

# **Patent Equity | MIT Collaboration Data Analysis**

Guo-Yi Jenny Wong, Yining Tang, Liqiu Li, Yuying Wu

November 19, 2021

## 1 Introduction

For the project of MIT Patent Equity Analysis, the main data source is the Faculty Research Collaboration Tool<sup>1</sup>, provided by MIT as a website. MIT provides this online database containing the data of its faculties, including the basic information including name, department, professional title; along with the number of articles they have published, the number of conference proceedings, grants, and the number of patents associated with each professor.

In this project, the main goal is to see if there is a correlation between the numbers, also if the race, gender, or tenure status of the professor makes any difference. We also wonder if race or gender has any impact on the numbers.

We have made several steps to reach the final result. First of all, to get the data ready to use in the analysis, we chose to scrape all information needed for analysis from the website.

## 2 Data Preprocessing

### 2.1 Web Scraping

In order to obtain the number of articles, conference proceedings, grants, and patents, we scrape the website of the collaboration tool. We first looked at the web page source code using the Chrome Developer Tools, and tried to call the get method from the Python library requests to generate a json file for further data analysis. However, this approach fails due to insufficient information returned. We can get correct numbers for articles, conference proceedings, and grants, but missing data of patents number. However, we found out that we can get the department ids and their associate professor ids using this method.<sup>2</sup>

After several trials on different methods, we finally decided to use the post method of requests, iteratively obtain the HTML source code of each professor, and find the data we need from the source code.<sup>3</sup>

Before we start scrape the website, we took a look at the faculty profile page of each department, and finds out that some departments do not necessarily have the patents or grants number. Thus, we only include the departments of engineering and natural sciences. As a result, we include 13 departments and 692 faculties in our initial faculty dataset.

---

<sup>1</sup><http://collaboration.mit.edu/>

<sup>2</sup> Code see ./code\_data/code/web\_scraping\_id.py

<sup>3</sup> Code see ./code\_data/code/web\_scraping\_main.py

## 2.2 Gender Classification

Since we do not have gender information in the online database, we decided to find a large database of first names and their corresponding gender to perform classification training to obtain a model. The database we used for training is a Gender by Name Data Set provided by the University of California, Irvine, Center for Machine Learning and Intelligent Systems (UCI).<sup>4</sup> In this dataset, there are 147269 rows, with 4 columns including first name, gender, count, and the probability. We used the first name and gender from this dataset, split them into training set and testing set with the same size after vectorization, then trained them with several different classification algorithms. See Table 1 below for the accuracy scores.<sup>5</sup>

Classification Algorithm	Training Accuracy	Testing Accuracy
Multinomial Naive Bayes	94.71%	58.01%
Bernoulli Naive Bayes	61.82%	61.64%
Random Forest	61.36%	61.44%

Table 1: Training and testing accuracy scores of algorithms used.

As we can see, the testing accuracy scores of all three algorithms do not have a big discrepancy, that the difference between best score and the worst score is only 3.63%. But if we look at the training accuracy scores, the result of multinomial NB classifier has a significantly higher accuracy score than other two classifiers. Thus, we decided to use the classifier which uses multinomial NB algorithm to classify the gender of the MIT faculties.<sup>6</sup>

## 2.3 Race Classification

For the same reasons mentioned in the previous section, we do not have the race information of each faculty. To classify the race, we refer the API released by the United States Census Bureau. This database contains the 151671 rows and 11 variables in total. Among the variables, I chose the surname, PCT2PRACE, PCTAIAN, PCTAPI, PCTBLACK, PCTHISPANIC, and PCTWHITE which are the races from the database, to train the classification model.<sup>7</sup>

We first tried to match the surnames in our faculty dataset with the surnames in the US Census database. The result turns out that 506 faculty surnames are included in the database, and 186 the faculty surnames are not included in the database.

---

<sup>4</sup> <https://archive.ics.uci.edu/ml/datasets/Gender+by+Name>

<sup>5</sup> Code see ./code\_data/code/gender\_classification.ipynb

<sup>6</sup> Data see ./code\_data/data/faculties\_with\_gender.csv

<sup>7</sup> See variable description here: <https://api.census.gov/data/2000/surname/variables.html>

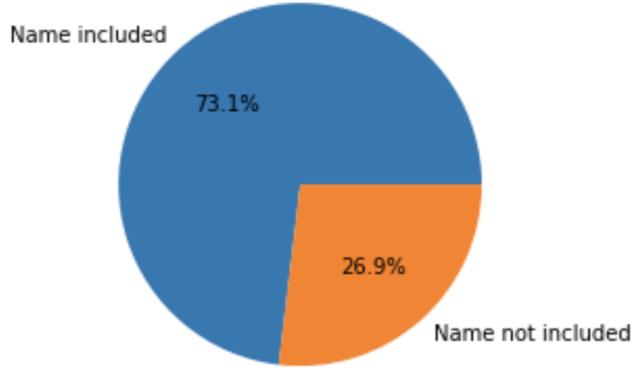


Figure 1: The pie chart of percentage names included in the database.

As we can see from Figure 1 above, 73.1% of the names matched, and 26.9% were not matched in the surname-race database. Thus, we split the surnames into two parts: for the surnames that are matched in the database, we simply assign the race with the highest percentage given in the race attributes, while for the surnames that are not matched in the database, we trained a classification model and apply the model on them.

