EDA of Gender & Diversity in Aerospace

Ashlyn Montgomery, Kameron Galm, Nam Jun Lee, Julian Rangel

2023-03-07

## Dataset Information

There are a total of five datasets to be used in our project, and we will perform exploratory data analysis to understand these datasets.  
Data 1 to 3 are CSV files listing values obtained by extracting specific words from pdf using “PdfExtract.R”.

Data 1: Average earnings by race and gender.  
Data 2: Ratio by gender in the aviation industry.  
Data 3: Gender leadership position ratio in the aviation industry.  
Data 4: Salary by gender in the aerospace industry.  
Data 5: pcgFull.

## Data Preprocessing

### Data 1

## 'data.frame': 8 obs. of 3 variables:  
## $ Race : chr "Asian" "White" "Asian" "Hispanic" ...  
## $ AvgEarning: chr "$103,300" "$90,600" "$88,600" "$73,000" ...  
## $ Gender : chr "men" "men" "women" "men" ...

After checking the structure of the first dataset, we know it consists of 3 columns and 8 rows.

## Race AvgEarning Gender   
## Asian :2 Min. : 57000 men :4   
## Black :2 1st Qu.: 63900 women:4   
## Hispanic:2 Median : 71100   
## White :2 Mean : 75612   
## 3rd Qu.: 89100   
## Max. :103300

## [1] "Total NA: 0"

In the case of “AvgEarning” variable, special characters are included, so they are removed and converted into numerical variable, and the attributes of each variable are converted into factorial.  
And as a result of checking the descriptive statistics, it can be seen that there are no missing values and the data are well organized.

### Data 2

## 'data.frame': 18 obs. of 3 variables:  
## $ Gender : chr "Women" "Women" "Women" "Women" ...  
## $ Ratio : num 5.6 5.4 7.2 2.6 11.6 16.8 19.7 79.7 16.7 94.4 ...  
## $ Occupation: chr "Aircraft Pilots and Flight Engineers" "Aircraft Mechanics and Service Technicians" "Remote Pilots" "MAINTENANCE TECHNICIAN" ...

After checking the structure of the second dataset, we know it consists of 3 columns and 18 rows.

## Gender Ratio Occupation  
## Men :9 Min. : 2.60 Aerospace Engineers :2   
## Women:9 1st Qu.:12.88 Air Traffic Controllers :2   
## Median :50.00 Aircraft Mechanics and Service Technicians:2   
## Mean :50.00 Aircraft Pilots and Flight Engineers :2   
## 3rd Qu.:87.12 Airport Managers :2   
## Max. :97.40 Dispatchers :2   
## (Other) :6

## [1] "Total NA: 0"

As a result of checking the descriptive statistics after converting the properties of each variable into factors, it can be seen that there are no missing values and the data are well organized.

### Data 3

## 'data.frame': 12 obs. of 3 variables:  
## $ Gender : chr "Women" "Women" "Women" "Women" ...  
## $ Ratio : chr "3%" "15%" "4%" "13%" ...  
## $ LeadershipPosition: chr "CEO" "CFO" "COO" "CCO" ...

As a result of confirming the structure of the third dataset, it was confirmed that it consisted of 3 columns and 12 rows.

## Gender Ratio LeadershipPosition  
## Men :6 Min. : 3 CCO :2   
## Women:6 1st Qu.:13 CEO :2   
## Median :50 CFO :2   
## Mean :50 CIO :2   
## 3rd Qu.:87 COO :2   
## Max. :97 HR Director:2

## [1] "Total NA: 0"

In the case of “Ratio” variable, special characters are included, so they are removed and converted into numerical variable, and the attributes of each variable are converted into factorial.  
And as a result of checking the descriptive statistics, it can be seen that there are no missing values and the data are well organized.

