Does a Gender Gap Exist in the Field of Data Science & Machine Learning?

Arushi Kapoor

DATS 6501 Data Science Capstone

Section 11

April 28, 2022

TABLE OF CONTENTS

GLOSSARY

1. **Gender Gap** - “The difference between women and men as reflected in social, political, intellectual, cultural, or economic attainments or attitudes” (*World Economic Forum*).
2. **Gender Pay Gap** - “The difference between median earnings of men and women relative to median earnings of men” (*Organization for Economic Cooperation and Development).*
3. **Data Preprocessing** - “Putting the data into the right shape and quality, for visualization and training. Preprocessing strategies include: data cleaning, balancing, replacing, imputing, partitioning, scaling, augmenting and unbiasing” (*Amazon Web Services*).
4. **Exploratory Data Analysis** - “An approach/philosophy for data analysis that employs a variety of techniques, mostly graphical, to -
    a. Maximize insight into a dataset;
    b. Uncover underlying structure;
    c. Extract important variables;
    d. Detect outliers and anomalies;
    e. Test underlying assumptions;
    f. Develop parsimonious models; and
    g. Determine optimal factor settings”

    (*National Institute of Standards and Technology, U.S. Department of Commerce*).
5. **Data Visualization** - The graphical representation of information and data, through visual elements, such as charts, graphs, and maps, to provide an

accessible way to see and understand trends, outliers, and patterns in data (*Tableau*).

6. **Training Data** - “The initial dataset used to train machine learning algorithms. Models create and refine their rules using this data. It's a set of data samples used to fit the parameters of a machine learning model to training it by example” (*G2*).
7. **Testing Data** - *“*The data used to evaluate the performance or accuracy of the model. It's a sample of data used to make an unbiased evaluation of the final model fit on the training data” (*G2*).
8. **Feature Importance** - “The value that quantifies the intensity of association between the model outcome and the predictor variable individually or as a set” (Wei, P. et. al, 2015).
9. **Light Gradient Boosting Machine Classifier** - Originally developed by *Microsoft Corporation*, Light Gradient Boosting Machine is a gradient boosting framework that uses tree based learning algorithms for ranking and classification in machine learning. It is designed to be distributed and efficient due to better accuracy, faster training speed, lower memory usage and capability to handle large-scale data (Microsoft Corporation).
10. **Random Forest Classifier** - Developed by *Scikit-Learn,* a free software machine learning library for the Python programming language, Random Forest Classifier is “a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting” (*Scikit-Learn*).

11. **K-Nearest Neighbors Classifier** - Originally developed by Evelyn Fix and Joseph Hodges in 1951, the K-Nearest Neighbors (KNN) algorithm is “a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to” (*G2*).
12. **Hyperparameter Tuning** - “A common machine learning workflow that involves appropriately configuring the data, model architecture, and learning algorithm to yield an effective model” (*Determined AI*).
13. **Cross Validation** - “Cross-validation provides information about how well a classifier generalizes, specifically the range of expected errors of the classifier” (*Scikit Learn*).
14. **Confusion Matrix** - A table, used to define, visualize and summarize the performance of a classification algorithm and consists of four numbers - True Positive, True Negative, False Positive and False Negative (Singh, P. et. al, 2021).
15. **Accuracy** - “The number of correctly predicted data points out of all the data points; ratio of the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives” (*Deep AI*).

INTRODUCTION

Background

In 1945, the United Nations Charter, enshrined “equality between men and women” as its core tenet. 75 years later, in 2022, girls and women continue to live in a world of widespread gender equality (*United Nations*). The difference between men and women, as reflected in social, political, intellectual, cultural, and economic attainments or attitudes across the world, is commonly referred to as the **Gender Gap** (*World Economic Forum*).

In the past decades, much has been accomplished to inspire girls and women to study and work in technical fields. As a result, women have felt empowered and have made strides in STEM (science, technology, engineering and mathematics) fields. However, even today, girls and women continue to be excluded from participating fully, as conveyed by the following three facts -

1. There’s a STEM gender gap
2. Only a fraction of female students select STEM-related fields in higher education
3. Bias and gender stereotypes can drive away women in STEM

   (*United Nations*).

Problem Statement

As discussed above, the gender gap continues to be one of the most persistent issues in the field of technology, where women have been historically excluded. Several factors, such as age, education level, job title, country of residence and others, can be explored to determine whether a gap, between men and women, exists in the field of

Data Science and Machine Learning. Therefore, the purpose of this study is to use visualization and machine learning techniques to examine and highlight such factors, understand the causes that contribute to the gap, and most importantly, make recommendations towards closing it.

Problem Elaboration

Through visualization and machine learning model implementation, this study seeks to answer the following questions in the field of Data Science & Machine Learning.

1. How do women compare to men, in terms of
    a. Age
    b. Education Level
    c. Job Title
    d. Country of Residence/Nationality
    e. Years of Coding/Programming Experience
2. How are men and women distributed over different compensation ranges?
3. Is there a pay gap between men and women?
4. If a gap does exist, has it changed over the years?
5. Is an individual's gender a primary determinant of his/her compensation?
6. What kind of models can be implemented on the given data to predict an individual's compensation?
7. What measures can be taken to reduce any existing pay gap between men and women?

Motivation

Gender equality and equity is a crucial component of a sustainable society. Pay gaps, between men and women, still exist in several fields of employment, especially technology. To address the same, it is necessary to understand the representation of women and girls and the attitudes towards women and girls in the field.

The main motivation behind this study is to use data science for social good. Women in data science are essential in today's data-driven world, to overcome internal biases while building machine learning algorithms. The first step in understanding the representation of women and girls in the field is to examine, visualize and analyze real data on the same through effective techniques such as data visualization, sentiment analysis, feature importance and machine learning model implementation.

Scope

The scope of this study is as follows -

1. Data Visualization - This includes exploratory data analysis, pay gap visualization and twitter sentiment analysis, conducted through R programming language. This is hosted through a Shiny Application, prepared on RStudio.
2. Feature Importance - This includes ranking of important variables in the chosen dataset through Light Gradient Boosting Machine Classifier and Random Forest Classifier algorithms, implemented through Python programming language.

3. Machine Learning - This includes prediction of the outcome or target variable, compensation, through the implementation of three models - K-Nearest Neighbors Classifier, Light Gradient Boosting Machine Classifier and Random Forest Classifier, using Python programming language.

LITERATURE REVIEW

Relevant Research

As data continues to shape the future of the economy, the world has observed a rise in women technologists, including data scientists. However, according to the *Global Gender Gap Report 2021*, published by the *World Economic Forum*, **women make up only 32% of the workforce in Data and AI** (p. 6)**.** According to *O'Reilly's 2021 Data/AI Survey*, which was limited to respondents in the United States and United Kingdom, "**women's salaries are significantly lower than men's, equating to 84% of the average salary for men regardless of education or job title. For example, at the executive level, the average salary for women was $163,000 versus $205,000 for men, O'Reilly found — a 20% difference**" (VentureBeat, 2021). These statistics indicate that women are drastically underrepresented in the field of Data Science and Machine Learning.

In 2021 *The Alan Turing Institute*, published a report, *Where are the Women?,* to map the gender job gap in the field of Artificial Intelligence and Data Science. The report highlighted the following -

1. **Limited availability of data:** The available data is "fragmented, incomplete and inadequate for investigating the career trajectories of women and men in the fields." Additionally, most data is collected through surveys, which not just lacks detailed information about job titles and pay levels, but is also prone to human error.
2. **Differences in career trajectories:** Men and women have different career trajectories in the field. Further, "women are more likely than men to

occupy a job associated with less status and pay in the data and AI talent pool, usually within analytics, data preparation and exploration, rather than the more prestigious jobs in engineering and machine learning."

3. **Representation in industries:** The report states that "women in data and AI are under-represented in industries which traditionally entail more technical skills (for example, the Technology/IT sector), and overrepresented in industries which entail fewer technical skills (for example, the Healthcare sector). Furthermore, there are fewer women than men in C-suite positions across most industries, and this is even more marked in data and AI jobs in the technology sector."
4. **Self-reporting skills:** Men are known to routinely self-report their skills on LinkedIn, more than women. This correlates with women's lower confidence levels in their own technical abilities.
5. **The qualification gap:** It has been reported that women in data and artificial intelligence have higher qualifications as compared to their male counterparts. Additionally, "the achievement gap is even higher for those in more senior ranks (i.e. for C-suite roles), and this 'over-qualification' aspect is most marked in the Technology/IT sector."
6. **Differences in participation in online platforms:** It has been reported that "women comprise only about 17% of participants across the online global data science platforms Data Science Central ('DS Central'), Kaggle and OpenML. On Stack Overflow, women are a mere 8%."

(Young, Erin et. al, 2021, p. 4-5)

Further research reveals that gender doesn't represent the most important predictor (feature) for compensation determination. This can be corroborated through the results of the three studies.

A study titled, *Can Gender Role Items Improve the Prediction of Income? Insight from Machine Learning,* conducted by Klara Raiber, at University of Mannheim, examines the issue of gender pay gap through the implementation of a Random Forest Regression model on the 2012 International Social Survey Program data. The model including sex (used as a proxy for gender) and other control variables, to evaluate prediction performance and variable importance, concludes that sex is the 10th most important variable (predictor) (2019). Another study, *Kagglers' Gender Pay Gap & Salary Prediction*, conducted by Theo Viel, a Kaggle Competitions Grandmaster, examines the issue of gender pay gap between male and female Kagglers using the 2018 Kaggle Machine Learning & Data Science Survey. The implemented Light Gradient Boosting Machine Classifier model reveals that profession is the most important feature for the purposes of salary prediction, while gender ranks last among all the chosen variables - Industry, Profession, Country, Age, Experience, Major of Study, Education Level & Gender (2018). A third study titled, *The confidence gap predicts the gender pay gap among STEM graduates,* published by Stanford Graduate School of Business, applied an Ordinary Least Squares Regression of the dependent variable, Log Annual Salary, on independent variables, Being Female, Engineering Self-Efficacy, Importance of Compensation and Importance of Workplace Culture. The results of the model indicate that gender does not predict the importance placed on salary. Also, the results suggest, "addressing cultural beliefs as manifested in

self-beliefs—that is, the confidence gap—commands attention to reduce the gender pay gap" (Sterling et. al, 2020, p. 3-4).

Additional research was conducted to understand the evaluation through accuracy and precision scores, of the machine learning prediction models on salary determination of survey respondents. Further, research has indicated that classifier models have proven to be more effective for the prediction of income levels of individuals based on attributes including education, gender, occupation, country and others. Also, these models have been able to show the top features in importance (Bekena, 2017). *Kagglers' Gender Pay Gap & Salary Prediction*, conducted by Theo Viel, involved the implementation of a Light Gradient Boosting Machine Classifier model to predict the respondents' salary class, ranging from 0 to 6, using the 2018 Kaggle Machine Learning & Data Science Survey. The model reflects an accuracy of approximately 49.86% on the testing data (2018).

Another project titled, *Kaggle-Survey-Salary-Prediction,* published on *Github*, involved the implementation of four regression algorithms, Gradient Boosting Regressor, K Nearest Neighbors, Random Forest Regressor and Lasso Regularization to predict the respondents' salary, a continuous value, using the 2018 Kaggle Machine Learning & Data Science Survey. The study reported that Random Forest Regressor has the highest accuracy score, of approximately 50.08%, on the testing data (2019).

Overall, research states, the Random Forest algorithm, tuned with important parameters, is known to generate the best efficiency and accuracy for salary prediction, based on the information provided by respondents in questionnaires. It is also useful to identify the key factors that explain the difference between high and low income.