In the model training step, we extract the surname as the data X and the corresponding race with the highest percentage possibility as the label y. After vectorization, we split the dataset into training and testing sets and applied the Multinomial Naive Bayes algorithm first since it has the best performance in our previous model. The result turns out to be surprisingly good, with a training accuracy score of 85.72% and a testing accuracy score of 86.33%.<sup>8</sup> After classification on race, we merge this result into the dataset with gender classified.<sup>9</sup>

## 2.4 Year of Publication

Since we need to answer the question of disparity over time, we would like to get the information of the year when the articles were published, when the conference proceedings were held, when the grants were given, and when the patents were published. We found out that in the website there are specific years and dates under each specific work, thus, we also obtain a dataset containing the years for each faculty, along with the count of each category.<sup>10</sup>

---

<sup>8</sup> Code see ./code\_data/code/race\_classification.ipynb

<sup>9</sup> Data see ./code\_data/data/faculties\_gender\_race.csv

<sup>10</sup> Data see ./code\_data/data/faculties\_gender\_race\_year.csv

## 2.5 List of Department

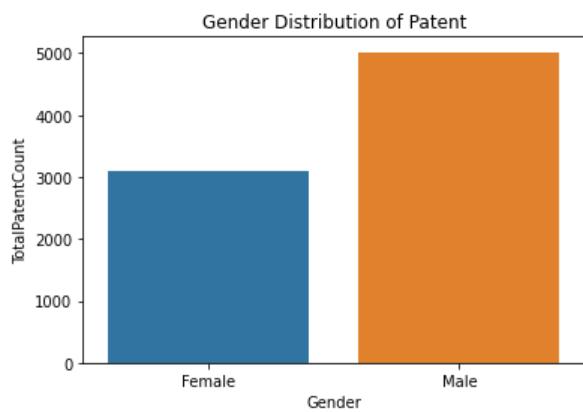
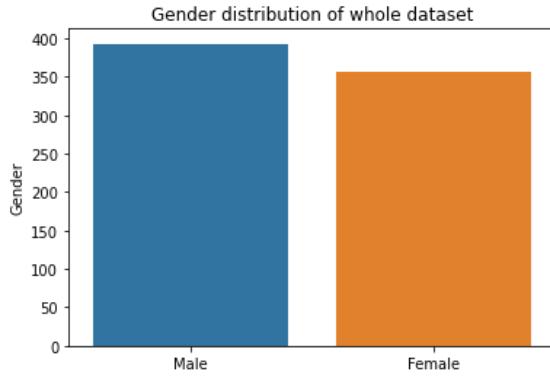
Since we found out that in some departments, the faculties do not necessarily have patent or grant works, to reduce the possible skewed distribution, we decided to eliminate those departments. Here is the list of departments included in our analysis:

Aeronautics and Astronautics  
Biological Engineering  
Biology Brain and Cognitive Sciences  
Chemical Engineering  
Chemistry  
Civil and Environmental Engineering  
Electrical Engineering and Computer Sciences  
Materials Science and Engineering  
Mechanical Engineering  
Media Arts and Sciences  
Nuclear Engineering  
Physics

### 3 Preliminary Analysis

#### 3.1 Preliminary Analysis by Gender

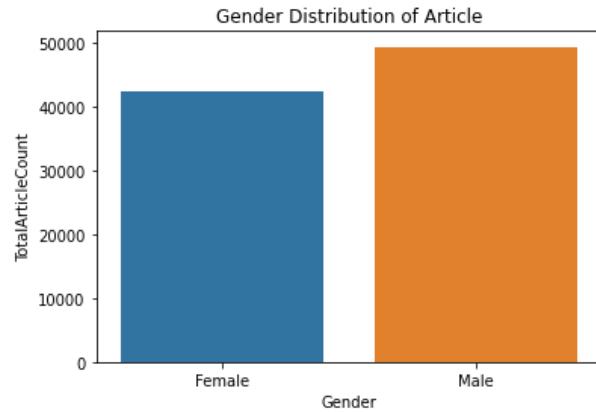
The gender distribution of the whole dataset is shown in the graph below, and we analyze the gender associated with all the heads in the dataset.



Patent Distribution:

Female 3086

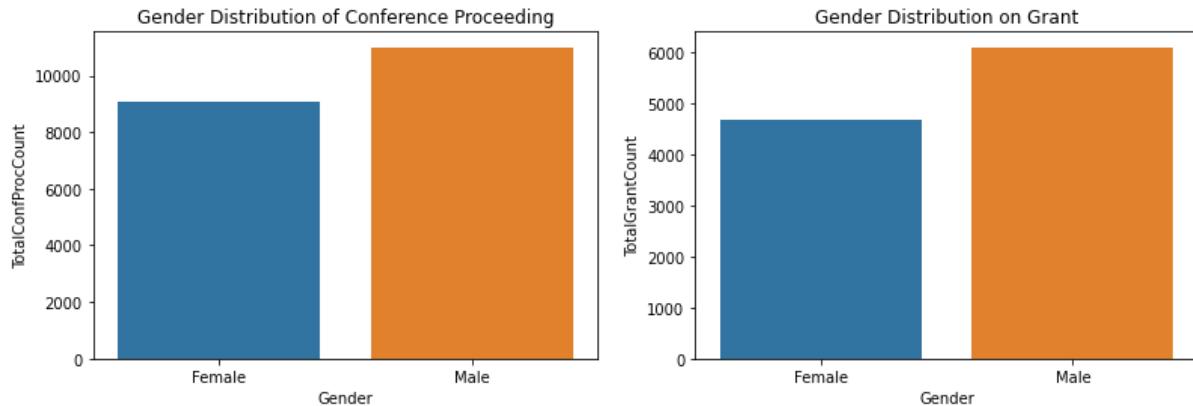
Male 5019



Article Distribution:

Female 42478

Male 49386



Conference Proceeding Distribution:

Female 9086

Male 10999

Grant Distribution:

Female 4676

Male 6102

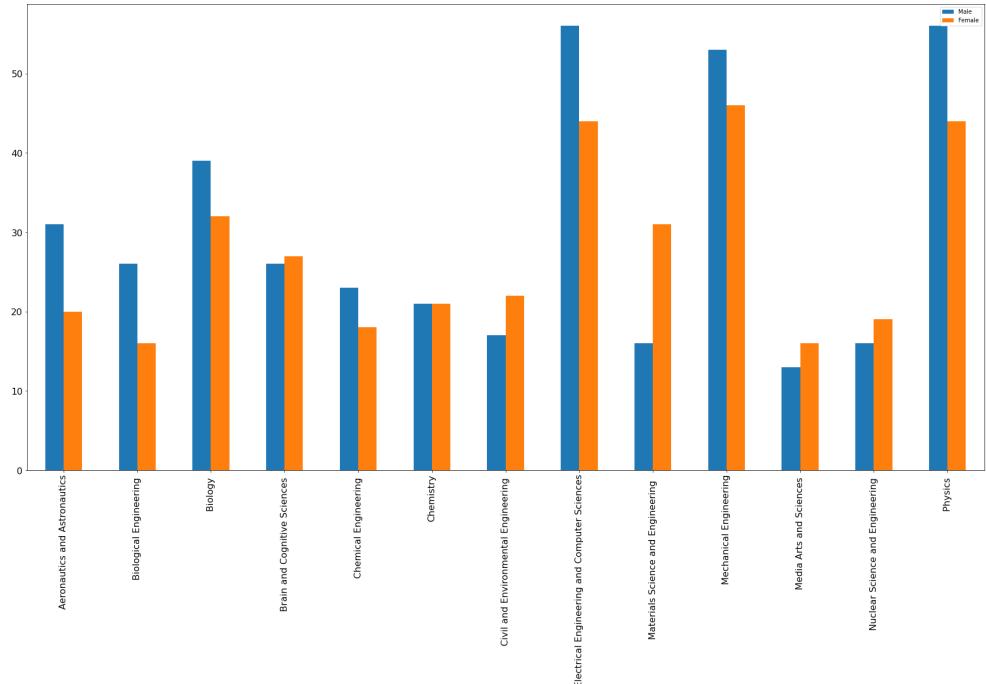


Figure: Gender distribution by gender and by the department.

Based on the graphical analysis based on the whole dataset by gender, by department, and by gender per department, we found out that there's an apparent comparison between the female and male results in all 4 categories, article, patent, grant, and conference proceeding, that we have looked at.

On an overall basis, the male has a comparatively larger number of people than females, but the difference in each department varies, as the graph right above shows.

### 3.2 Preliminary Analysis by Race

In general there are four types of race in the dataset:

"PCTAPI", 'PCTWHITE', 'PCTHISPANIC', 'PCTBLACK'.

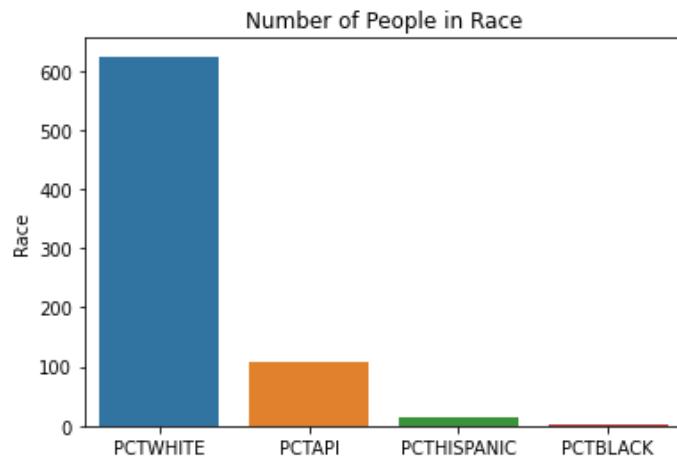
The quantity of them are listed:

PCTWHITE 624

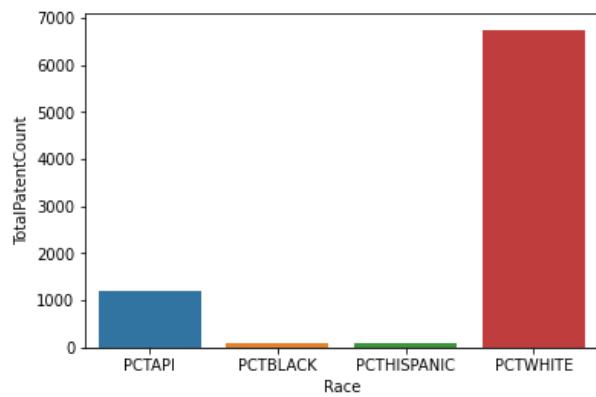
PCTAPI 107

PCTHISPANIC 15

PCTBLACK 3



Same as gender, we analyze the race associated with all the heads in the dataset.



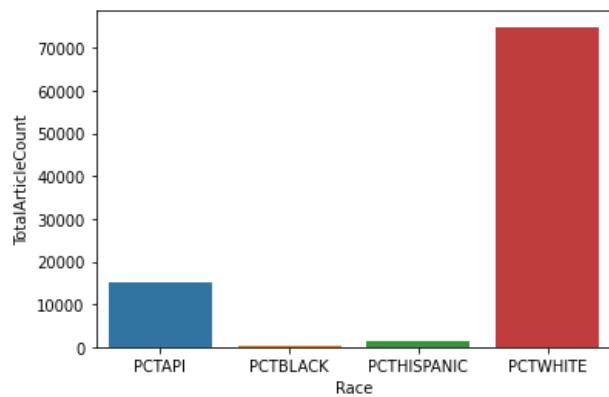
Patent by Race:

PCTAPI 1194

PCTBLACK 79

PCTHISPANIC 81

PCTWHITE 6751



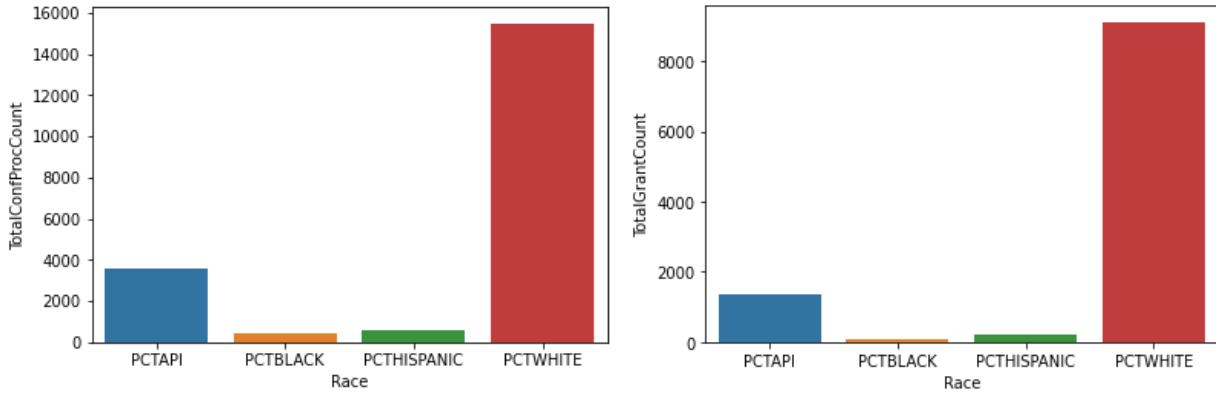
Article by Race:

PCTAPI 15239

PCTBLACK 216

PCTHISPANIC 1484

PCTWHITE 74925



#### Conference Proceeding by Race:

PCTAPI	3605
PCTBLACK	432
PCTHISPANIC	557
PCTWHITE	15491

#### Grant by Race:

PCTAPI	1374
PCTBLACK	66
PCTHISPANIC	217
PCTWHITE	9121

In general, PCTWHITE weighs the most in all aspects of race analysis. The sum of three other races is approximately one-third of the PCTWHITE.