### Data 4

## 'data.frame': 46 obs. of 17 variables:  
## $ ID.Year : int 2020 2019 2018 2017 2016 2015 2014 2020 2019 2018 ...  
## $ Year : int 2020 2019 2018 2017 2016 2015 2014 2020 2019 2018 ...  
## $ ID.Workforce.Status : chr "true" "true" "true" "true" ...  
## $ Workforce.Status : chr "true" "true" "true" "true" ...  
## $ Record.Count : int 62 69 111 172 269 336 411 502 576 855 ...  
## $ Average.Wage : num 87142 97561 91652 91284 90340 ...  
## $ Average.Wage.Appx.MOE : num 27384 27400 23248 19538 15756 ...  
## $ Total.Population : int 1212 1329 2143 3300 5235 6591 7614 9686 11175 16593 ...  
## $ Total.Population.MOE.Appx: num 855 895 1137 1410 1776 ...  
## $ PUMS.Occupation : chr "Aerospace engineers" "Aerospace engineers" "Aerospace engineers" "Aerospace engineers" ...  
## $ ID.PUMS.Occupation : chr "172011" "172011" "172011" "172011" ...  
## $ Slug.PUMS.Occupation : chr "aerospace-engineers" "aerospace-engineers" "aerospace-engineers" "aerospace-engineers" ...  
## $ PUMS.Industry : chr "Aerospace products & parts manufacturing" "Aerospace products & parts manufacturing" "Aerospace products & parts manufacturing" "Aerospace products & parts manufacturing" ...  
## $ ID.PUMS.Industry : chr "33641M2" "33641M2" "33641M2" "33641M2" ...  
## $ Slug.PUMS.Industry : chr "aerospace-products-parts-manufacturing" "aerospace-products-parts-manufacturing" "aerospace-products-parts-manufacturing" "aerospace-products-parts-manufacturing" ...  
## $ Sex : chr "Female" "Female" "Female" "Female" ...  
## $ ID.Sex : int 2 2 2 2 2 2 2 1 1 1 ...

As a result of confirming the structure of the third dataset, it was confirmed that it consisted of 17 columns and 46 rows. There are some columns that are not to be used for our project, so it needs to be removed.

## [1] 46 6

## Year Record\_Count Average\_Wage Total\_Population  
## Min. :2014 Min. : 42.00 Min. : 38811 Min. : 822   
## 1st Qu.:2016 1st Qu.: 79.25 1st Qu.: 51806 1st Qu.: 1513   
## Median :2018 Median : 163.00 Median : 91151 Median : 3236   
## Mean :2018 Mean : 368.96 Mean : 85774 Mean : 7176   
## 3rd Qu.:2019 3rd Qu.: 327.25 3rd Qu.:111437 3rd Qu.: 6298   
## Max. :2020 Max. :2924.00 Max. :168674 Max. :55838   
## PUMS\_Occupation Sex   
## Aerospace engineers :14 Female:23   
## Inspectors, testers, sorters, samplers, & weighers:14 Male :23   
## Other assemblers and fabricators : 6   
## Other managers : 6   
## Software developers : 6   
##

## [1] "Total NA: 0"

After deleting unnecessary columns, changing the names of each column, and converting attributes of gender and occupational variables into factors. And checking technical statistics shows that there are no missing values and the data are well organized.  
Also, it can be seen that the finally organized dataset consists of 6 columns and 46 rows.

### Data 5

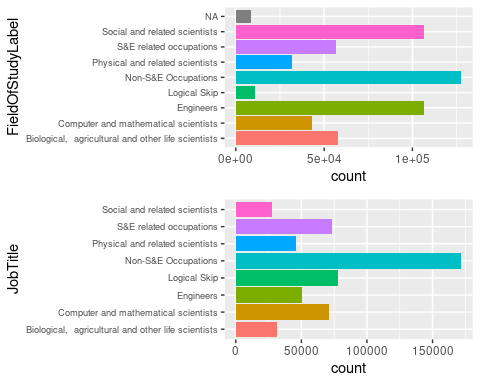
## 'data.frame': 549398 obs. of 11 variables:  
## $ AGE : int 36 38 42 48 37 52 34 35 61 47 ...  
## $ GENDER : chr "M" "F" "M" "F" ...  
## $ Field.of.Study : int 1 4 4 2 4 4 4 4 2 7 ...  
## $ Field.of.Study.Label: chr "Computer and mathematical scientists" "Social and related scientists" "Social and related scientists" "Biological, agricultural and other life scientists" ...  
## $ SALARY : num 50000 24000 200000 38147 70000 ...  
## $ Salary.Group : chr ">=100k, <=200k" "<50k" ">=50k, <100k" "<50k" ...  
## $ Job.Satisfaction : chr "2" "1" "2" "3" ...  
## $ Place.of.Birth : chr "Y" "Y" "Y" "Y" ...  
## $ Job.code : int 1 3 3 1 3 3 3 1 2 3 ...  
## $ Job.Title : chr "Computer and mathematical scientists" "Physical and related scientists" "Physical and related scientists" "Computer and mathematical scientists" ...  
## $ Year : int 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...

