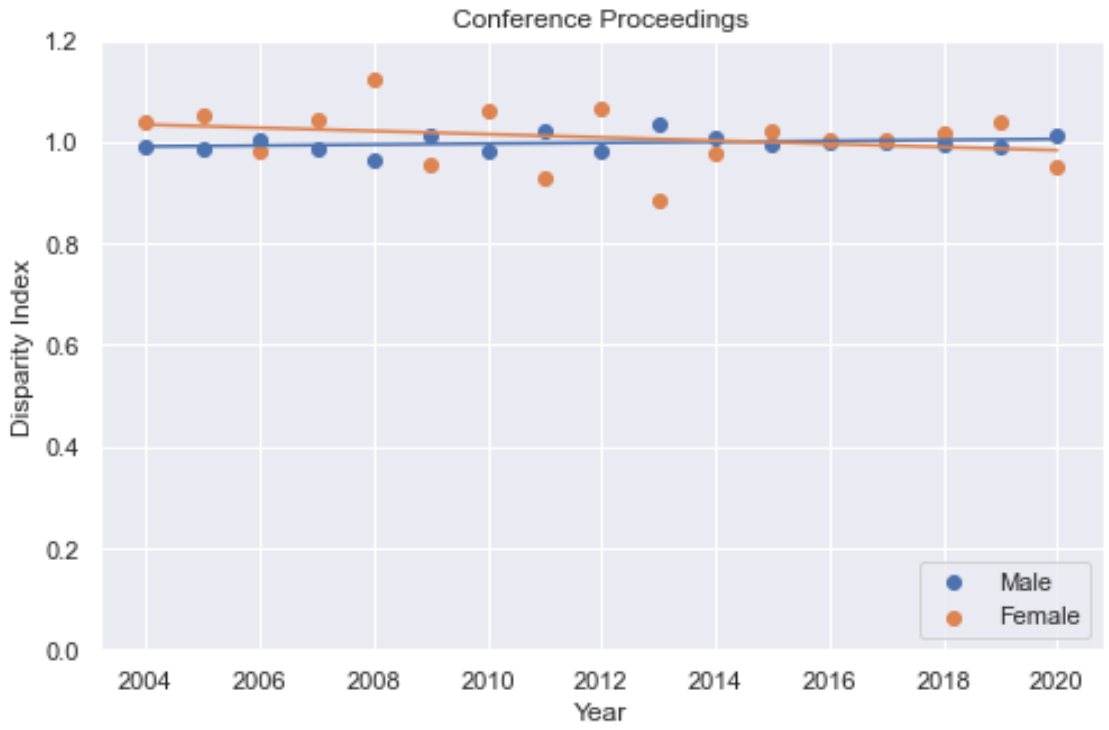

Figure 15. The graph shows the Disparity Index of male and female professors in proceedings from 2004 to 2020 and the trend line for each gender.

Key Question 3

"Are newer professors (professors who have been at MIT less than ten years) experiencing similar gaps to older professors?"

As discussed in earlier deliverables, partitioning faculty members based on work year did not yield favorable results. An alternative method was to partition the faculty members based on their tenure status. We've decided to compare faculty members with professor status versus faculty members with associate professor, assistant professor, lecturer, and other statuses. This is based on the theory that it takes several years to reach a status of professor, whereas it would take less time to achieve other statuses. In the following analysis, faculty members with professor status will be addressed as older faculty members, and faculty members with the status that is not a professor will be addressed as newer faculty members.