## 4 Statistical Disparity Significance

To answer the first analysis question, we looked at the disparity between gender and among races respectively overall at MIT. We first computed the distribution to see if they have a similar shape, and then check the variance equality. If we obtain a result that the variances of two groups are equal, we then perform a hypothesis test using a t-test.

All the tests are performed at the significance level  $\alpha = 0.05$ .

### 4.1 Statistical Test - Disparity over Gender

The hypothesis tests have the following hypotheses:

$H_0$ : The means of the two groups are the same.

$H_1$ : The means of the two groups are significantly different.

#### 4.1.1 Article Count by Gender

Square Root Transformed Distribution of Male Faculty on Article Count

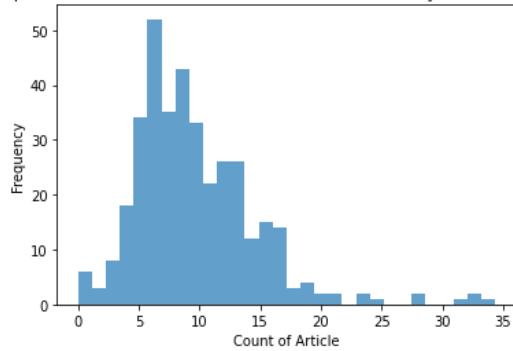


Figure 2: Histogram of Article Count of Male Faculty

Square Root Transformed Distribution of Female Faculty on Article Count

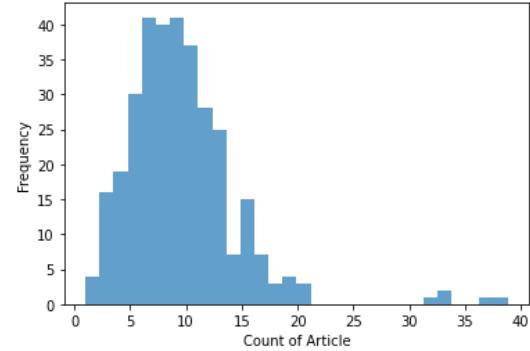


Figure 3: Histogram of Article Count of Female Faculty

The distributions show that the counts of articles for male and female faculties are in a similar shape. Since the distribution plot on raw data does not follow a normal distribution, we transform the data with square root and obtain normally distributed samples. Performing the Levene test on the square root transformed data, we got the result  $p - value = 0.43 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the t-test, we got the result  $p - value = 0.81 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing a significant disparity between male and female faculties in the total patent count.

#### 4.1.2 Conference Proceeding Count by Gender

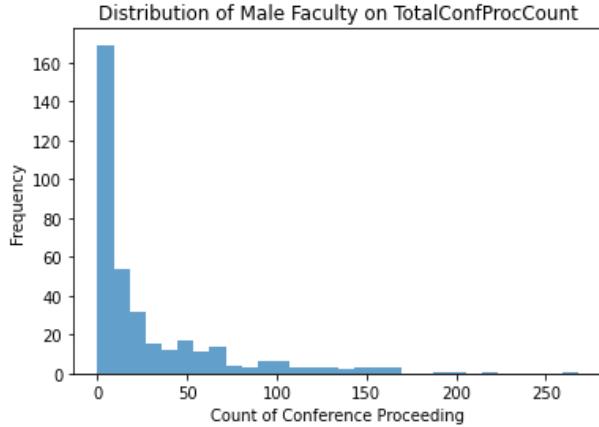


Figure 4: Histogram of Conference Count of Male Faculty

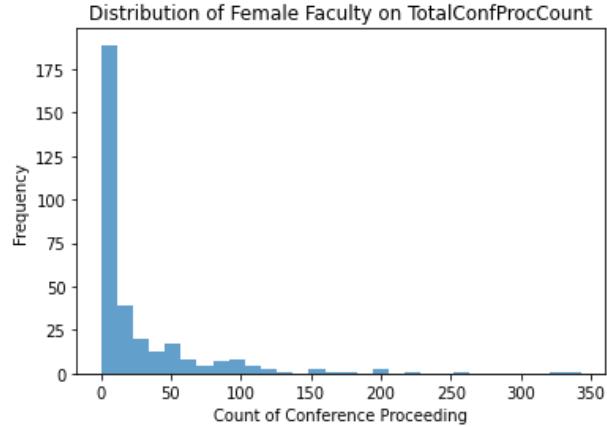


Figure 5: Histogram of Conference Count of Female Faculty

The distributions show that the counts of conference proceedings for male and female faculties are in a similar shape. Even though the distributions are not normal, we fail to transform them into a normal distribution. Performing the Levene test, we got the result  $p - value = 0.81 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the t-test, we got the result  $p - value = 0.62 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing a significant disparity between male and female faculties in the total conference proceeding count.