METHODOLOGY

Dataset Description

For this study, the chosen dataset comprises observations from 2019, 2020 and 2021 Kaggle Machine Learning & Data Science Survey. The survey was conducted by *Kaggle Inc.*, a subsidiary of *Google LLC* and the largest online community of data scientists and machine learning practitioners. It consists of multiple choice questions and "presents a truly comprehensive view of the state of data science and machine learning" (Kaggle, 2021). It includes "raw numbers about who is working with data, what's happening with machine learning in different industries, and the best ways for new data scientists to break into the field" (Kaggle, 2021). The responses have been received from "data scientists and ML engineers, on their backgrounds and day to day experience – everything from educational details to salaries to preferred technologies and techniques" (Kaggle, 2021).

Data Collection

For the purpose of this study, the following answers' were selected from 2019, 2020 and 2021 survey multiple choice questions and were consolidated into a Master dataset -

1. Age Group
2. Gender
3. Country of Residence/Nationality
4. Job Title

5. Education Level
6. Programming/Coding Experience in Years
7. Programming Language Known: Python
8. Programming Language Known: R
9. Programming Language Known: SQL
10. Programming Language Known: Java
11. Programming Language Known: C
12. Programming Language Known: C++
13. Compensation Amount

Data Preprocessing

Data preprocessing was conducted using the R programming language on RStudio. The process involved the following steps -

1. Appending the **2019**, **2020** and **2021** responses for the above chosen variables.
2. Renaming the missing observations as **Not Provided**.
3. Filtering the appended dataset to two genders - **Man** & **Woman**.
4. Creating the following new variables -
    a. **Compensation Declared** - Yes or No; based on whether a respondent has selected a compensation range.
    b. **Compensation Group** - A revised range; based on the respondent's declared compensation range in the survey. Missing compensation groups have been renamed to **Unknown**.
    c. **Lower** - The lower level of the compensation group.

d. **Upper** - The upper level of the compensation group.

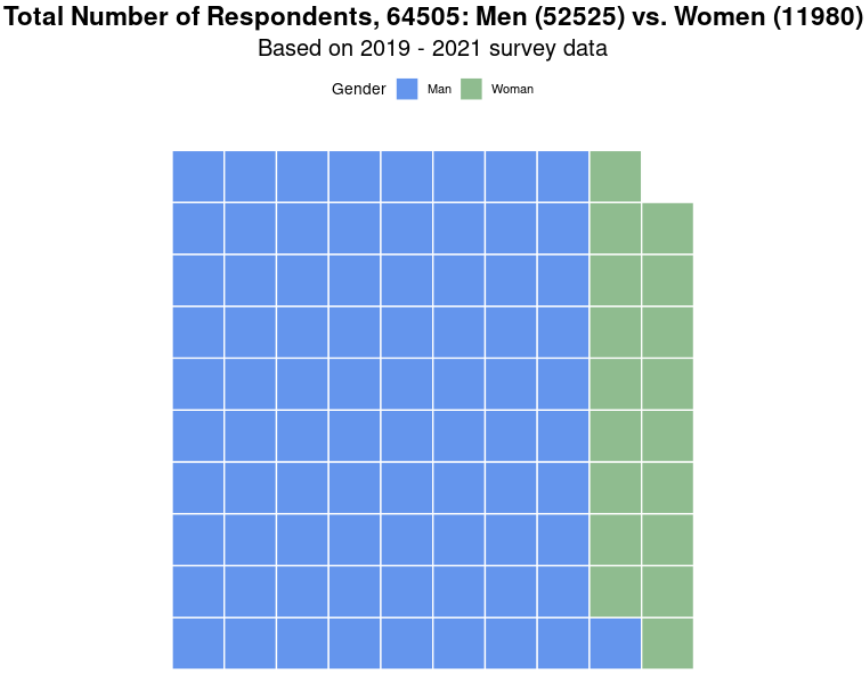
e. **Mid** - The average of the lower and upper levels of the compensation group.

f. **Survey Year** - 2019, 2020 and 2021.

Exploratory Data Analysis

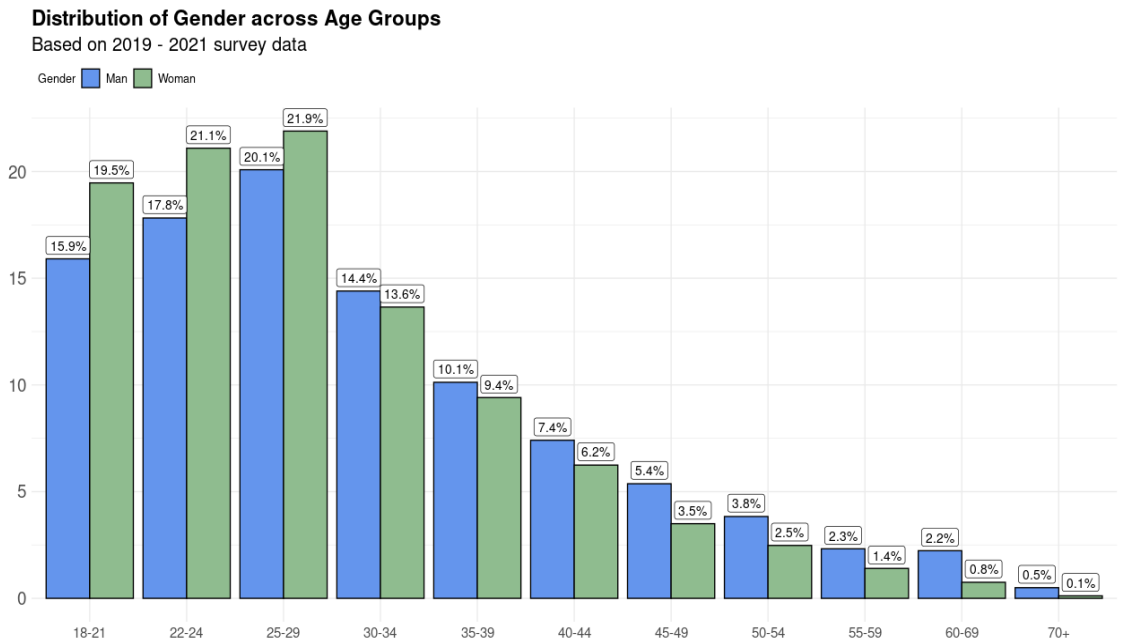
The purpose of exploratory data analysis is to visually examine the distribution of both men and women respondents, with respect to the following variables -

1. Number of Respondents - There are a total of **64,505** respondents, of which **52,525** are men and **11,980** are women.

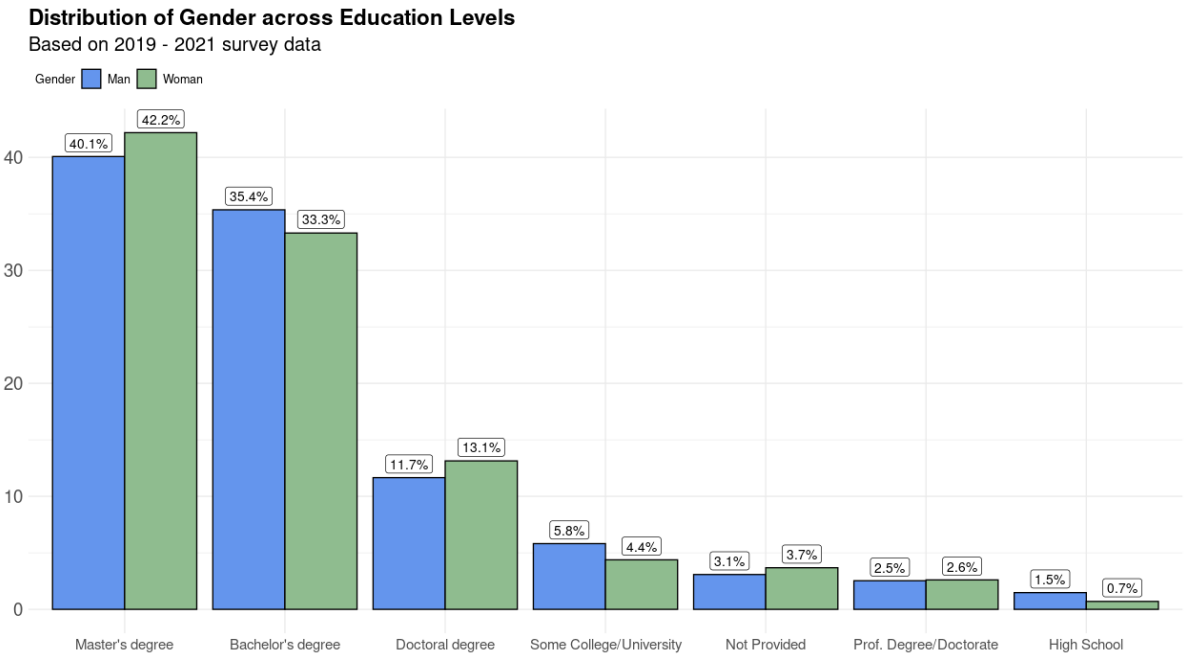


2. Age Group - Of the total 52,525 male respondents, **the highest percentage of male respondents, approximately 20%, is in the age group 25 - 29**. Similarly,

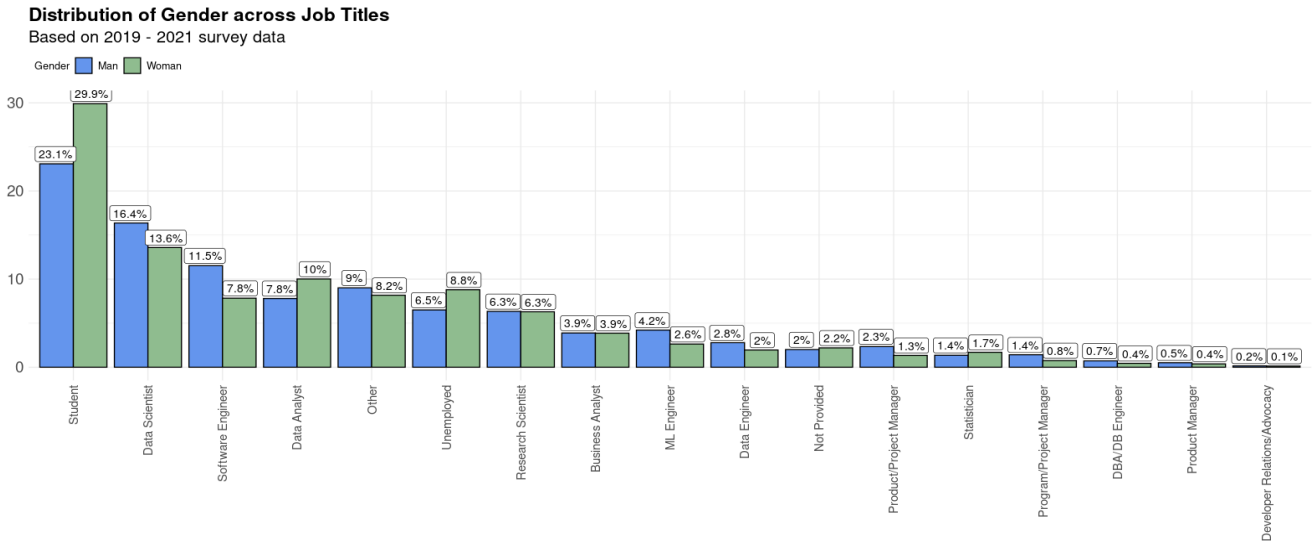
of the total 11,980 female respondents, **the highest percentage of female respondents, approximately 22%, is in the age group 25 - 29**.