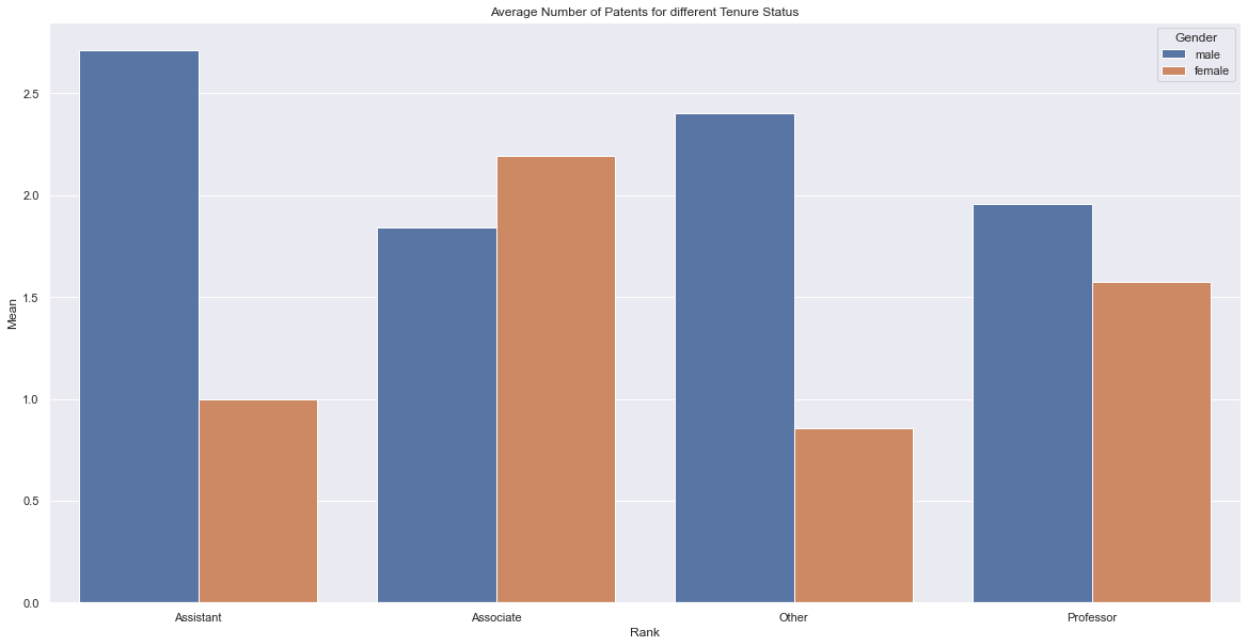

Figure 16. The plot shows the average number of patents for different tenure status

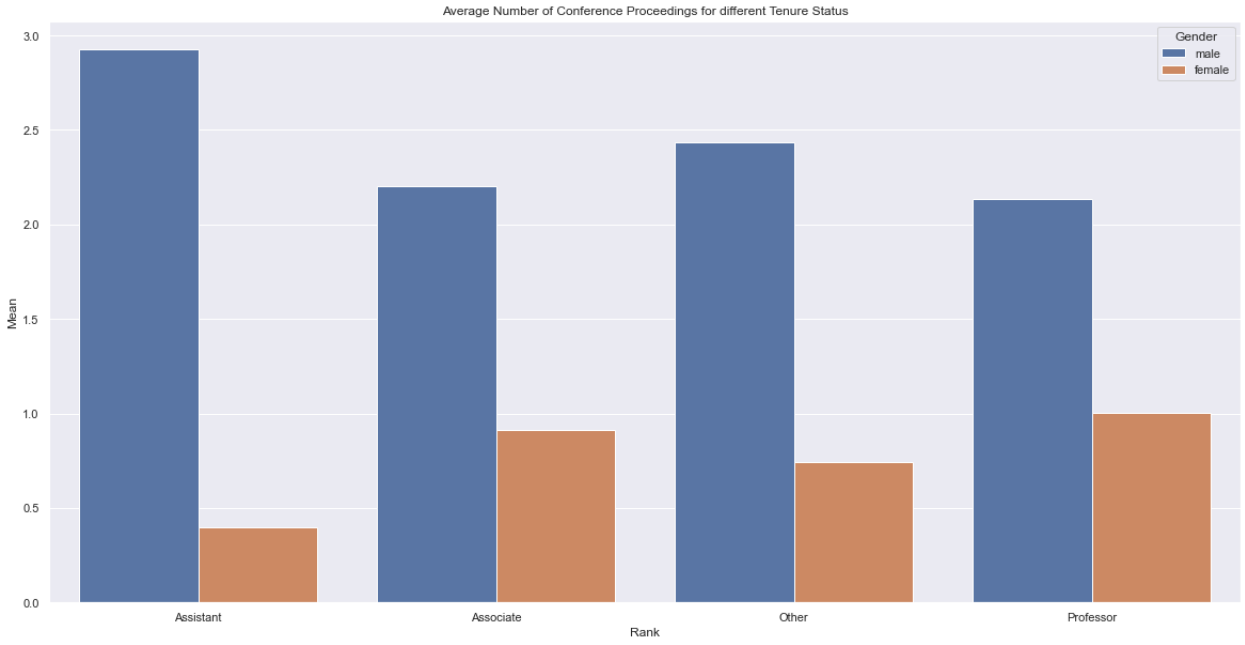

Figure 17. The plot shows the average number of conference proceedings for different tenure status

