Kameron Galm

Nam Jun Lee

Ashlyn Montgomery

Julian Rangel

DATA 424

Professor Rhonda Crate

22 April 2023

Gender and Diversity in the Aerospace Industry

# Abstract

Many studies have been conducted on the wage gaps between men and women, along with between people of varying races, but few have been sourced from technological industries. In this study, we analyzed several reports from the aerospace industry regarding diversity and inclusion goals across several years, as well as surveys that were filled out anonymously. By manipulating the data by extracting gendered text, it became possible to determine important characteristics, such as job position and pay, by gender. When possible, race was also considered alongside gender identity. Several different programs were used to enable this, including both Microsoft Excel and R-Studio. It was then determined that there is a notable gap in both wages and higher-level jobs within companies between men and women, up to approximately $20,000 across all positions. Because of this, it can be safely assumed that there is still a slight stigma in technology-heavy industries towards diversity, despite recent trends indicating a strong increase in the number of women and minorities hired.

**Contents**

[Abstract 1](#_Toc133173297)

[Introduction 4](#_Toc133173298)

[Literature Review 5](#_Toc133173299)

[Methods 6](#_Toc133173300)

[Results 12](#_Toc133173301)

[Conclusion 20](#_Toc133173302)

[Works Cited 21](#_Toc133173303)

# Introduction

Workplace diversity refers to equal representations of different genders and races. Unfortunately, the aerospace industry has long lacked this quality. There are numerous benefits to promoting gender and racial diversity in the workplace, some of which include garnering new perspectives, filling gaps in a growing industry with technically skilled workers, and further growing the economy. When only a small subset of all genders and ethnicities are represented in a scientific field, it impacts marginalized communities. This is due to it ignoring the differences in gender, such as the different healthcare needs of women, non-binary people, and racial and ethnic minorities. Because of this, research was done regarding the issue of gender and diversity in the aerospace industry through various analytical methods and analyses.

To achieve this, various business questions that fall under the umbrella of diversity in the aerospace industry have been pinpointed. They are as follows: What is the ratio of women to men in aerospace? Is there a difference between men and women graduating from STEM fields in aerospace? Is there a difference in how men and women are treated in aerospace, including by race? Does workplace productivity change based on the ratio or treatment of women within the workplace? Additionally, work has been done to predict and project the number of female employees in the aerospace industry in the next ten years. Given this, our goal is to identify patterns and differences between men and women in the aerospace sector so that overall STEM-related workplaces can hire employees equally without gender and race biases prior to recruitment. This is important because achieving this goal and answering these business questions will help create an inclusive environment in STEM-related workplaces and a workplace culture in which individuals are respected equally.

# Literature Review

When we began analyzing differences in how women and men are treated in aerospace, we reviewed existing scholarly articles that have been written on the subject. We found that in the current STEM field of society, treatment by gender and race is significantly different. According to a report, "STEM workers often earn more than other workers, but the average STEM worker has a significant wage gap by gender, race, and ethnicity" (Fry, R et al., 2021). We hope that our analysis will help uncover the severity of this gap. Our next step was to try to understand workplace productivity variations related to the ratio of women within the workplace. In most cases, STEM-field companies have shown poor gender diversity, especially in the more technical occupations. This lack of diversity affects the profitability of these companies, and it was discovered that "the most gender-diverse companies are 21% more likely to experience above-average profitability" (Meta). Therefore, our productivity analysis will look at profitability as a proxy and explore profitability data to try and find a correlation with diversity and profitability.

Currently, efforts are being made in the STEM field throughout the United States regarding recruitment. “[W]omen’s representation in our workforce increased to 23.2% in the United States and 24.6% internationally, both because of hiring efforts and stronger retention” (Boeing.). As such, companies in many regions are currently striving to correct the gender ratio balance in hiring. We will expand on this existing research and show, through an ARIMA model, the number of female employees projected to be in the STEM field over the next 10 years. This will help companies in STEM plan and prepare in advance to fill the recruitment quota and provide equal opportunities for gender diversity in recruitment.