After checking the structure of the first dataset, we know it consists of 11 columns and 549398 rows. and checking the data properties, almost all data is made into characters, so it seems necessary to convert it.

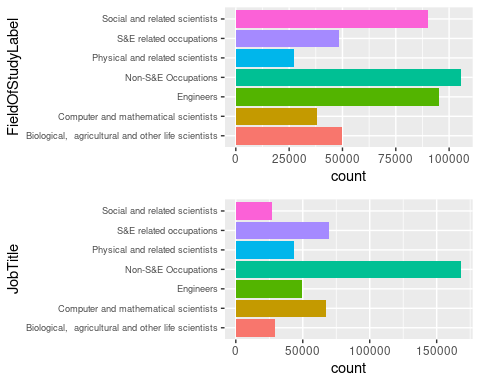
## AGE GENDER FieldOfStudy FieldOfStudyLabel   
## 0 0 8762 8762   
## SALARY SalaryGroup JobSatisfaction JobCode   
## 0 14282 0 0   
## JobTitle Year   
## 0 0

## AGE GENDER FieldOfStudy   
## Min. :16.00 F:249040 Min. :1.000   
## 1st Qu.:33.00 M:300358 1st Qu.:4.000   
## Median :44.00 Median :5.000   
## Mean :45.29 Mean :4.684   
## 3rd Qu.:56.00 3rd Qu.:7.000   
## Max. :76.00 Max. :8.000   
## NA's :8762   
## FieldOfStudyLabel SALARY   
## Non-S&E Occupations :127731 Min. : 0   
## Social and related scientists :106571 1st Qu.: 50000   
## Engineers :106441 Median : 80000   
## Biological, agricultural and other life scientists: 57581 Mean :1462535   
## S&E related occupations : 56922 3rd Qu.: 140000   
## (Other) : 85390 Max. :9999998   
## NA's : 8762   
## SalaryGroup JobSatisfaction JobCode   
## <50k :134748 1:204370 Min. :1.000   
## >=100k, <=200k:199150 2:205678 1st Qu.:3.000   
## >=50k, <100k : 95594 3: 36804 Median :6.000   
## >200k :105624 4: 10266 Mean :5.281   
## NA's : 14282 L: 92280 3rd Qu.:7.000   
## Max. :8.000   
##   
## JobTitle Year   
## Non-S&E Occupations :172024 Min. :2003   
## Logical Skip : 77860 1st Qu.:2010   
## S&E related occupations : 73518 Median :2013   
## Computer and mathematical scientists: 71053 Mean :2013   
## Engineers : 50368 3rd Qu.:2017   
## Physical and related scientists : 45554 Max. :2019   
## (Other) : 59021

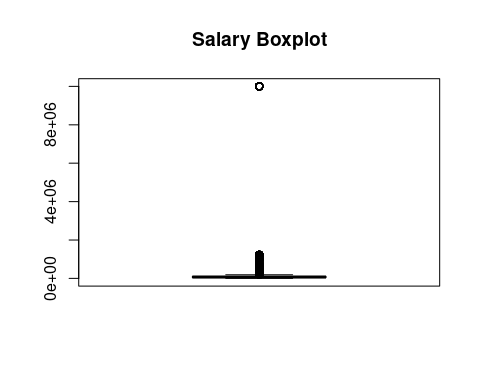
First, the “Place.of.Birth” variable is an unnecessary column, so it was deleted, and the data attributes of each variable were first converted to properly check the data. And as a result of checking the sum of the missing values in each column, it can be seen that there are missing values in the “FieldOfStudy”, “FieldOfStudyLabel”, and “SalaryGroup” variables.  
To confirm more specifically, we checked the descriptive statistics and found that outliers exist when looking at the minimum and maximum values of the “SALARY” variables and that the “JobTitle” variable also has characters that replace missing values.