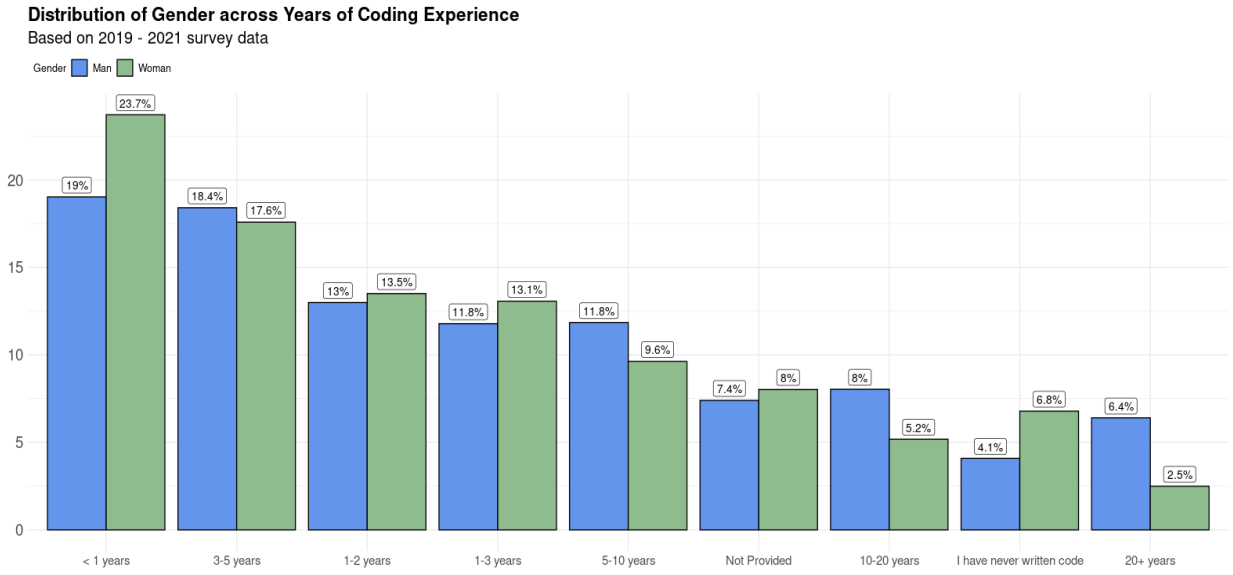
3. Education Level - Of the total 52,525 male respondents, **the highest percentage of male respondents, approximately 40%, have attained an education level equivalent to a Master's degree**. Similarly, of the total 11,980 female respondents, **the highest percentage of female respondents, approximately 42%, have attained an education level equivalent to a Master's degree.** It can also be observed that of the total number of female respondents, approximately 13%, have attained an educational qualification of a doctoral degree, whereas, of the total number of male respondents, approximately 11.7% have attained an educational qualification of a doctoral degree.



4. Job Title - Of the total 52,525 male respondents, **the highest percentage of male respondents, approximately 23%, are students**. **The second highest percentage of respondents, approximately 16.4%, work as a Data Scientist**. Similarly, of the total 11,980 female respondents, **the highest percentage of female respondents, approximately 30%, are students. The second highest percentage of respondents, approximately 13.6%, work as a Data Scientist**.

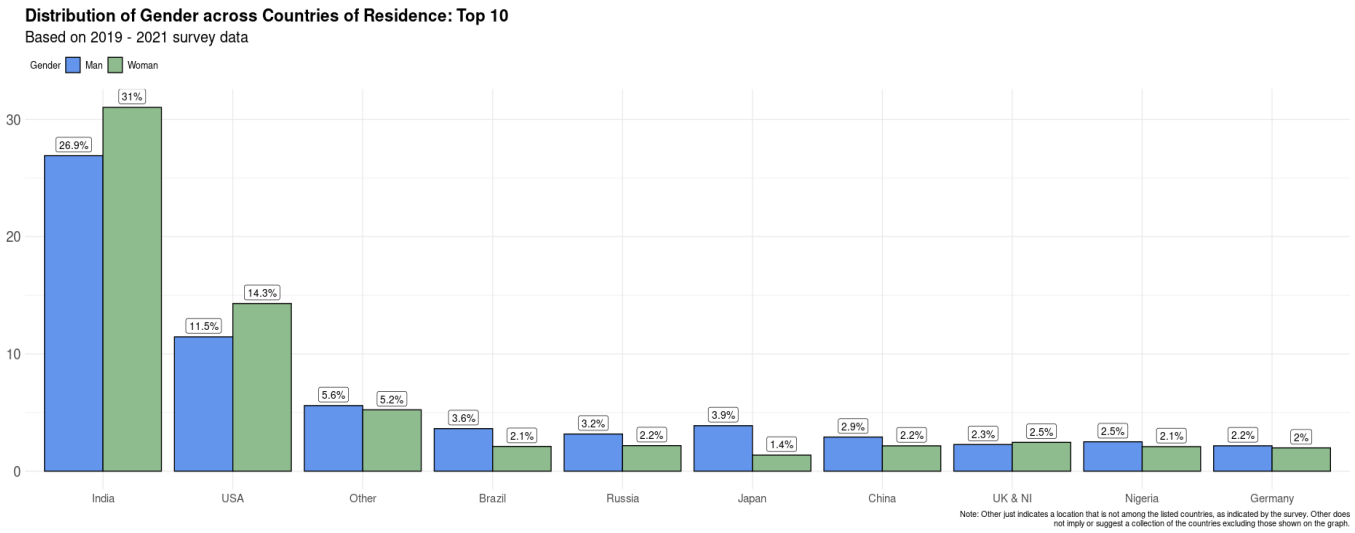


5. Programming/Coding Experience in Years - Of the total 52,525 male respondents, **the highest percentage of male respondents, approximately 19%, have less than 1 year of coding experience**. **The second highest percentage of respondents, approximately 18.4%, have 3 - 5 years of experience.** Similarly, of the total 11,980 female respondents, **the highest percentage of female respondents, approximately 23.7%, have less than 1 year of coding experience. The second highest percentage of respondents, approximately 17.6%, have 3 - 5 years of coding experience**.

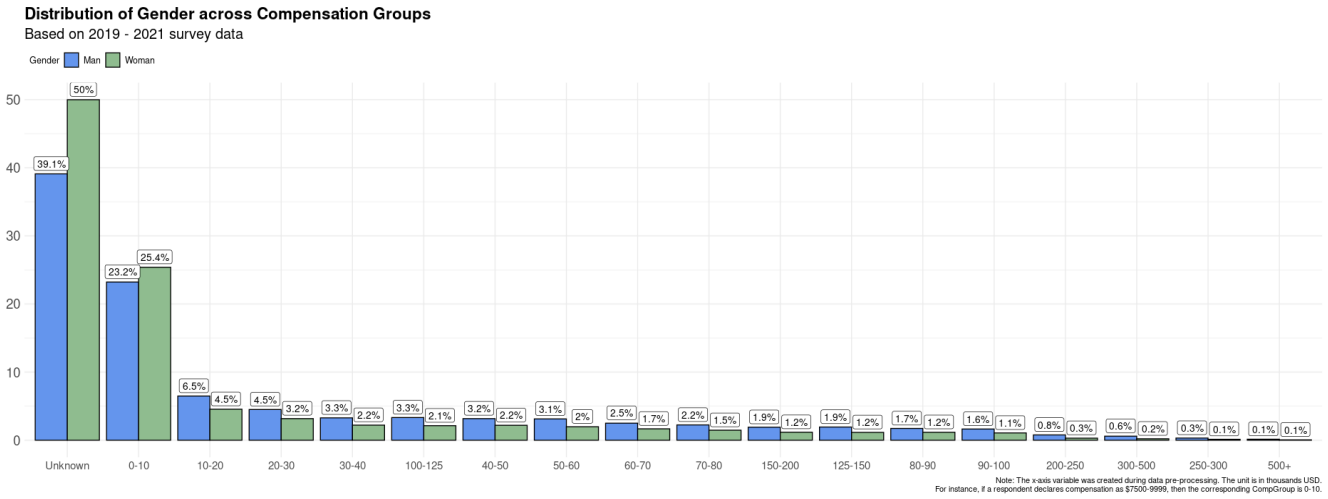


6. Country of Residence/Nationality - Of the total 52,525 male respondents, **the highest percentage of male respondents, approximately 27%, are from India**. **The second highest percentage of respondents, approximately 11.5%, are from the United States of America.** Similarly, of the total 11,980 female respondents, **the highest percentage of female respondents, approximately 31%, are from India. The second highest percentage of respondents, approximately 14.3%, are from the United States of America.**

The following visual depicts the top 10 countries from which the respondents are based. It should be noted that **Other** just indicates a location that is not among the listed countries, as indicated by the survey. **Other does not imply or suggest a collection of the countries excluding those displayed**.



7. Compensation Group - Of the total 52,525 male respondents, **approximately 39%, have not declared their compensation group.** Similarly, of the total 11,980 female respondents, **approximately 50%, have not declared their compensation group. Most respondents fall within the 0 - 10K USD group.**



Data Modeling

Data modeling was conducted on the 2021 Kaggle Data Science and Machine Learning survey data to determine the most and the least important features that contribute to compensation prediction and to determine the accuracy of the data to predict a respondent's compensation.

The chosen dataset was pre-processed for training and testing purposes using the Python programming language on Jupyter Notebook, as follows -

1. Remove observations that did not consist of compensation information.
2. Calculate the respondents' average age using the lower and upper range value.
3. Calculate the respondents' programming experience using the lower and upper range value.
4. Classify the target variable, Compensation Group, into seven classes -
    a. Less than 10K
    b. Between 10K and 30K
    c. Between 30K and 50K
    d. Between 50K and 80K
    e. Between 80K and 100K
    f. Between 100K and 500K
    g. More than 500K
5. Label the column values through category encoding.
6. Split into training (70%) and testing (30%) sets.

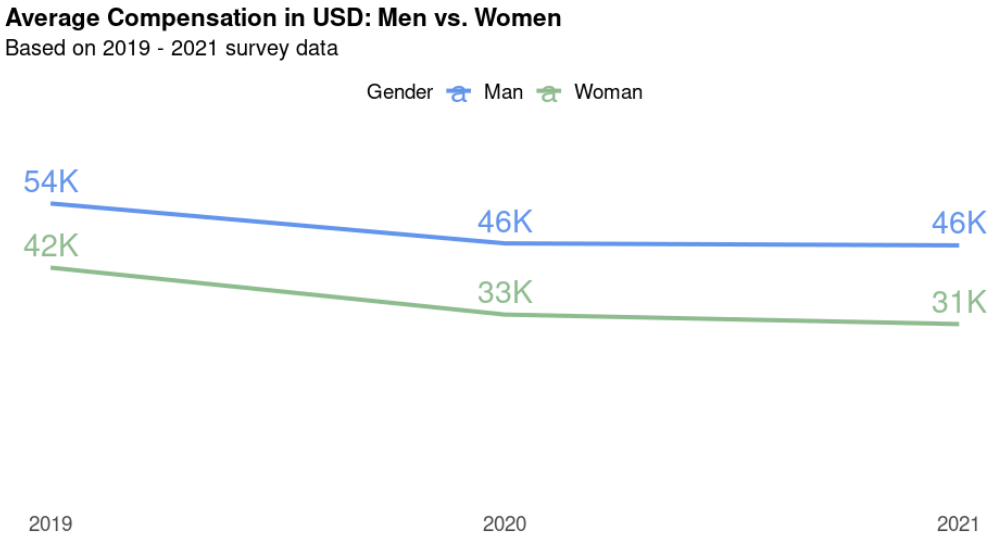
The preprocessed dataset was then used for machine learning purposes.