# Methods

For this project, six datasets were used, and while each dataset is unrelated to the other, they are used to find answers to specific questions related to the problem statement and goals. Several .CSV files were extracted from .PDF documents using the R “pdftools” package, the first of which deals with data on the average wage gap in STEM fields by gender and race. Similarly, the second file deals with data on leadership positions by gender in the aerospace industry and the third .CSV file deals with data on wages by gender in the aerospace industry. Lastly from this group of files was the fourth .CSV, which deals with data on the gender distribution of pilots by year. The fifth .CSV, which is unrelated to the priorly mentioned ones, was collected from the DATAUSA webpage, and deals with data on gender distribution in the aerospace industry by year. It consists of a total of 46 rows and 17 columns. After the preprocessing step of removing unnecessary columns, it is only composed of 46 rows and 6 columns. The sixth and final .CSV we collected is a survey dataset for STEM field workers or non-STEM field workers, consisting of a total of 549,398 rows and 11 columns.

After examining the descriptive statistics of each dataset, it was discovered that there are several variables that require pre-processing in the survey dataset. Upon observing the minimum and maximum values of the "SALARY" variable, it was found that there are outliers present. Additionally, missing values were found in the "JobTitle", "SalaryGroup", and "FieldOfStudyLabel" variables (Figure 1).

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

Figure 1: Descriptive statistics of survey dataset

Through the boxplot of the "SALARY" variable, it can be observed that there are many outliers. This suggests that the dataset contains a significant amount of abnormal or extreme values, which may affect the overall distribution and analysis of the data (Figure 2).

차트이(가) 표시된 사진

자동 생성된 설명

Figure 2: Check outliers present in salary variables

To identify the outliers, the decision range was obtained using the formula (Equation 1).

Equation 1: Decision range equation (Chaudhary)

Any data point less than the lower bound or more than the upper bound is considered as an outlier (Chaudhary). After applying the designated formula, the resulting values for the lower and upper bounds are -41232.75 and 188721.25, respectively, which are indicative of any potential outliers in the data. However, since a negative salary value is not possible, such values have been replaced with null values falling outside the acceptable range (0 < Salary < 188721.25). Upon identifying the outliers, it has been observed that there are a total of 32814 instances that fall outside the interquartile range. It was determined that replacing outliers with an average value was better than removing 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 (Figure 3).

차트, 도표이(가) 표시된 사진

자동 생성된 설명

Figure 3: Salary variables after replacing outliers with average value

Through exploratory analysis, a pre-processed dataset was obtained and an ARIMA model approach will be used to predict the distribution of female pilots for the next few years. ARIMA is a statistical model used for time series data prediction, where the independent variables are lagged values of the dependent variable and/or lagged errors of the forecasted variable. The equation for the ARIMA model is given by the following equation (Equation 2).

Equation 2: ARIMA model equation (Shweta)

This equation has three terms. AR stands for autoregressive analysis, where the time series is regressed on its previous values, such as y(t - 1), y(t - 2), etc. The order of the lag is denoted by p. I stands for integration, which transforms the time series into a stationary state using differencing. The order of the difference is denoted by d. MA stands for moving average, where the time series is regressed on the residuals of past observations, and the order of the lagged errors is denoted by q. In the above equation, y'(t) is the differenced series, ϕ t is the error term (Shweta). Furthermore, ARIMA works properly only with stationary time series, so it is necessary to check whether the data is stationary before using the ARIMA model. If the time series is non-stationary, it needs to be transformed into a stationary time series using differencing.

Prior to applying the ARIMA model, the data on yearly distribution of female pilots was transformed into a time series dataset. However, the transformed dataset exhibited non-stationary time series, indicating that it deviates from the stationary assumptions of constant mean, variance, and autocorrelation over time (Figure 4).

차트이(가) 표시된 사진

자동 생성된 설명

Figure 4: Time series of female pilot distribution

Therefore, to address the issue of non-stationarity, the "auto.arima" function in the R Forecast package was employed to automatically estimate the orders and coefficients of the ARIMA model. As a result, the ARIMA(1, 2, 0) model was found to be the most suitable, with an AIC value of 262.5353 (Figure 5).