As a result of checking the FieldOfStudyLabel and JobTitle bar plots, it was confirmed that the FieldOfStudyLabel has a NA element, and the two variables contain the same element “Logical Skip”. Through this, it was determined that this factor was not answered by the survey respondents, so it was decided to remove it because it was the same as the missing value.



As a result of re-checking the graph to see if the missing values were well removed, it can be seen that they were well processed.



As a result of checking the “SALARY” variable through the boxplot, it can be confirmed that there are very many outliers.

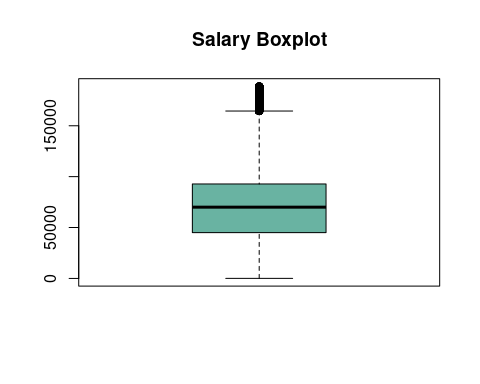
## [1] "Minimum: -41232.750000, Maximum: 188721.250000"

## [1] "Total NA: 32814"

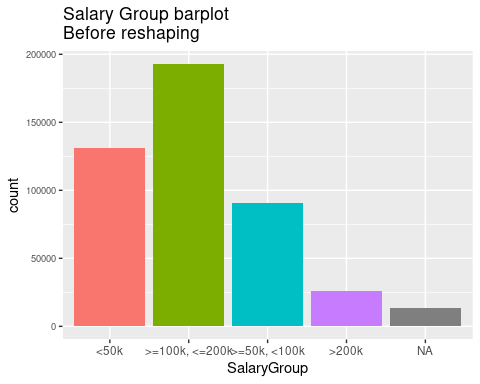
In order to identify the outliers, the interquartile range was obtained using the formula:

or

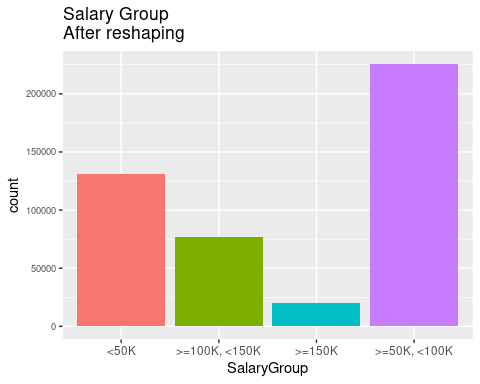
Therefore, interquartile range is minimum: -41232.750000, Maximum: 188721.250000. If you look at the minimum value, it is a negative number, but salary can’t be negative. So replaced with null values outside the interquartile range (0 < Salary < 188721.25). And when we checked the number of outliers, it turned out to be 32814.



It was determined that it was better to replace outliers with an average value than to remove them. So after replacing it with an average value we checked the outliers with a box plot, and as a result, there are still a few outliers existing but we decided to ignore them. Because in some cases where the salary exceeds $200,000.



As a result of checking the “SalaryGroup”, it can be seen that there are five levels, and among them, there are NA levels. There was a change in the “SALARY” variable while removing outliers, so we decided to relocate the values.



After relocating, it can be seen from the bar graph that the missing values were well removed and the values were well divided into four levels.

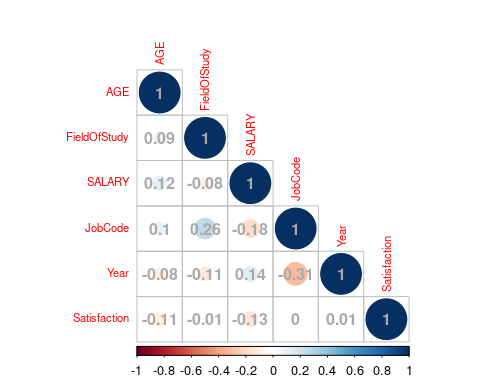
## AGE GENDER FieldOfStudy FieldOfStudyLabel   
## 0 0 0 0   
## SALARY SalaryGroup JobSatisfaction JobCode   
## 0 0 0 0   
## JobTitle Year StemStudy StemJob   
## 0 0 0 0