RESULTS & ANALYSIS

Dissecting the Pay Gap

The first step towards dissecting the pay gap between the two genders is to examine the average compensation trend from 2019 to 2021 for both men and women. It is important to note that the displayed compensation values may not reflect current market salary rates. This can be attributed primarily due to the currency conversion factor. Since the survey was conducted worldwide, respondents were asked to indicate their salaries in United States Dollar. Therefore, respondents outside of the United States may indicate salaries which seem quite lower than the average salary of a professional in the field in the United States of America.

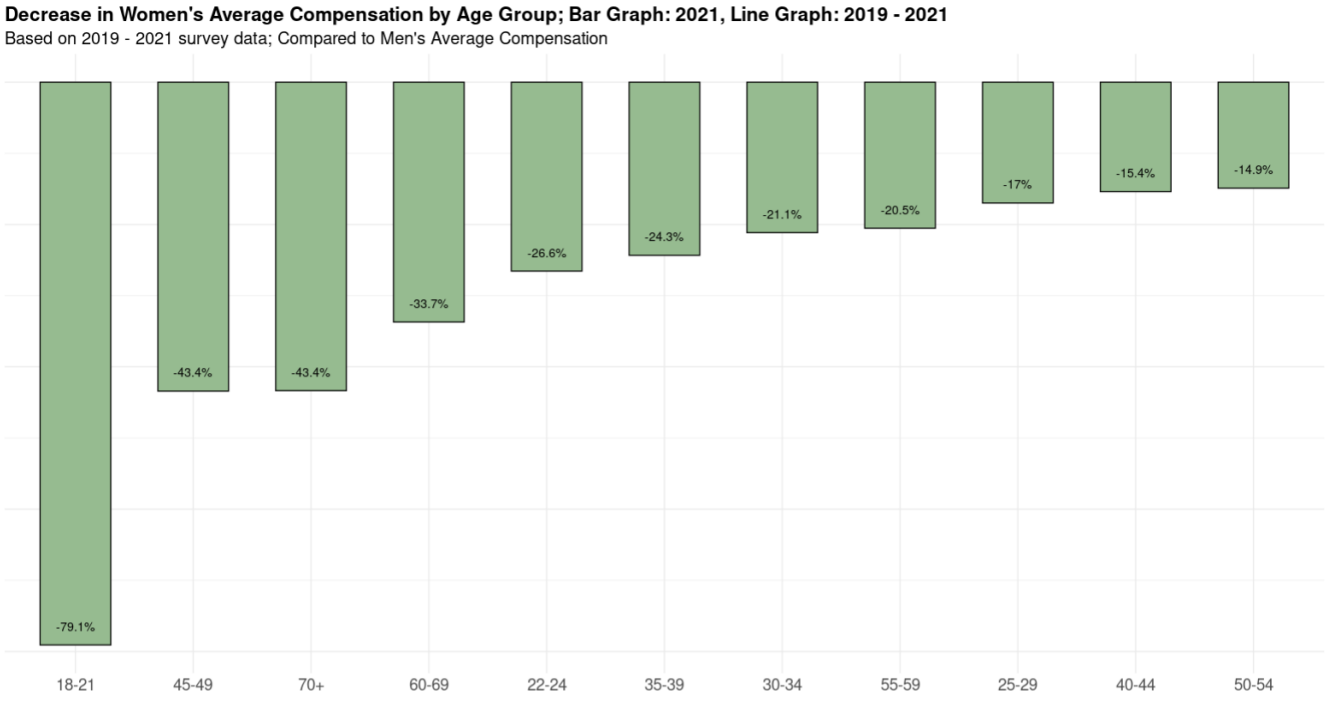
The following visual depicts the trend between the compensations of men and women across the given years. Compared to men's average compensation, women's average compensation has continued to decrease. The gap has been wider in 2020 and 2021, when the COVID-19 pandemic was at play.



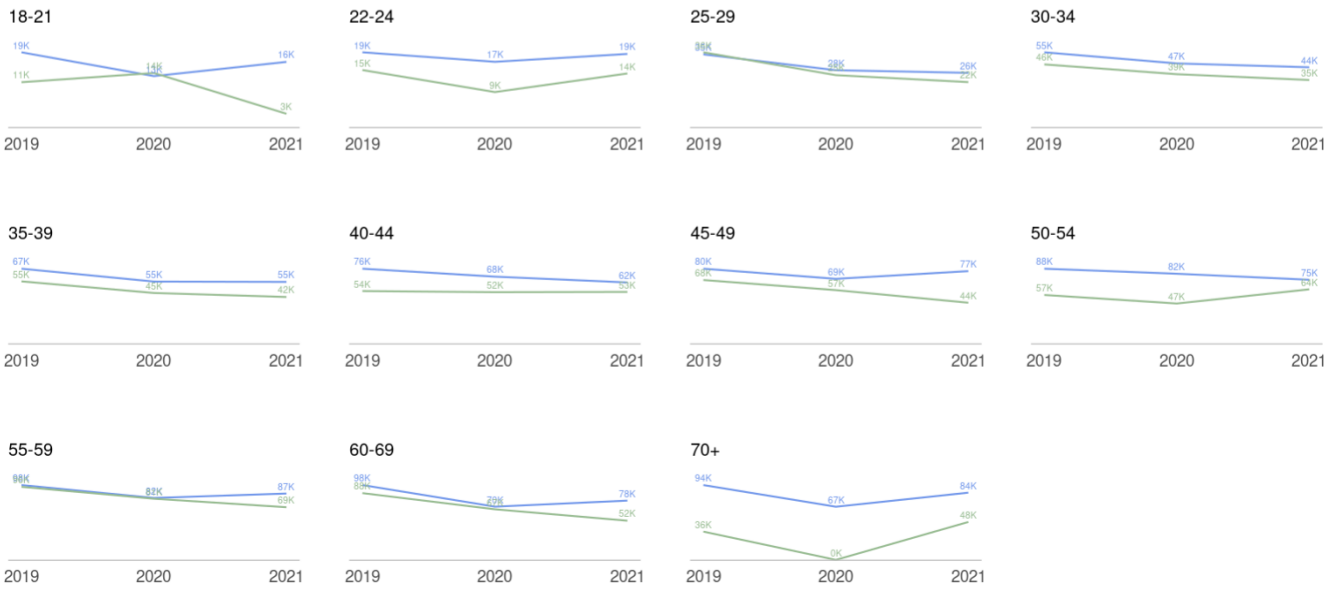
The next step is to dissect the pay gap between men and women respondents across the following factors -

1. Age

Based on the given bar graph, it can be seen that, **in 2021**, **women in the age range 18 - 21 had the widest pay gap.** Their average compensation is approximately 80% less than the men in the same age range.

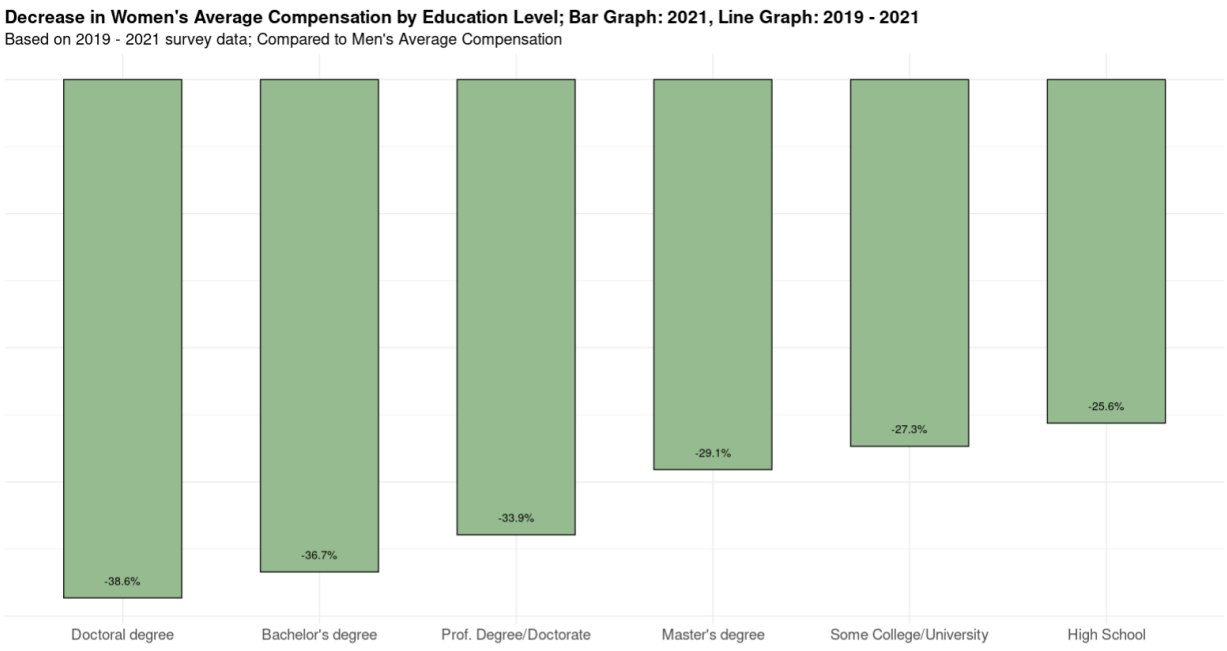


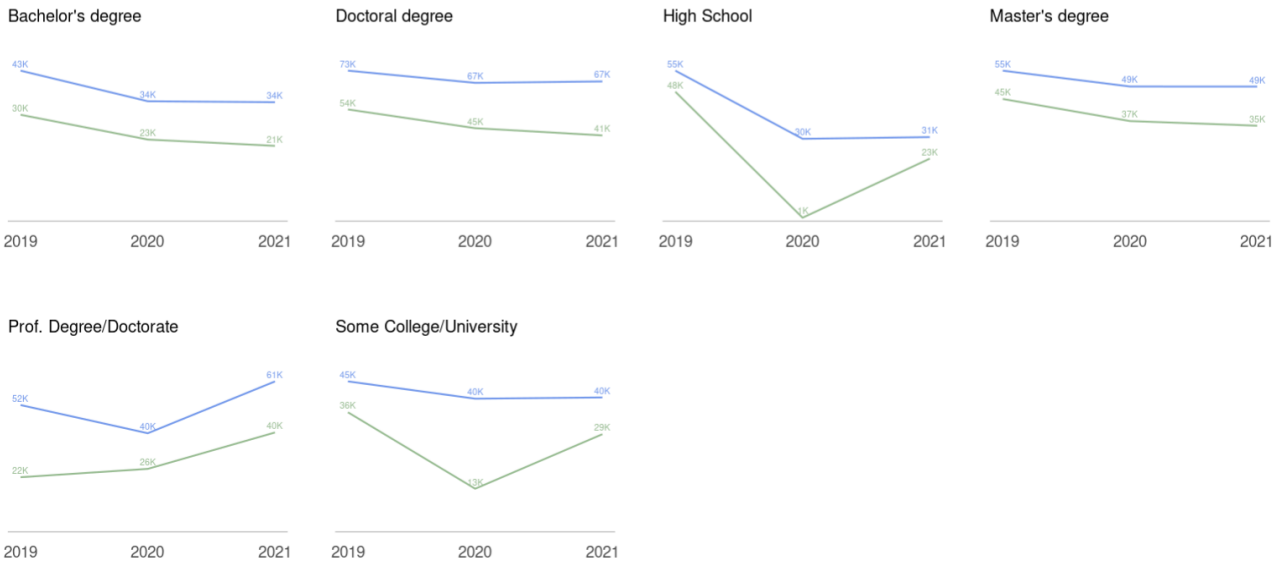
The following line graph shows there exists a pay gap between men and women across most age groups throughout the given time period. However, **the gap between men and women was at its minimum for the age groups 18 - 21 and 55 - 59 in the year 2020.** Strikingly, the gap has widened in 2021 for most age groups.