테이블이(가) 표시된 사진

자동 생성된 설명

Figure 5: Best ARIMA model

After applying the ARIMA model, the time series data was found to exhibit stationarity (Figure 6) and follows a normal distribution (Figure 7).

차트이(가) 표시된 사진

자동 생성된 설명차트이(가) 표시된 사진

자동 생성된 설명

Figure 6: Time series after ARIMA model application

Figure 7: Normal Distribution

Furthermore, it can be observed from Figure 8 that the ACF and PACF of this ARIMA model fall within the normal range, indicating the presence of white noise. This is an important characteristic of a good ARIMA model, as white noise indicates that the model has successfully captured all the systematic patterns and trends in the data and that there are no remaining patterns or trends that can be exploited for further forecasting.

차트이(가) 표시된 사진

자동 생성된 설명

Figure 8: ACF & PACF

# Results

The data showed that 59.62% of the respondents were male and 40.38% were female (Figure 9). Men consistently outnumbered women in each year, but the ratio between the two genders remained relatively constant. The largest difference between the genders occurred in 2003 (Figure 10).

차트, 파이 차트이(가) 표시된 사진

자동 생성된 설명차트이(가) 표시된 사진

자동 생성된 설명

Figure 9: Gender distribution in STEM

Figure 10: Gender distribution of STEM by year

After graduating from STEM fields, the average annual salary for men is $20,000 higher than that of women. Overall, salaries increased until 2017 and then decreased in 2019 for both genders (Figure 11).

차트이(가) 표시된 사진

자동 생성된 설명

Figure 11: Average wage gap after graduation by year

When looking at the gender ratio based on average salary, women have a higher proportion than men in the lower salary group, while men have a higher proportion in the higher salary group (Figure 12). In addition, groups that receive an average salary of more than $150,000 are mainly distributed among people in their 30s to 60s, while those in their teens and 20s belong to the lower salary group (Figure 13).

차트이(가) 표시된 사진

자동 생성된 설명차트이(가) 표시된 사진

자동 생성된 설명

Figure 12: Gender ratio of Salary Group

Figure 13: Age ratio of Salary Group

In the field of aerospace, Asians on average receive higher salaries than other races, and Black people receive the lowest salaries (Figure 14). Additionally, there is an overall gender pay gap of over $20,000 (Figure 15).

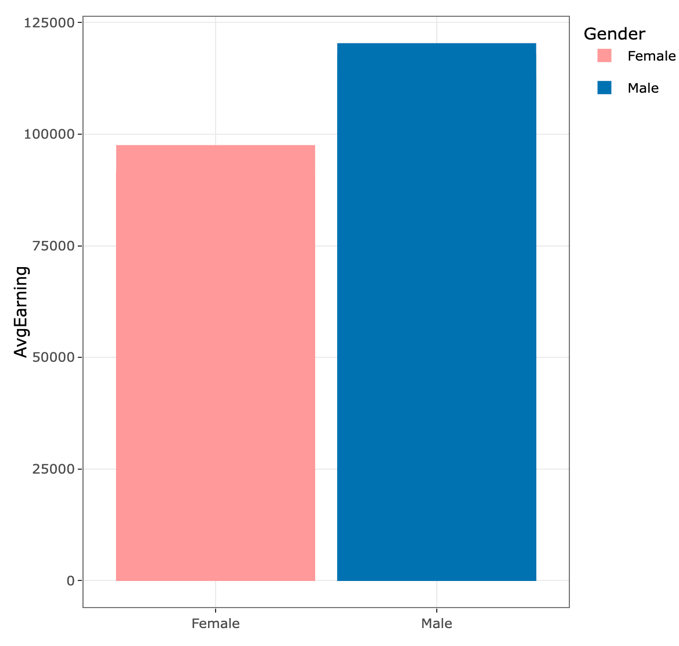
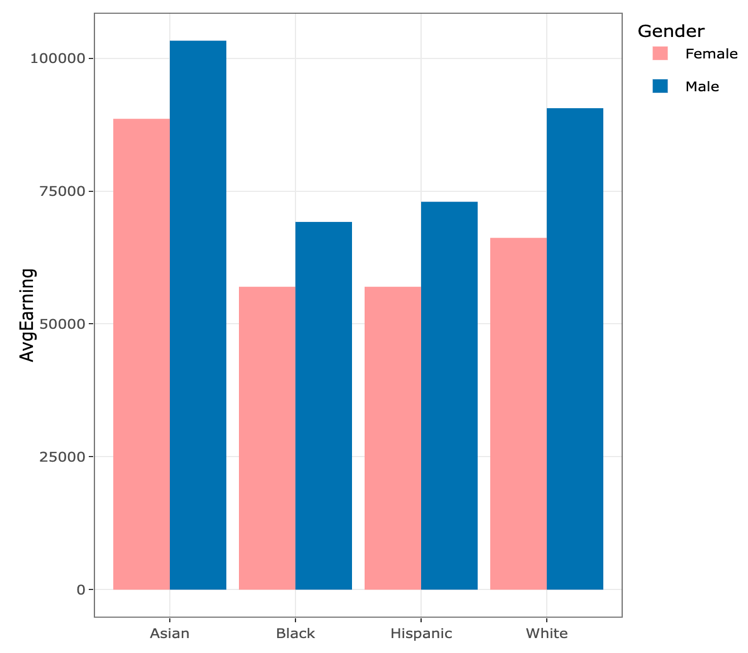