#### 4.1.3 Grant Count by Gender

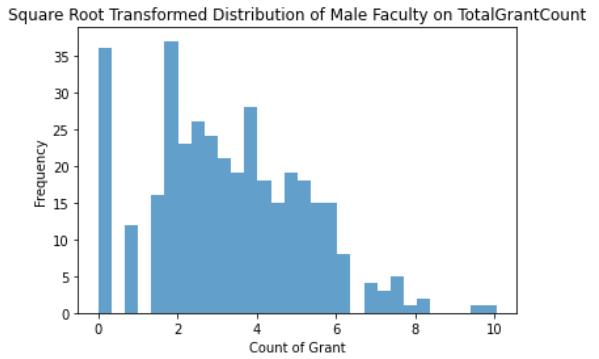


Figure 6: Histogram of Grant Count of Male Faculty

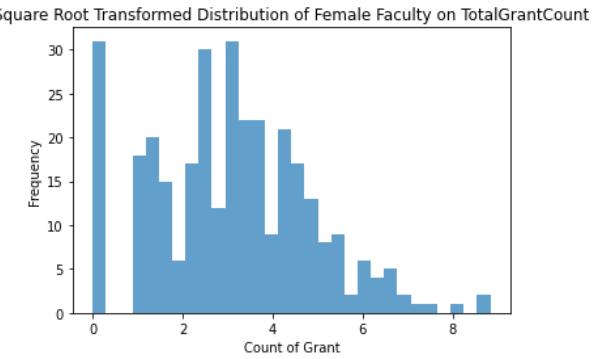


Figure 7: Histogram of Grant Count of Female Faculty

The distributions show that the counts of grants for male and female faculties are in a similar shape. Since the distribution plot on raw data does not follow a normal distribution, we transform the data with square root and obtain normally distributed samples. Performing the Levene test,

we got the result  $p - value = 0.066 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the t-test, we got the result  $p - value = 0.106 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing a significant disparity between male and female faculties in the total grant count.

#### 4.1.4 Patent Count by Gender

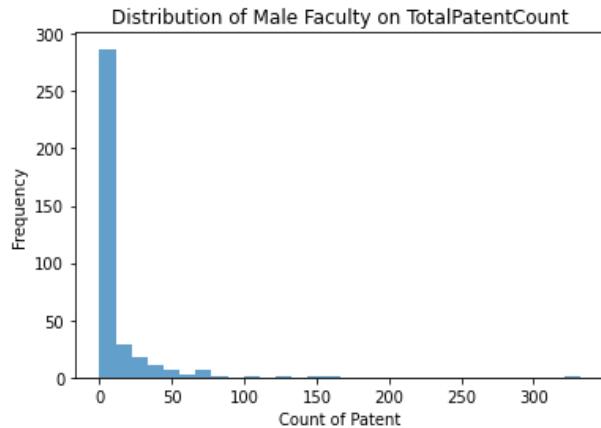


Figure 8: Histogram of Patent Count of Male Faculty

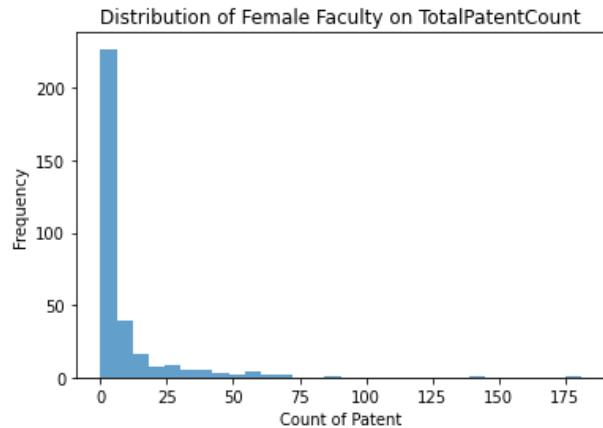


Figure 9: Histogram of Patent Count of Female Faculty

The distributions show that the counts for male and female faculties are in a similar shape. Even though the distributions are not normal, we fail to transform them into a normal distribution. Performing the Levene test, we got the result  $p - value = 0.18 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the t-test, we got the result  $p - value = 0.15 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing a significant disparity between male and female faculties in the total patent count.

## 4.2 Statistical Test - Disparity over Race

The hypothesis tests have the following hypotheses:

$H_0$ : The means of all sample groups are the same.

$H_1$ : At least two of the means in all sample groups are significantly different.

In this part of the test, we chose to use ANOVA F-test since there are more than 2 groups of data. Even though we utilized all races in the database, it turns out that only 4 races are matched and

predicted in our faculty dataset. Thus, for all tests below, there are 4 groups of data each representing a distinct race class.

#### 4.2.1 Article Count by Gender

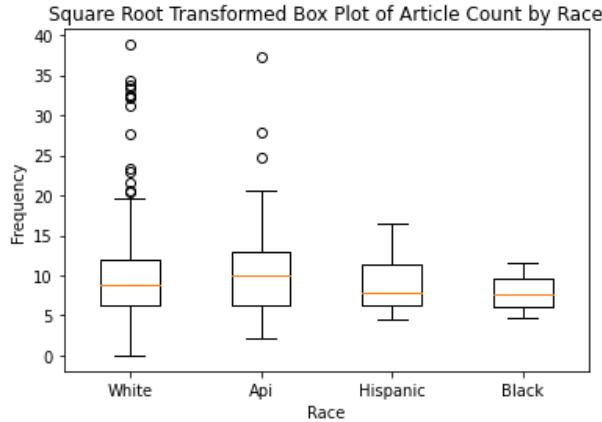


Figure: Box Plot of Article Count by Race.

As the box plot shown above, we can see that the range of all 4 races is quite similar in terms of boxes. However, we noticed there are plenty of “outliers” detected in the race group WHITE and API. This is the result of a large difference between the number of samples we have in each race. Performing the Levene test, we got  $p - value = 0.74 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the Analysis of Variance (one-way ANOVA), we got the  $p - value = 0.55 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing any two of the races have a significant disparity in the total patent count.

#### 4.2.2 Conference Proceeding Count by Gender

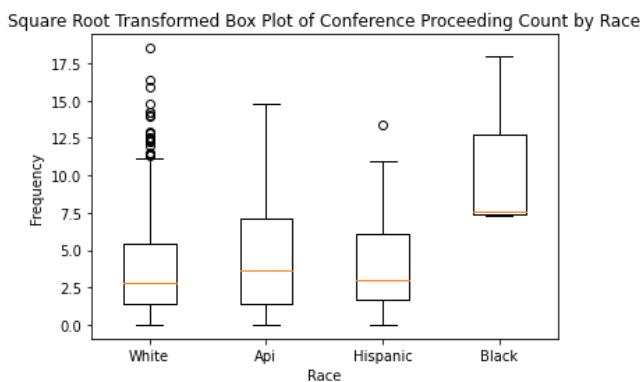


Figure: Box Plot of Conference Proceeding Count by Race.