2. Education Level

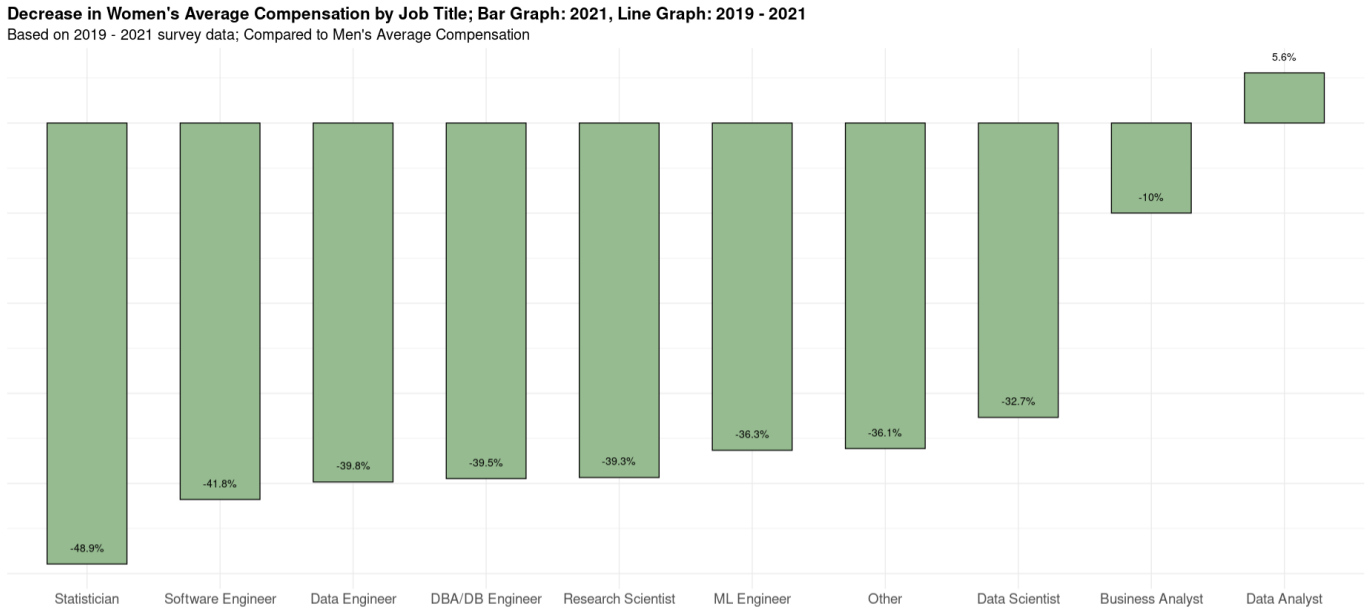
It can be seen that, **in 2021, the gap between men and women was the widest for those with a doctorate degree**. The line graph shows that a gap exists between men and women across all degree levels, throughout the given time period.





3. Job Title

It can be seen that, **in 2021, the gap between men and women was the widest for Statisticians**. Interestingly, **in 2021, women made approximately 6% more than men as Data Analysts**.

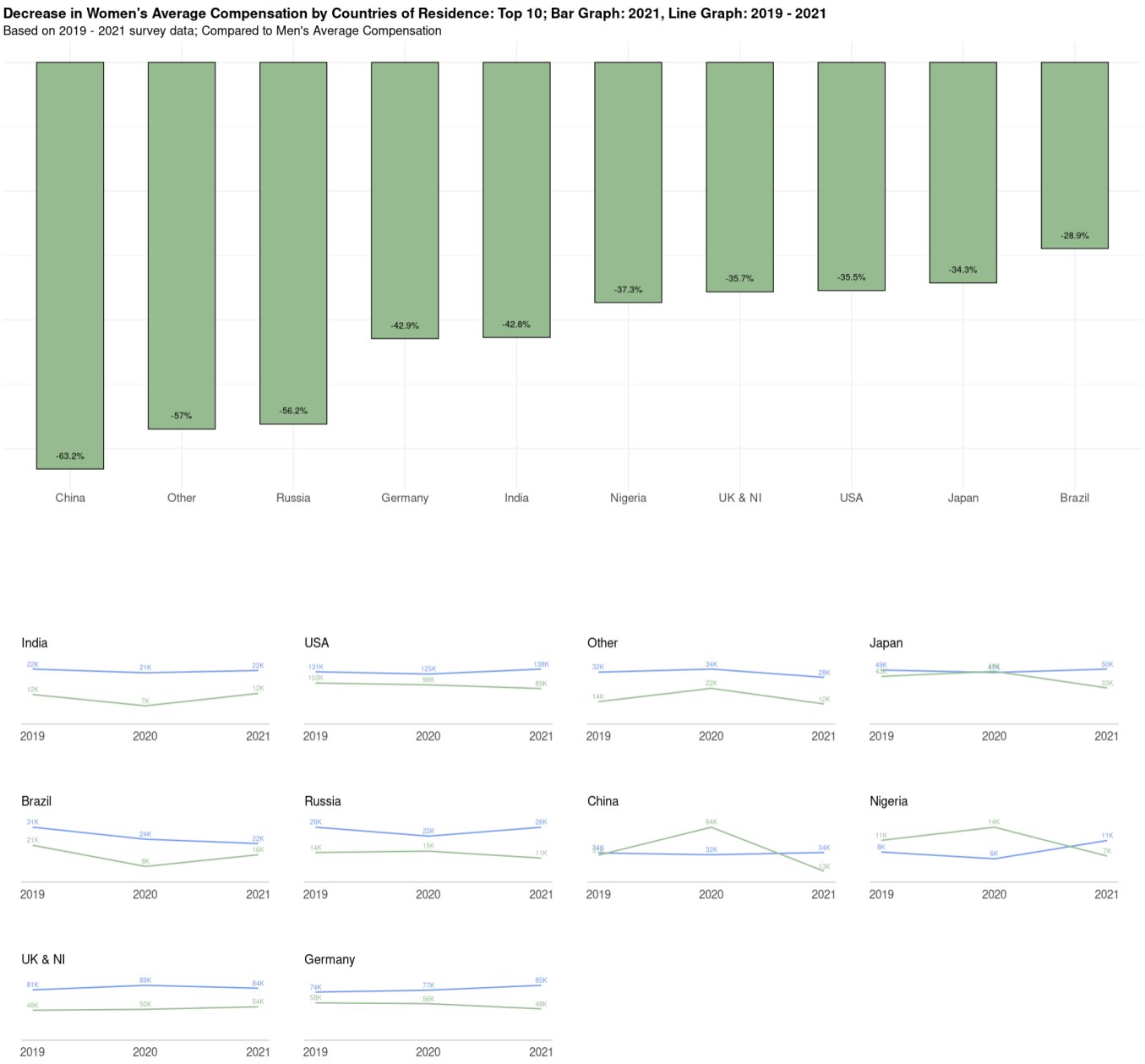


The line graph shows that a gap exists between men and women across all positions, throughout the given time period. However, the gap seems to have narrowed between male and female Product/Project Managers in 2020.

4. Country of Residence/Nationality

It can be seen that, **in 2021, the gap between men and women was the widest in China. The gap between men and women was the lowest in Brazil in 2021**.

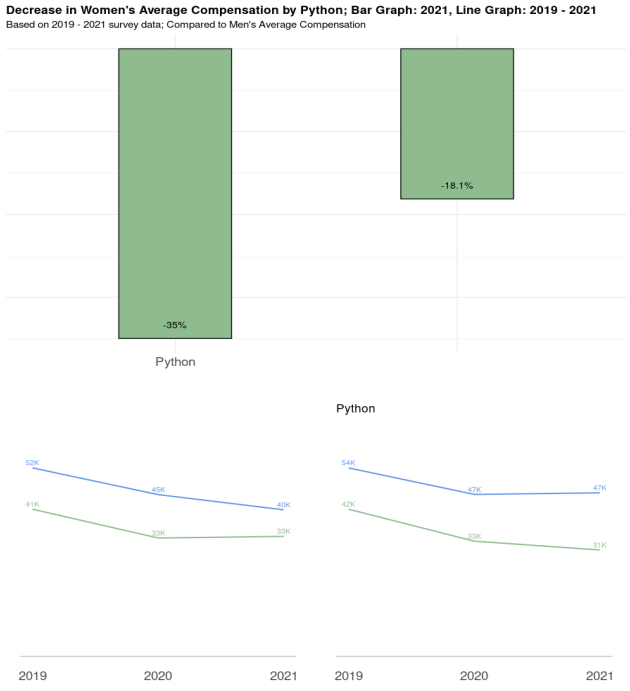
The line graph shows that a gap exists between men and women across most examined countries of residence/nationality, throughout the given time period. Interestingly, women respondents from China and Nigeria earned a lot more than their male counterparts in the year 2020. Similarly, the gap was non-existent between male and female respondents from Japan in 2020. Interestingly, the gap became more prevalent among males and females in 2021, when COVID-19 pandemic was still at play.

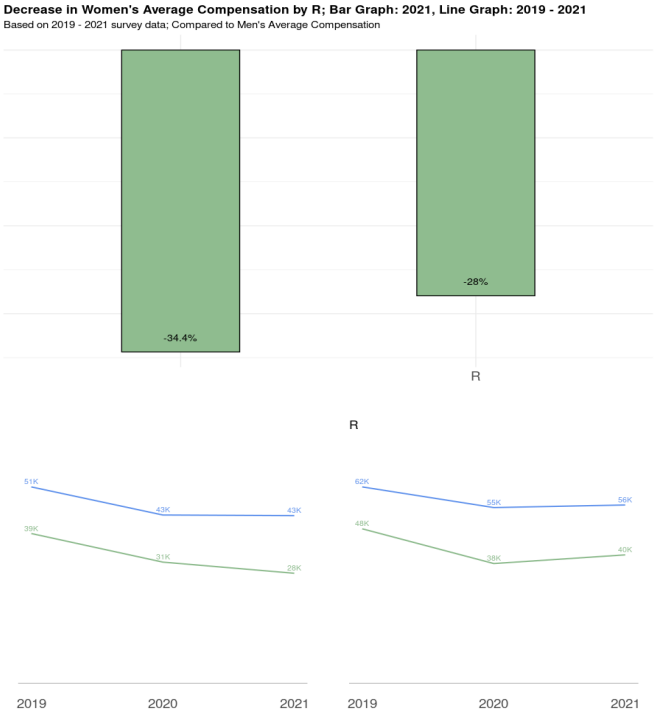


5. Programming/Coding Experience in Years

It can be seen that **in 2021, women who reported to have the same programming skills as men, earned less than their male counterparts**. This can be observed across all the examined programming languages: Python, R, SQL, Java, C and C++. Strikingly, **women who indicated the absence of knowledge of programming in Python, R, Java, C and C++, suffered a lesser pay gap as**

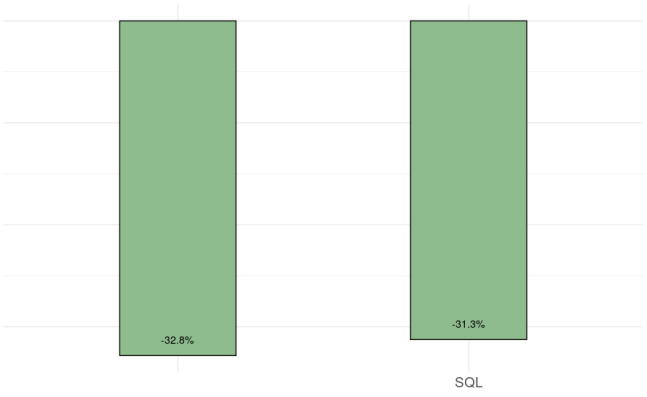
compared to those who indicated the presence of knowledge of programming languages.