Figure 14: Race & Gender wage gap in Aerospace

Figure 15: Average gender wage gap in Aerospace

Men are generally much more heavily represented than women in various machinery-oriented job positions within the aerospace industry, while women are more heavily represented than men in the service-oriented role of flight attendants (Figure 16). Additionally, in higher-level positions, men are predominantly more prevalent than women (Figure 17).

차트이(가) 표시된 사진

자동 생성된 설명차트이(가) 표시된 사진

자동 생성된 설명

Figure 16: Ratio of Gender in Aerospace occupations

Figure 17: Gender leadership position ratio in Aerospace

After analyzing the projected distribution of female pilots in the future, it is anticipated that the proportion of women in this profession will steadily rise, and beyond 2025, there is a possibility of over a twofold increase in the number of women who are employed as pilots, as compared to the figures recorded in 2005.

차트이(가) 표시된 사진

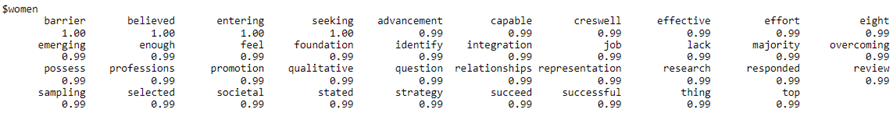
자동 생성된 설명

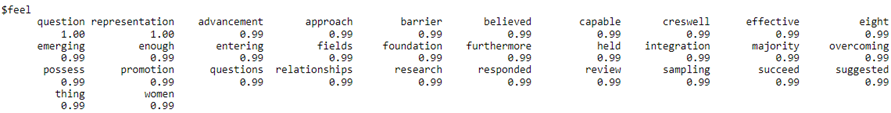
Figure 18: Predict after few years distribution of women pilots in aerospace

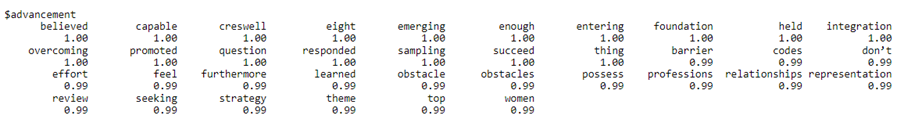
We were also given additional PDF text data to analyze, we were given 115 documents about women in STEM majors and 21 documents about women in aerospace. At first, we were going to analyze all the documents. However, due to our difficulty working with text analysis we decided to use the 21 relevant documents to answer our questions specific to women in aerospace. We decided to use the Syuzhet Package by Matthew Jockers, this package allowed us to use a sentiment extraction tool to compute sentiment data from text. Our goal in using this package is to see how women are talked about in academic papers and get some visualization on our 21 documents. Unfortunately, we were not able to apply the Syuzhet library to the 115 documents due to the errors we were getting when trying to read the 115 documents. In order to confirm that the 21 documents are talking about the topics we are analyzing, we developed a word cloud.