## AGE GENDER FieldOfStudy   
## Min. :16.00 F:200882 Min. :1.000   
## 1st Qu.:32.00 M:253593 1st Qu.:3.000   
## Median :42.00 Median :5.000   
## Mean :43.56 Mean :4.592   
## 3rd Qu.:54.00 3rd Qu.:6.000   
## Max. :76.00 Max. :7.000   
##   
## FieldOfStudyLabel SALARY   
## Non-S&E Occupations :105767 Min. : 1   
## Engineers : 95417 1st Qu.: 45000   
## Social and related scientists : 90142 Median : 70000   
## Biological, agricultural and other life scientists: 49618 Mean : 71102   
## S&E related occupations : 48326 3rd Qu.: 92800   
## Computer and mathematical scientists : 37975 Max. :188700   
## (Other) : 27230   
## SalaryGroup JobSatisfaction JobCode   
## <50K :131274 1:197261 Min. :1.000   
## >=100K, <150K: 77019 2:197994 1st Qu.:3.000   
## >=150K : 20330 3: 35585 Median :6.000   
## >=50K, <100K :225852 4: 9963 Mean :4.859   
## L: 13672 3rd Qu.:7.000   
## Max. :7.000   
##   
## JobTitle Year   
## Non-S&E Occupations :168485 Min. :2003   
## S&E related occupations : 69925 1st Qu.:2010   
## Computer and mathematical scientists : 67482 Median :2015   
## Engineers : 49223 Mean :2013   
## Physical and related scientists : 43701 3rd Qu.:2017   
## Biological, agricultural and other life scientists: 28987 Max. :2019   
## (Other) : 26672   
## StemStudy StemJob   
## Non-Stem:105767 Non-Stem:168485   
## Stem :348708 Stem :285990   
##   
##   
##   
##   
##

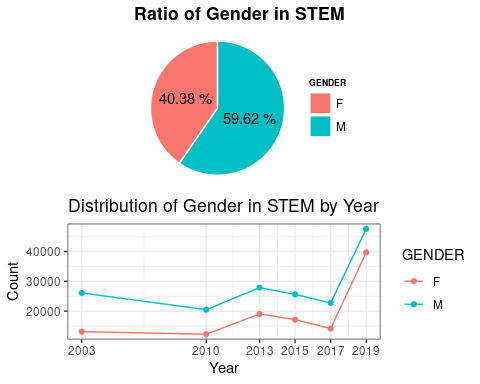
## [1] 454475 12

The “FieldOfStudy” and “JobCode” variables are composed of seven levels, through which we added derivatives by dividing them into two levels as to whether they were STEM or not.  
And finally, as a result of checking whether the data is well organized, it can be confirmed that there are no missing values in each column and all variables are well organized. Also, it can be seen that the finally organized dataset consists of 12 columns and 454475 rows.

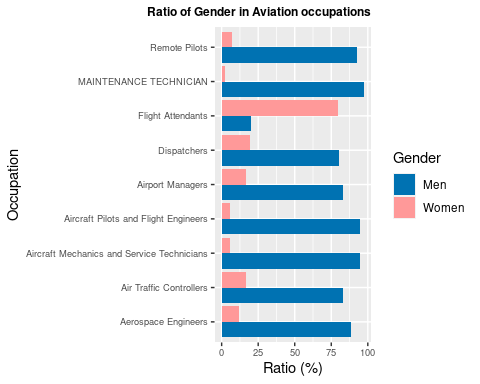
## Visualization



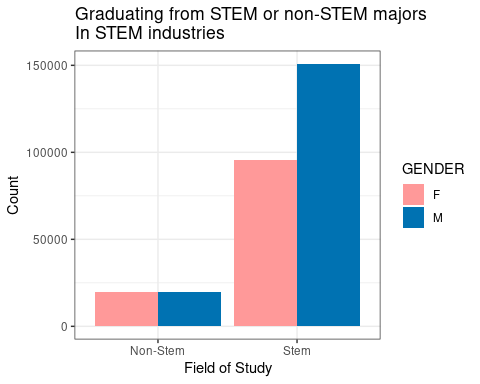
Using numerical variables to determine the correlation between each variable, it was found that “FieldOfStudy” and “JobCode” had a slight positive correlation, and the rest of the variables were not related.