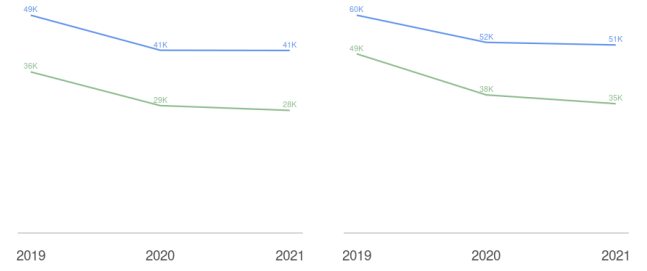


Decrease in Women's Average Compensation by SQL; Bar Graph: 2021, Line Graph: 2019 - 2021

Based on 2019 - 2021 survey data; Compared to Men's Average Compensation

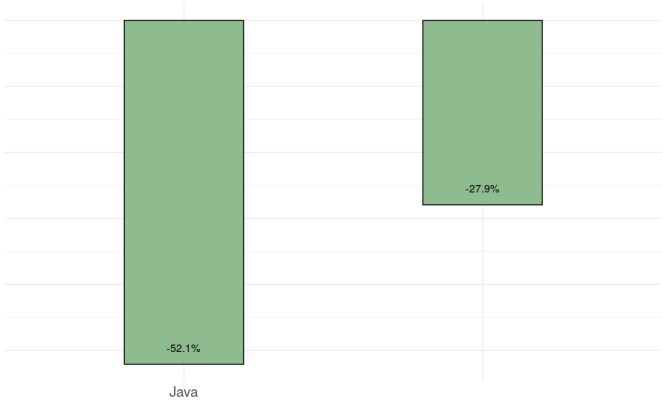


SQL

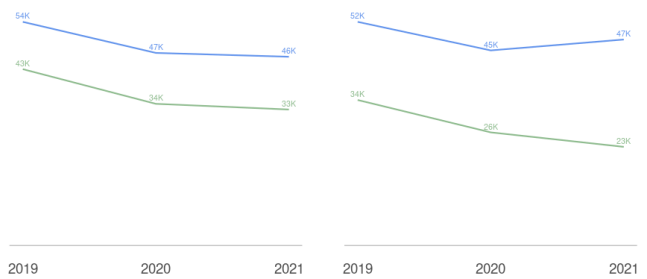


Decrease in Women's Average Compensation by Java; Bar Graph: 2021, Line Graph: 2019 - 2021

Based on 2019 - 2021 survey data; Compared to Men's Average Compensation

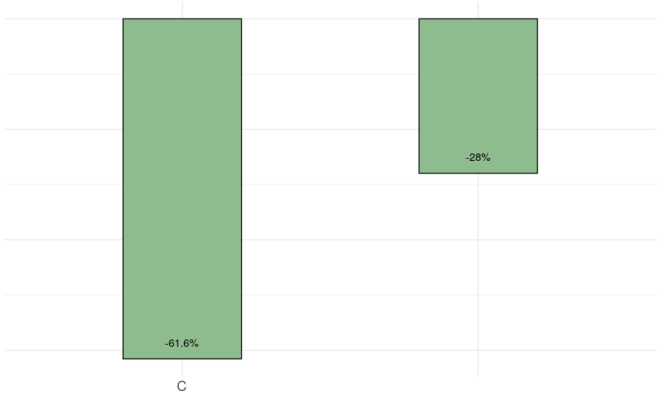


Java

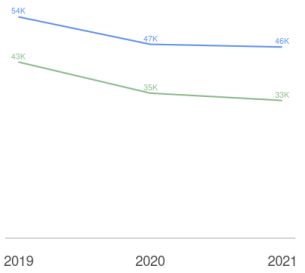


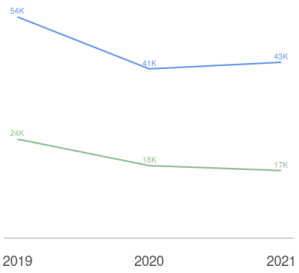
Decrease in Women's Average Compensation by C; Bar Graph: 2021, Line Graph: 2019 - 2021

Based on 2019 - 2021 survey data; Compared to Men's Average Compensation



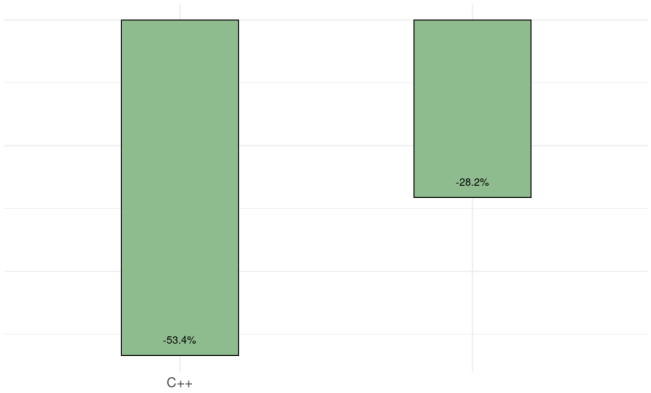
C



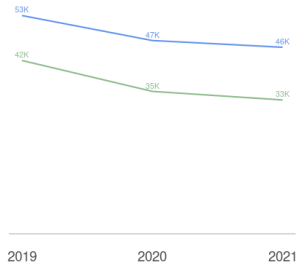


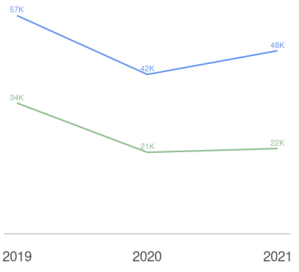
Decrease in Women's Average Compensation by C++; Bar Graph: 2021, Line Graph: 2019 - 2021

Based on 2019 - 2021 survey data; Compared to Men's Average Compensation



C++

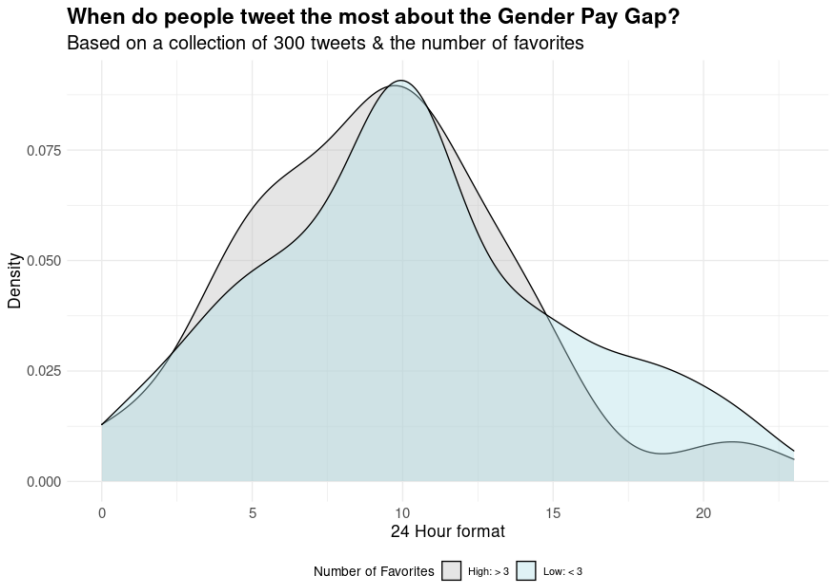


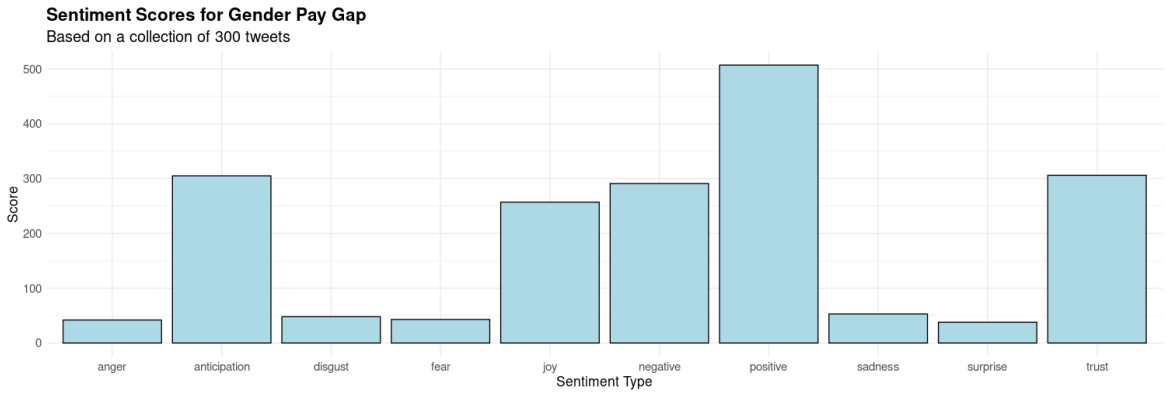


Sentiment Analysis

The purpose of sentiment analysis is to analyze the engagement of social media users, particularly Twitter users, towards the topic of gender pay gap. To achieve the same, the latest 300 tweets on the given topic were scraped using RStudio.

The density plot displays the times when most tweets on the topic are favorited or liked by users. **In the morning and afternoon, tweets on the topic of gender pay gap get more favorites**. It can be concluded that **users actively discuss the topic during earlier hours of the day**. The sentiment scores highlight the prevalence of each type of sentiment from the scraped tweets.





Machine Learning: Feature Importance

The purpose of feature importance is to rank the analyzed variables from most to least important for predicting the target variable. The target variable in the chosen dataset is the classified Compensation Group. The variables or the features in focus are Age, Gender, Industry, Education Level, Job Title, Country of Residence and Programming Experience in Years.

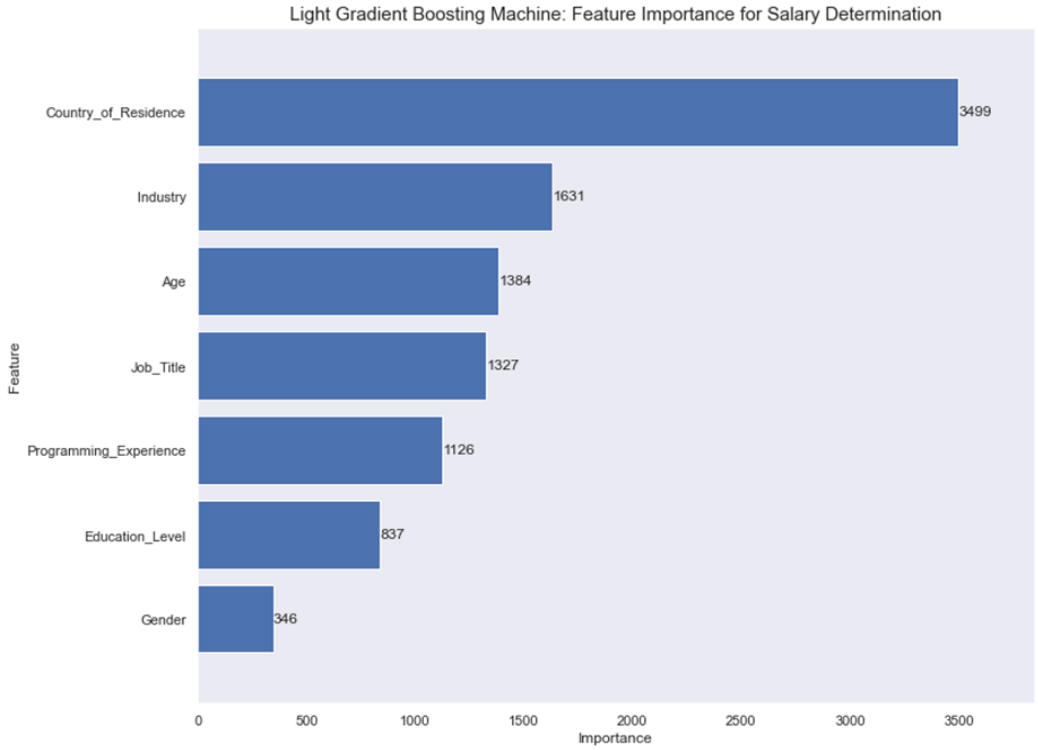
Light Gradient Boosting Machine and Random Forest were implemented on the chosen dataset to rank the features from most to least important.