As the box plot shown above, we can see that the range among WHITE, API and HISPANIC is quite similar in terms of boxes, while the BLACK class has a relatively higher median compared to others. It is obvious to see a lot of outliers detected in the race class of WHITE. This is the result of a relatively bigger sample compared to other races. Performing the Levene test, we got  $p - value = 0.33 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the Analysis of Variance (one-way ANOVA), we got the  $p - value = 0.00063 < \alpha$ . At  $\alpha = 0.05$  level of significance, we can reject the null hypothesis, that there is evidence showing at least two of the races have a significant disparity in the total patent count.

Multiple Comparison of Means - Tukey HSD, FWER=0.05						
group1	group2	meandiff	p-adj	lower	upper	reject
PCTAPI	PCTBLACK	6.3064	0.0113	1.0453	11.5674	True
PCTAPI	PCTHISPANIC	-0.0776	0.9	-2.5676	2.4124	False
PCTAPI	PCTWHITE	-0.8425	0.1236	-1.8274	0.1424	False
PCTBLACK	PCTHISPANIC	-6.384	0.0203	-12.0601	-0.7079	True
PCTBLACK	PCTWHITE	-7.1489	0.0024	-12.3439	-1.9539	True
PCTHISPANIC	PCTWHITE	-0.7649	0.8142	-3.1121	1.5823	False

Table: Pairwise Comparisons of Means in Conference Proceeding Counts.

The table above shows the pairwise comparisons between each race. We can see that there are significant disparities between BLACK and API, between BLACK and HISPANIC, and between BLACK and WHITE. And for the rest pairs, we fail to reject the null hypothesis, that no significant disparity exists.

#### 4.2.3 Grant Count by Gender

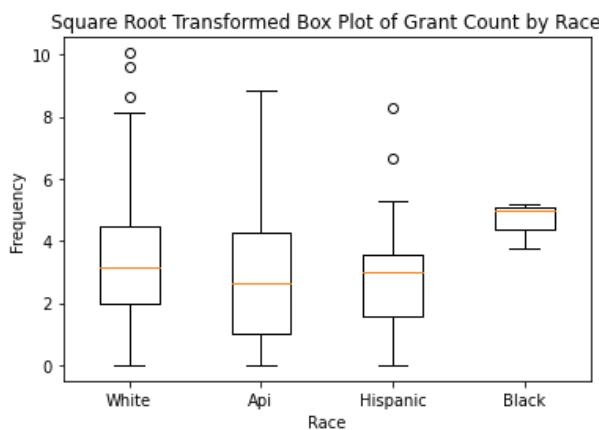


Figure: Box Plot of Grant Count by Race.

As the box plot shown above, we can see that the range of WHITE, API and HISPANIC is quite similar in terms of boxes, while BLACK is slightly higher than others in terms of the median. Performing the Levene test, we got  $p - value = 0.097 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the Analysis of Variance (one-way ANOVA), we got the  $p - value = 0.116 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing any two of the races have a significant disparity in the total patent count.

#### 4.2.4 Patent Count by Gender

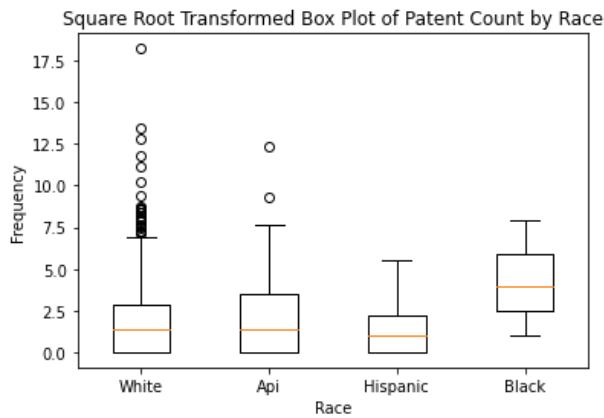


Figure: Box Plot of Patent Count by Race.

As the box plot shown above, we can see that the range of WHITE, API and HISPANIC is quite similar in terms of boxes, while BLACK is slightly higher than others in terms of the median. Performing the Levene test, we got  $p - value = 0.598 > \alpha$ . We can conclude that the two groups have equal variances.

Performing the Analysis of Variance (one-way ANOVA), we got the  $p - value = 0.28 > \alpha$ . At  $\alpha = 0.05$  level of significance, we fail to reject the null hypothesis, that there is no evidence showing any two of the races have a significant disparity in the total patent count.

### 4.3 Answer to First Analysis Question

To conclude the statistical tests performed above, we fail to reject the null hypothesis for all tests by gender. That is, according to our tests, we found that there is no statistically significant disparity between male and female faculties in all four categories we examined.

For the tests by race, we only successfully reject the null hypothesis for the conference proceeding count and fail to reject the rest three categories. That is, we found that there is no statistically significant disparity among races in articles, grants, and patents; there is a significant disparity for conference proceedings.

## 5 Disparity Over Time

In the previous sections, our team calculates, visualizes, and compares the number of patents, grants, articles, and conferences proceedings in different genders and races. Since female professors are less than male professors, the overall trend of the counts for male professors is more than those of female professors. And it is similar for different races. Even though we did not see a statistically significant disparity in genders and in races except for the conference proceedings by race, we still want to see an overall trend of the counts over time. In addition to the trend constructed using raw numbers, we put the disparity index into account this time in order to offset the influence of the raw data and reveal the fact that individual genders/ races contribute.

### 5.1 Disparity by Gender

In this part, we first clean and consolidate the data we have. We eliminate duplicates-- professors who are counted in more than one department but are the same professors. We also delete outliers in the “Year” parameter, which are “No Issue”, “1”, “2”, “Supp1”, and “First Serie.1”.