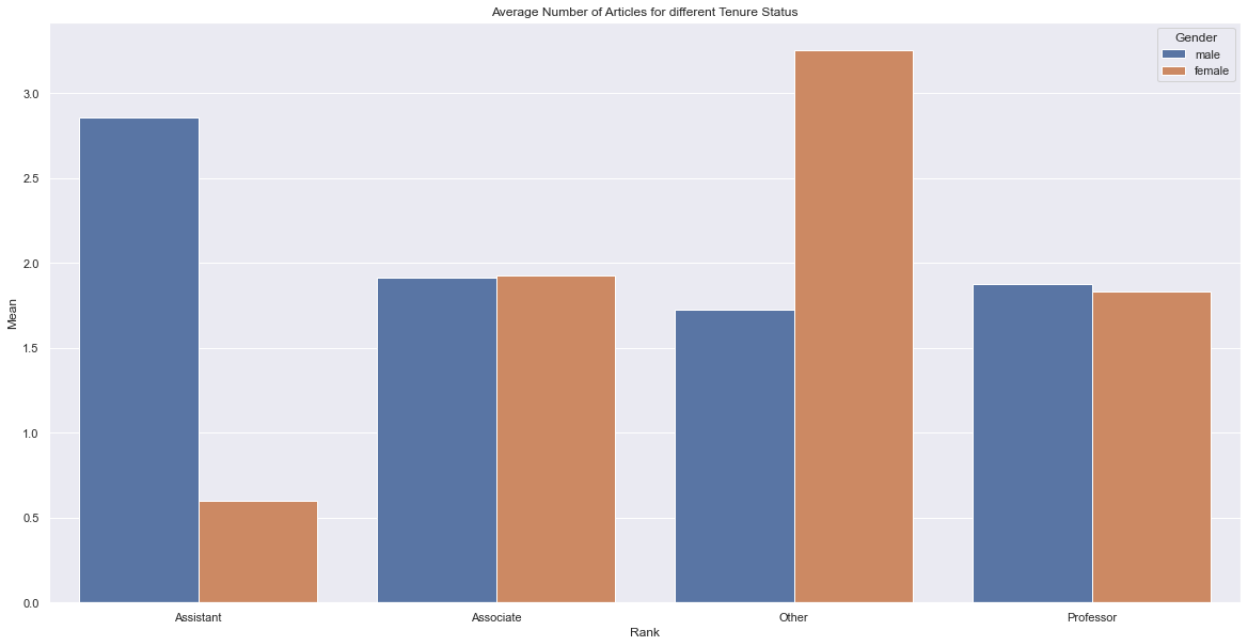

Figure 18. The plot shows the average number of articles for different tenure status

Figures 16, 17 & 18 show the average amount of publications for each tenure status. We can see that although the disparity between gender for each rank still exists, the difference is much less than the total value displayed in the preliminary section.

We decided to perform a disparity test for gender between older and newer faculty members independently and see if the disparity persists. We also tested for disparity over the union of the two subsets as a reference.

We used methods similar to those described for key question 1 to determine if there is a significant difference between the mean of the two groups. We have the two hypotheses as follow:

Null Hypothesis: mean of male = mean of female

Alternative Hypothesis: mean of male ≠mean of female

By using the T-Test, we got the following result(Figure 19):

```
Summary

For Total Patent Count
Overall:  True  Rank1:  True  Rank2:  True
For Total Article Count
Overall:  True  Rank1:  True  Rank2:  True
For Total Conference Proceedings Count
Overall:  True  Rank1:  True  Rank2:  True

Legend: True := H0 not rejected, no evidence for disparity, False := H0 rejected, significant disparity
```

Figure 19. Raw test result for gender disparity for overall, older and newer.

- For patents, we failed to reject the null hypothesis for the overall sample size and the two separated groups, which means we do not have enough evidence to show that there is a gender disparity within older faculty members or newer faculty members.
- For articles, we failed to reject the null hypothesis for the overall sample size and the two separated groups, which means we do not have enough evidence to show that there is a gender disparity within older faculty members or newer faculty members.
- For conference proceedings, we failed to reject the null hypothesis for the overall sample size and the two separated groups, which means we do not have enough evidence to show that there is a gender disparity within older faculty members or newer faculty members.

Limitations

We want to note that the list of professors we acquired is likely incomplete compared to the dataset available on MIT's end due to inherent flaws in the MIT collaboration website's API although we demonstrated in the *Data Collection* section that we took several approaches to maximize the dataset available to us.

Additionally, we had to use an API to predict the gender of the faculty members, while doing so, we are sacrificing some accuracy. However, as we demonstrated in the *Data Collection* section, the *gender-api* achieves good accuracy across the board, and we manually acquired the gender of the faculty members that *gender-api* is uncertain with.