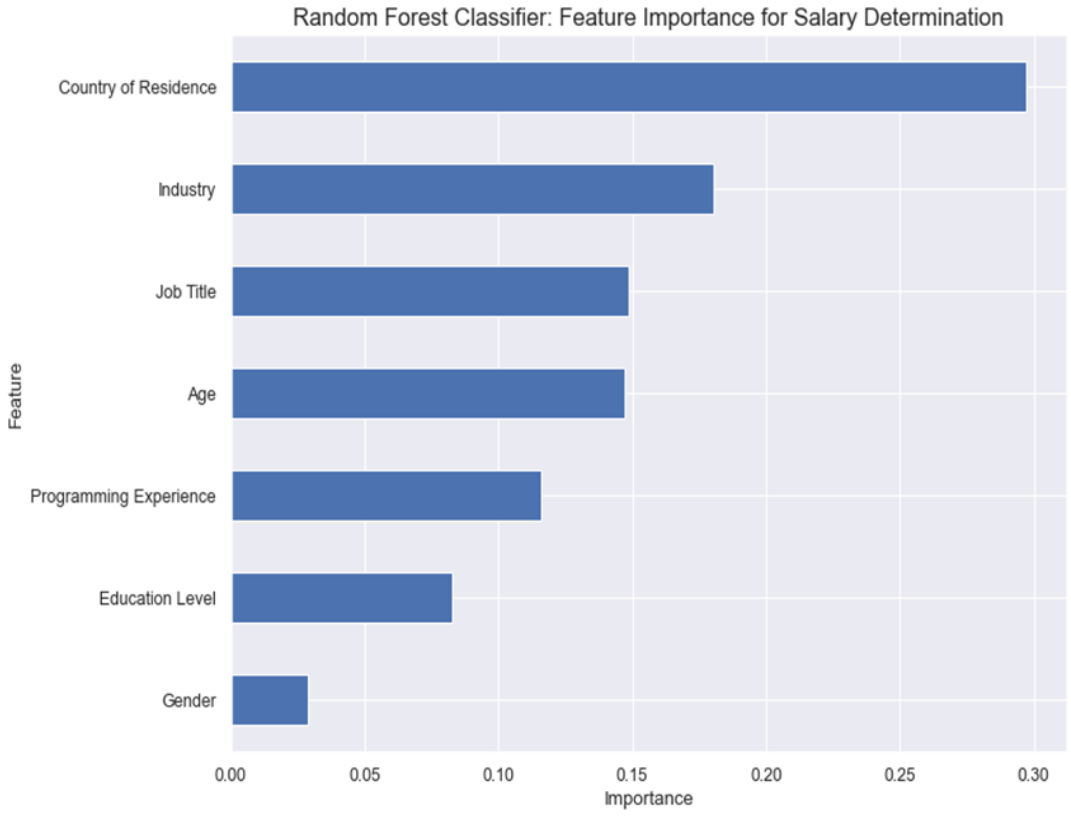
Light Gradient Boosting Machine Classifier

Based on the implementation and the obtained plot (below), it can be seen that **Country of Residence is the most important feature in the dataset. Industry ranks second.** Gender is the least important feature in the dataset. The x-axis on the plot indicates the frequency measure, which is the number of times a feature was splitted in trees.

Random Forest Classifier

Based on the implementation and the obtained plot (below), it can be seen again that **Country of Residence is the most important feature in the dataset. Industry ranks second**. Gender is the least important feature. The x-axis on the plot indicates the importance of each feature in the order in which the features are arranged in the training dataset for the prediction of the target variable.


Light Gradient Boosting Machine: Feature Importance for Salary Determination
Country_of_Residence
Industry
Age
Job_Title
Programming_Experience
Education_Level
Gender
3499
1631
1384
1327
1126
837
346
Feature
0
500
1000
1500
2000
2500
3000
3500
Importance


Random Forest Classifier: Feature Importance for Salary Determination
Country of Residence
Industry
Job Title
Age
Programming Experience
Education Level
Gender
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance

Machine Learning: Salary Prediction

The purpose of using machine learning techniques is to check the accuracy of the dataset for predicting the target variable. In the chosen dataset, the target variable is respondents' compensation group. For the purpose of this study, three models were applied to the pre-processed 2021 Data Science and Machine Learning survey dataset - K-Nearest Neighbors, Random Forest Classifier and Light Gradient Boosting Machine Classifier.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) algorithm is a supervised machine learning technique and "a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to" (*G2*). KNN works by "calculating the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label in the case of classification" (Harrison, 2018). The purpose of this model in the given study is to predict a salary group class, 0 to 6, for a respondent.

Random Forest Classifier

Random Forest Classifier is a supervised machine learning algorithm that builds decision trees on different samples and takes their majority vote for classification purposes (*Analytics Vidhya*, 2021). For classification purposes, the algorithm takes n number of random records from a data set having k number of records. Then, individual decision trees are constructed for each sample. Each decision tree generates an output

and the final output is based on majority voting or averaging (*Analytics Vidhya*, 2021). Random forests are also reported to be more efficient and effective for predictions as compared to decision trees. The purpose of the Random Forest Classifier in the given study is to predict a salary group class, 0 to 6, for a survey respondent.

Light Gradient Boosting Machine Classifier

Light Gradient Boosting Machine is a gradient boosting framework based on decision trees and is based on two types of techniques - Gradient Based on Side Sampling or GOSS and Exclusive Feature Bundling or EFB. It is proven to have a higher accuracy, a faster training speed, low memory utilization and handles overfitting much better while working with smaller datasets (*Analytics Vidhya*, 2021).

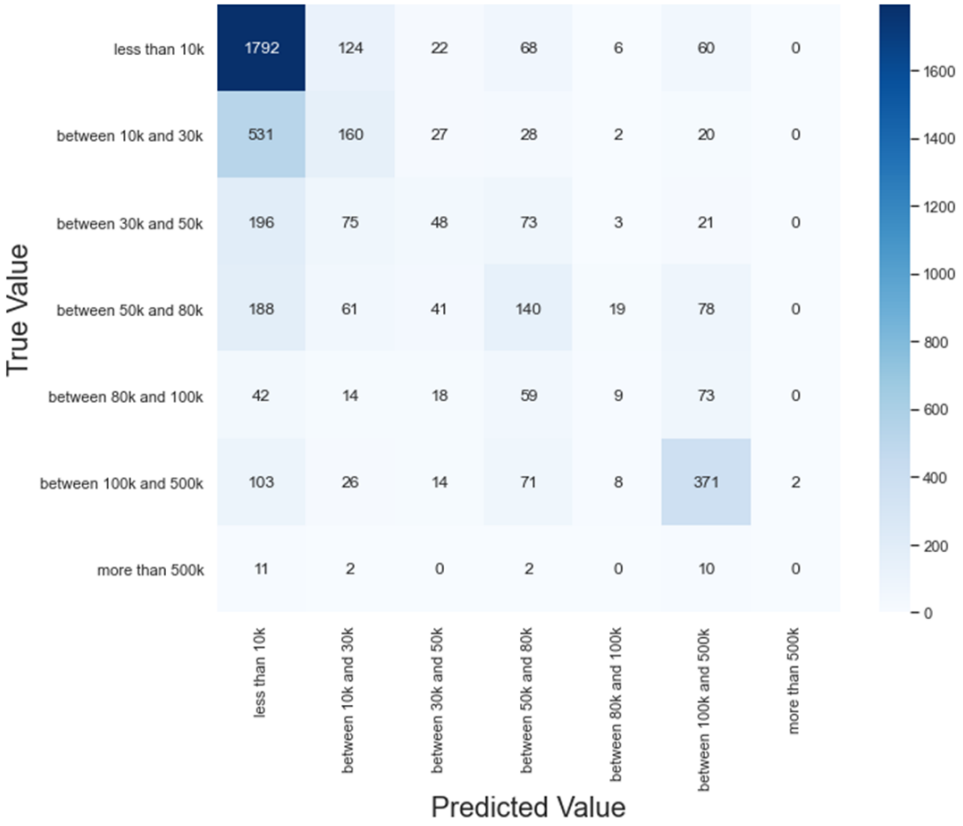
GOSS excludes a significant proportion of data instances with small gradients, and only uses the rest to estimate the information gain. Since the data instances with larger gradients play a more important role in the computation of information gain, GOSS can obtain quite accurate estimation of the information gain with a much smaller data size. EFB bundles mutually exclusive features (i.e., they rarely take nonzero values simultaneously), to reduce the number of features (Ke et. al, 2017).

Results: Accuracy Scores & Confusion Matrices

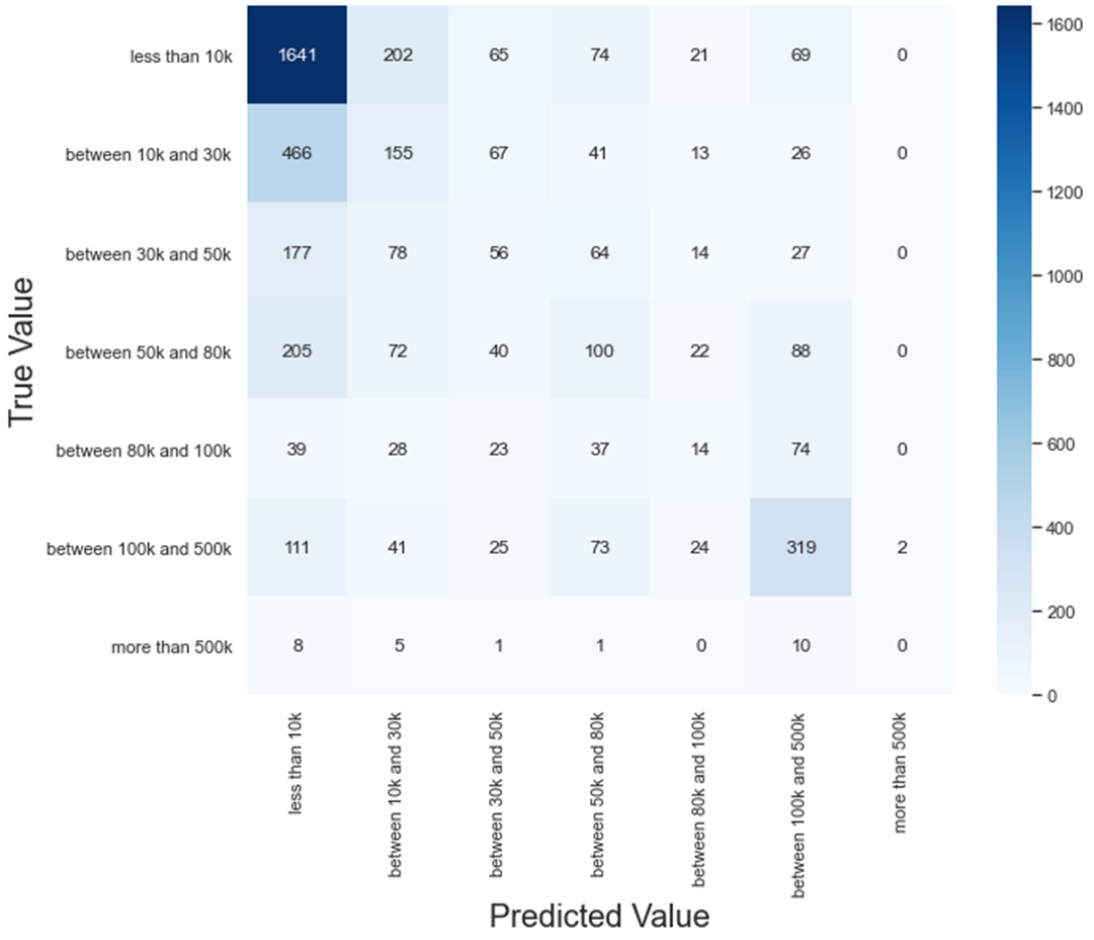
The three models were implemented on the 70% training set. Then, they were tested on the 30% testing set. The following accuracy scores were obtained on the testing set. Based on the given scores, it can be seen that the Light Gradient Boosting Machine Classifier obtained the highest accuracy rate of approximately 54.56%.