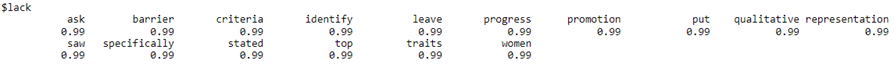
  
**Figure 19: Word cloud of Women in Aerospace**

As we can see in the word cloud the articles talked about women, aerospace, stem, participant, industry, management, gender, executive, and positions. Some words that stood out were advancement, management, executive, leadership, career, lack skills, and found bias. Although we don’t really have any context for these words, for the most part they tend to be positive with the exception of the words “lack of skills”. In order to get a better interpretation of these terms, we used the findAssocs() function which would help us find associations in a document-term matrix. For this function we set the corlimit to 0.99 to reduce the associated words displayed.

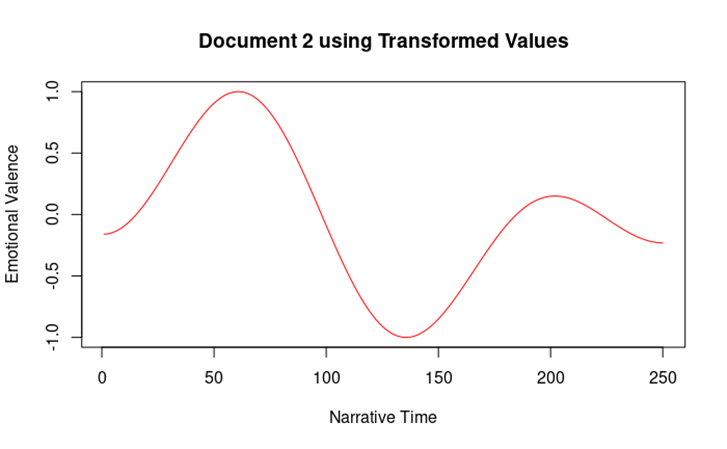
  
**Figure 20: Associations of the word “women”**

  
**Figure 21: Associations of the word “feel”**

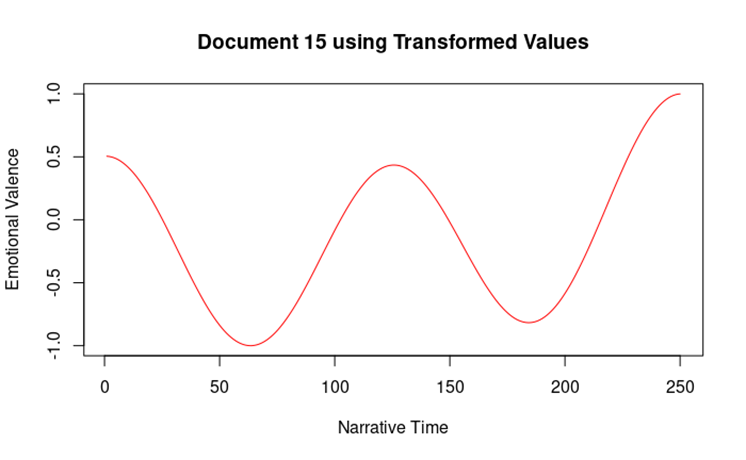
  
**Figure 22: Associations of the word “advancement”**

  
**Figure 23: Associations of the word “lack”**

After we saw the association for our words of interest, we only selected the ones that had better results because for some of our words we would have too many associations. We believe this was due to it being in the word cloud and because the same words would come up multiple times throughout the articles. We were glad to see that for the most part it was positive terms that were associated like advancement, capable, effective, promotion, and integration. However, we did notice some negative terms like barrier, overcoming, obstacle, and something that stood out was that the term “lack” was associated with “women” because in the word cloud we saw the terms “lack skills” came up. Therefore, we interpreted that women were lacking skills and was seen as a negative. Now, although it seemed there was a lot of positive terms associated with women in aerospace, that was bit always the case. As we started using the Syuzhet library, we were having difficulties applying the get\_sentiment() function due to the data frame where we had all the text data not applying correctly to the function. So, we had to manually copy and paste the text we were wanting to analyze and then save that information as a PDF file. Due to some articles being more than 50 pages, we were only able to format 14 out of the 21 documents in order to analyze the text. After we applied the get\_sentiment() function to the different articles and saved the result of our emotional valence of the text into a vector, we used the sum function on that vector to get an idea of what the overall emotional valence of the text was. It was not surprising that the article (Document 15) with the lowest emotional valence value was written in 1995 with a minimal score of -24.85, which meant that for the most part the emotional valence of this article was negative. As for the most recent article, it had an overall emotional valence of 30.95 which is not much of a difference, but it is positive instead of negative emotional valence and this article was written in 2021 (Document 2).