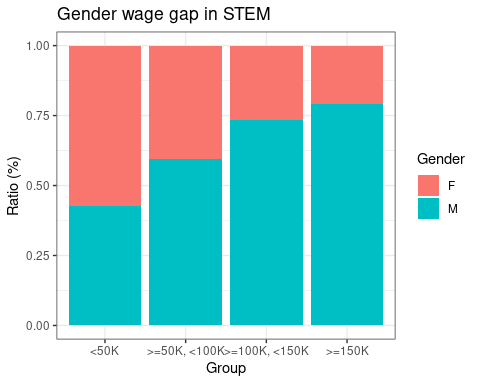
As can be seen from the above graph, in the case of STEM fields, the gender ratio is slightly more distributed for men, and as a result of checking the annual distribution, it can be seen that the number of employees in STEM fields has increased since 2017.  
But noticeably, the proportion of women is better than that of men.



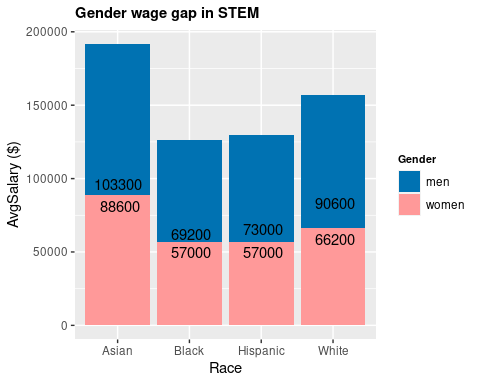
Therefore, as a result of checking the ratio of women and men in the aviation industry to further subdivide it, it can be confirmed that the gender ratio of the aviation industry is far more numerous to men except for flight attendants.



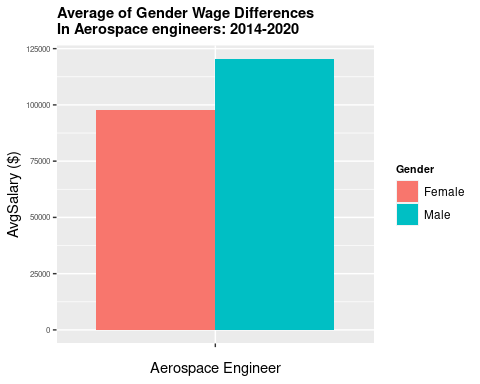
And as a result of identifying the distribution of STEM employees according to the graduation rate in related fields, the proportion of men and women who graduated from the STEM field was not equally large, and in the case of employees who graduated from the STEM field, men were more engaged in the STEM industry than women.



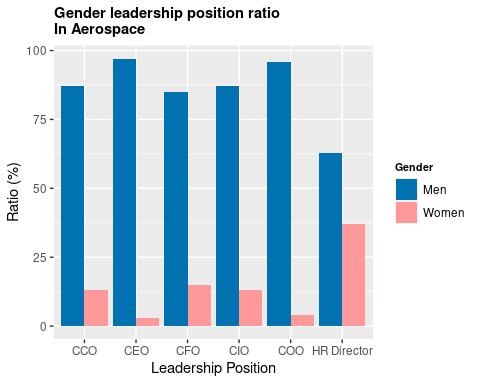
In addition, visualizing gender pay by the salary group confirmed that many women were paid less than 50,000 and that most of the employees who were paid more than 50,000 were male.



As a result of looking at the above graph, we can see that Asians are paid the most, and Black is paid the least. Also, when checking all the average salaries by each race, it can be seen that women are paid less than men.



To further refine it, we checked the average annual salary of a gender aerospace engineer from 2014 to 2020, and found that female employees were paid less than 100,000, and men were paid more than 120,000.



Finally, as a result of checking the gender ratio in the leadership position by gender, it can be seen that men are overwhelmingly more distributed than women.

## Conclusion

Through this analysis, STEM still prefers men to women, and even if they graduate from the same field, it can be judged that there is gender discrimination in STEM (add a little more specifically, aerospace) because there is a wage gap between gender and race.