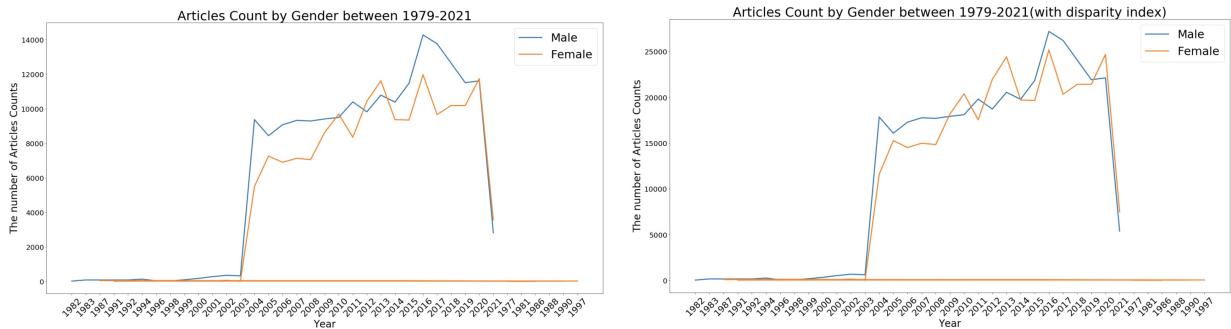


Figure: Articles Count for professors of different genders (without and with disparity index)

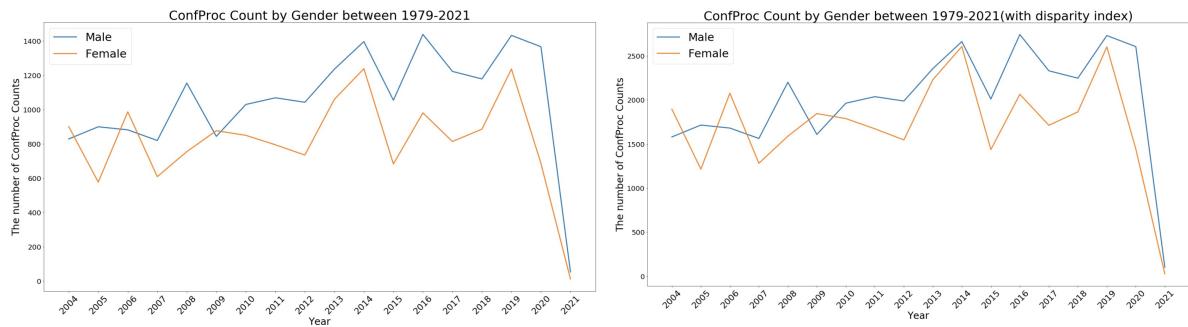


Figure: Conferences Proceedings Count for professors of different genders (without and with disparity index)

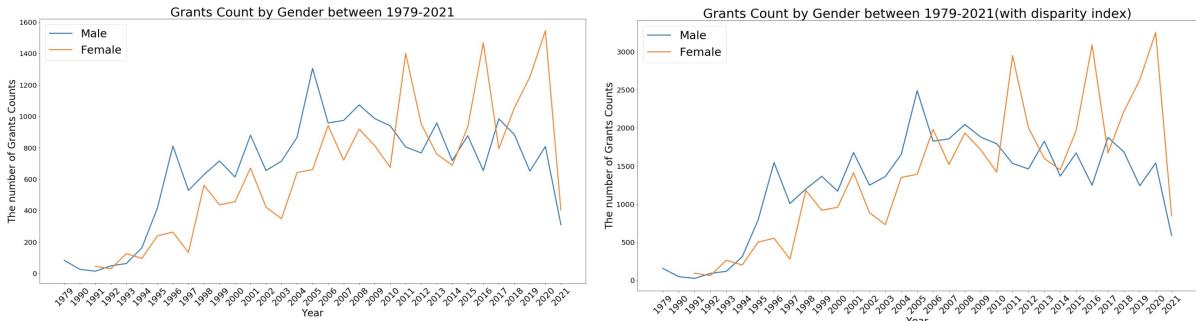


Figure: Grants Count for professors of different genders (without and with disparity index)

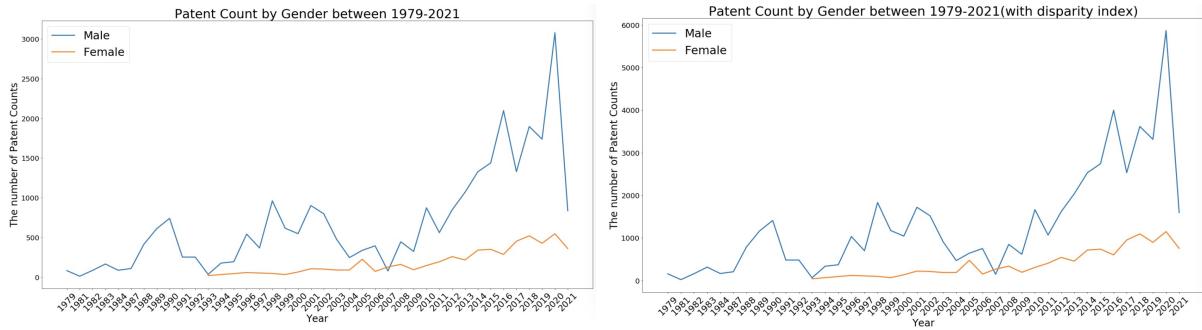


Figure: Patents Count for professors of different genders (without and with disparity index)

Put disparity index into account when we consider whether different genders can exert influence on the result. The overall trend does not change because of the disparity index, but we can see that the differences in the number of grants between male and female professors become smaller.

Before considering the disparity index, the picture shown indicates that male professors have more article counts, but after dividing the article count by each gender's ratio, we can see female professors don't do worse than female professors in terms of the article number they have. It helps reveal the individual contribution each professor has rather than the whole group contributions which could lead to too general conclusions and bias.

One of the reasons the disparity index doesn't change the overall shape/trend of different parameters is that the number of female and male professors is more or less half and half, so dividing the raw number we have in deliverable 2 by about 0.5 doesn't change the plotted trend. Will the change be the same for different races when taking the disparity index into account? Our initial guess is that the situation will become different as PCTWHITE has a large percentage compared to the super low percentage of PCTBLACK. Let's see the actual trending when we consider the disparity index.

## 5.2 Disparity by Race

Similarly to the disparity by gender, we also remove the duplicates, and remove the invalid years in the dataset, and construct graphs in raw numbers and with disparity index for each category.