**Figure 24: Document 2 emotional valence throughout the article.**



**Figure 25: Document 15 emotional valence throughout the article.**

When trying to plot the data, we would end up with a lot of noise in the graph that made it hard to read the emotional valence of the article. We used the get\_dct\_transform() function on our emotional valence vector and then plot those values. This function applies discrete cosine transformation which helped better represent edge values by smoothing the values from our emotional valence vector. Something that really stood out was that for Document 15 the overall curve was under the 0 value on the emotional valence vector, while Document 2 was mainly above that 0 value. Something to note is it seems like Document 2 starts positive and ends by seeming to converge on 0 or neutral and Document 15 starts negative but ends positive. The reason we think there is much more alternating oscillations is because at the time in 1995, women in aerospace might have been having more difficult time in the workplace than in 2021. This gave us a similar inference from our earlier analysis of the PDFs text, where the overall message of women in aerospace is positive yet there could be some room for improvement due to the small but present negative emotional valence in the field.

# Conclusion

Our findings indicate that there is a notable gap in wages between men and women in the same position and of the same seniority. Additionally, it indicated that there is also a gap present between people of different races, discounting gender identity. However, both gaps have been seen to be shrinking in recent years. This is likely due to a societal shift in how diversity is seen, with more companies pushing to have diverse workforces due to public demand. Additional rationale for this shift could include government regulations for diversity and younger generations entering the workforce, which could be slowly changing the workplace culture to one that is more welcoming of others. While the results of this data may not be surprising to some, it is important to note that there is marked improvement across all datasets that were analyzed. In the future, it would be interesting to see how those with non-binary identities would be included in these datasets, as well as those who are transgender. However, it is unlikely that there will be enough data to analyze within the next few years, leaving it as merely a thought exercise.

## Works Cited

Boeing. “2022 Global Equity, Diversity & Inclusion Report.” *2022 GEDI Report*, <https://www.boeing.com/principles/diversity-and-inclusion/annual-report/index.page>.

Chaudhary, Shivam. “Why ‘1.5’ in IQR Method of Outlier Detection?” *Medium*, Towards Data Science, 26 Aug. 2020, <https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097>.

Fry, Richard, et al. “Stem Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity.” *Pew Research Center Science & Society*, Pew Research Center, 1 Apr. 2021, <https://www.pewresearch.org/science/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/>.

Meta. “7 Benefits of Gender Diversity in the Workplace.” *Workplace from Meta*, <https://www.workplace.com/blog/diversity-in-the-workplace>.

Shweta. “Introduction to Time Series Forecasting - Part 2 (Arima Models).” *Medium*, Towards Data Science, 30 July 2021, <https://towardsdatascience.com/introduction-to-time-series-forecasting-part-2-arima-models-9f47bf0f476b#:~:text=The%20ARIMA%20equation%20is%20a,(t%E2%88%92q)%20%2B%20%CE%B5t>.