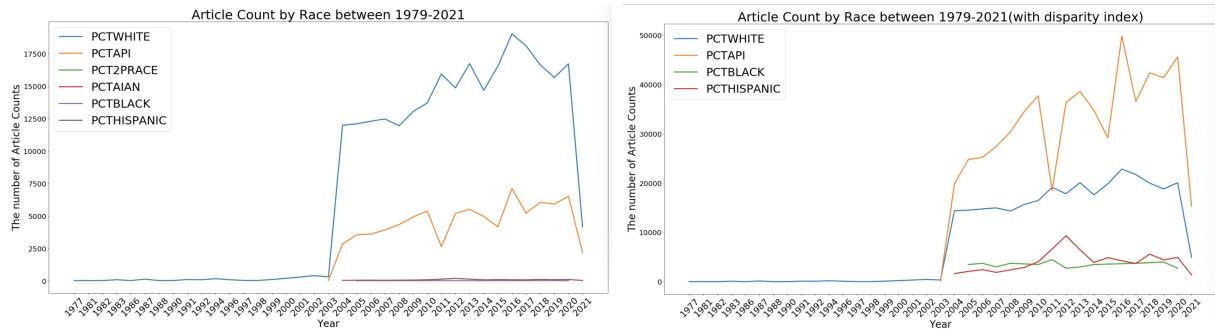


Figure: Articles Count for professors of different races(without and with disparity index)

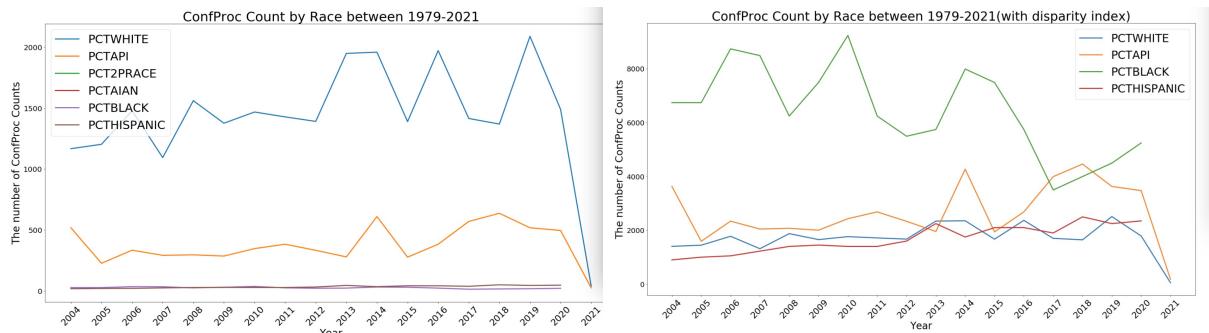


Figure: Conferences Proceeding Count for professors of different races(without and with disparity index)

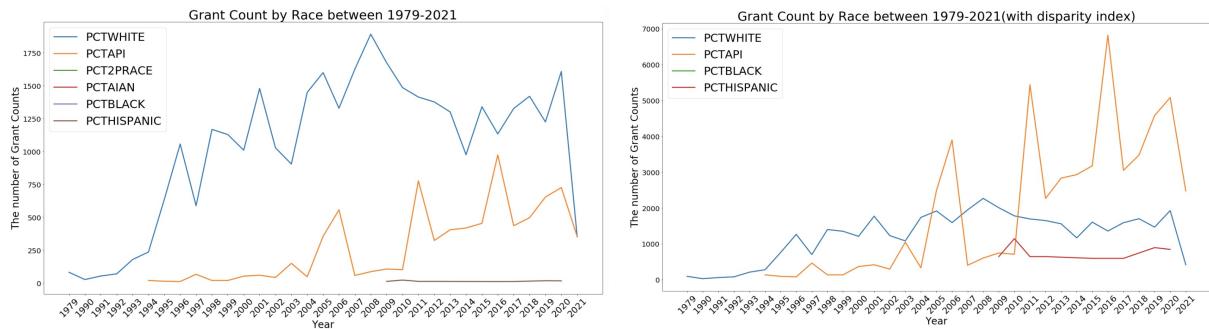


Figure: Grants Count for professors of different races(without and with disparity index)

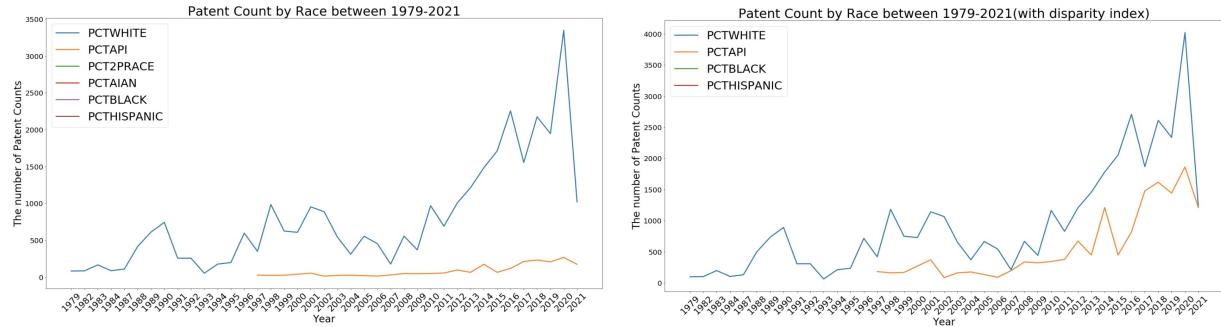


Figure: Patents Count for professors of different races(without and with disparity index)

After using the disparity index, we can see more vivid trending of the races that have very few professors. Especially PCTAPI professors and PCTBLACK professors, their contributions are shrunk when we use raw data to do data visualization because of their small population bases. In fact, their individual contributions are even larger than PCTWHITE professors. We can see this by comparing the plots of with and without disparity index, and also by comparing the parameter counts of professors of different races from the “with disparity index” plot.

## 6 Gap Between Newer and Older Faculties

### 6.1 Disparity on Patent

In deliverable 2, we can conclude that the disparity of Patent Count by Gender between 1983-2021 is increasing. And the disparity of Article Count by Gender between 2001-2021 is decreasing. We now use the titles of professors to divide whether a professor is new or old, e.g. Professor, Associate professor, Assistant professor, etc. It takes more than 8 years to become a professor in general. Thus, we have 428 old professors and 321 new professors using this classification way.

Data of total Rank:

Professor	428	Associate	115	Assistant	83	Other	123
-----------	-----	-----------	-----	-----------	----	-------	-----