	Model Accuracy Scores on Testing Data
Light Gradient Boosting Machine Classifier	0.545691
Random Forest Classifier	0.494803
K-Nearest Neighbors Classifier	0.462971

For further analysis, confusion matrices were generated for each of the applied models. Based on the given confusion matrix generated via the Light Gradient Boosting Machine Classifier, it can be concluded that the model was able to predict the salary or compensation class, less than 10k, for 1792 respondents correctly. However, it was able to predict the compensation class, between 100k and 500k, for only 371 respondents correctly.

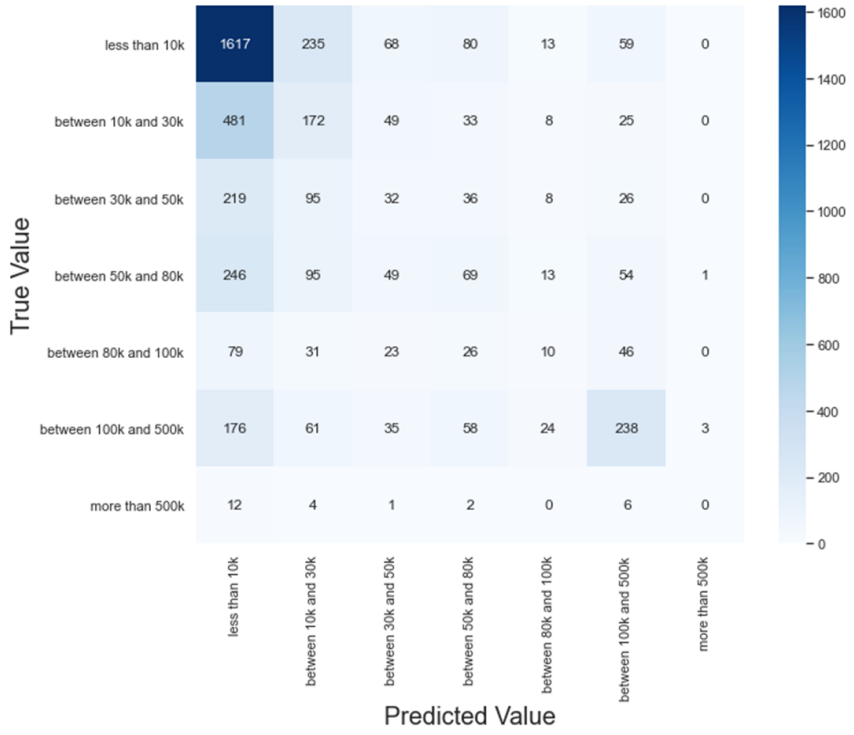


Based on the given confusion matrix generated via the Random Forest Classifier, it can be seen that the model was able to predict the salary or compensation class, less than 10k, of 1641 respondents correctly whereas it was able to predict the compensation class, between 100k and 500k, for only 319 respondents correctly.



Lastly, the confusion matrix generated from the K-Nearest Neighbors model reveals that the model was able to predict the salary or compensation class, less than 10k, of 1617 respondents correctly whereas it was able to predict the compensation class, between 100k and 500k, for only 238 respondents correctly.

The confusion matrices prove that the Light Gradient Boosting Machine Classifier is more efficient and effective than the Random Forest Classifier and K-Nearest Neighbors Classifier models.



Hyperparameter Tuning

The purpose of hyperparameter tuning is to choose an optimal set of parameters for an algorithm. The tuning process was conducted on the three models to improve the accuracy rates.

Grid Search Cross Validation was performed on the three models. Cross Validation is defined as “the process of training learners using one set of data and testing it using a different set” (Albon, 2017). Grid Search “evaluates all the

combinations from a list of desired hyper-parameters and reports which combination has the best accuracy" (Norena, 2018). When Grid Search CV is implemented on the classifier model, the performance of the selected parameters is evaluated using cross-validation and the performance of the model is optimized by "nested" cross-validation (*Scikit Learn*).

Once Grid Search CV was implemented, the models were trained on the 70% training set with the obtained optimal parameters. Then, the models were tested on the 30% testing set and the following accuracy scores were obtained. Based on the results, it can be concluded that the Light Gradient Boosting Machine Classifier provided the highest accuracy rate of approximately 55.34%. Through hyperparameter tuning, the accuracy scores of the three models observed a slight improvement.

	Model Accuracy Scores on Testing Data
Light Gradient Boosting Machine Classifier	0.553486
Random Forest Classifier	0.522304
K-Nearest Neighbors Classifier	0.498051

CONCLUSION

Insights

Based on the conducted analysis, the following can be concluded -

1. A significant gender pay gap does exist in the field of Data Science and Machine Learning, across all the examined factors - Age, Education Level, Job Title, Programming Experience in Years, Country of Residence.
2. Study shows that though women have successfully closed the gap at entry level positions, men continue to dominate senior positions with higher pay. This corroborates with previously conducted research as reported in the literature review section of this report.
3. Women tend to be more qualified than men. This corroborates with previously conducted research that women tend to have more educational degrees than their male counterparts in the field.
4. Study shows that gender is not the most important feature for salary prediction purposes, in the given dataset.
5. Lastly, study corroborates with conducted research that classifier models tend to produce better results to predict compensation groups in survey data.

Project Limitations

Through this study, the following key limitations can be identified -

1. The analyzed data is sparse and fragmented. Survey data is prone to human error. Since a large number of respondents are male, the results can be biased.
2. Compensation information does not account for inflation or purchasing parity. For better results, it is necessary to account for a country's gross domestic product and inflation.

Future Research

This research can be advanced by accounting for countries' gross domestic product and inflation rates in classification models. Additionally, enhanced social media analysis of professional networking websites such as LinkedIn and Untapped can reveal further insights on the potential reasons for a pay gap between males and females in the field of Data Science and Machine Learning. Lastly, further research should be conducted to explore the reasons why women are drawn more towards entry-level positions in the field, despite being more educationally qualified than their male counterparts.

[1] Github Repository

[1] https://github.com/arushik1994/DATS-6501---Data-Science-Capstone---Arushi-Kapoor

REFERENCES

- *1.1.1. What Is Eda?,* NIST, https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm.
- *2019 Kaggle Machine Learning & Data Science Survey,* Kaggle LLC*,* https://www.kaggle.com/c/kaggle-survey-2019/code.
- *2020 Kaggle Machine Learning & Data Science Survey,* Kaggle LLC*,* https://www.kaggle.com/c/kaggle-survey-2020/code.
- *2021 Kaggle Machine Learning & Data Science Survey,* Kaggle LLC*,* https://www.kaggle.com/c/kaggle-survey-2021/code.
- *2019 Kaggle Machine Learning & Data Science Survey*, Kaggle LLC, 2019, https://www.kaggle.com/c/kaggle-survey-2019/data.
- *2020 Kaggle Machine Learning & Data Science Survey*, Kaggle LLC, 2020, https://www.kaggle.com/c/kaggle-survey-2020/data.
- *2021 Kaggle Machine Learning & Data Science Survey*, Kaggle, Nov. 2021, https://www.kaggle.com/c/kaggle-survey-2021.
- "3.1. Cross-Validation: Evaluating Estimator Performance." *Scikit Learn*, Scikit-Learn Developers, https://scikit-learn.org/stable/modules/cross_validation.html.
- Albon, Chris. "Cross Validation with Parameter Tuning Using Grid Search." *Chris Albon*, Github, 20 Dec. 2017, https://chrisalbon.com/code/machine_learning/model_evaluation/cross_validation_parameter_tuning_grid_search/.

- Bekena, Sisay Menji. “Using Decision Tree Classifier to Predict Income Levels.” *Munich Personal RePEc Archive*, University Library LMU Munich, 30 July 2017, https://mpra.ub.uni-muenchen.de/83406/.
- *Data Preparation - Machine Learning Lens - Docs.aws.amazon.com*. Amazon Web Services, Inc. , https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/data-preparation.html.
- DeepAI. “Accuracy (Error Rate).” *DeepAI*, DeepAI, 17 May 2019, https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate#:~:text=Accuracy%20in%20Machine%20Learning&text=Accuracy%20is%20the%20number%20of,false%20positives%2C%20and%20false%20negatives.
- “Earnings and Wages - Gender Wage Gap - OECD Data.” *Gender Wage Gap*, OECD, https://data.oecd.org/earnwage/gender-wage-gap.htm.
- *Global Gender Gap Report 2021*. World Economic Forum, Mar. 2021, https://www3.weforum.org/docs/WEF_GGGR_2021.pdf.
- Joby, Amal. “What Is Training Data? How It's Used in Machine Learning.” *Learn Hub*, G2, 30 Apr. 2021, https://learn.g2.com/training-data#training-test-validation-data.
- “Kaggle-Survey-Salary-Prediction/kaggle_salary_prediction.Ipynb at Master · Mustardn/Kaggle-Survey-Salary-Prediction.” *GitHub*, Github, 1 Apr. 2019, https://github.com/mustardn/Kaggle-Survey-Salary-Prediction/blob/master/Kaggle_Salary_Prediction.ipynb.

- Ke, Guolin, et al. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. 31st Conference on Neural Information Processing Systems , 2017, https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
- Norena, Sebastian. “Python Model Tuning Methods Using Cross Validation and Grid Search.” *Medium*, Medium, 15 June 2018, https://medium.com/@sebastiannorena/some-model-tuning-methods-bfef3e6544f0.
- Patwardhan, Sai. “KNN Algorithm: What Is KNN Algorithm: How Does Knn Function.” *Analytics Vidhya*, Analytics Vidhya, 21 Apr. 2021, https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/.
- Raiber, Klara. *Can Gender Role Items Improve the Prediction of Income? Insight from Machine Learning*. July, 2019, https://www.europeansurveyresearch.org/conf2019/uploads/142/2901/89/Raiber_GRI_in_income_prediction.pdf.
- “Sklearn.ensemble.randomforestclassifier.” *Scikit Learn*, Scikit-Learn Developers, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
- Sterling, Adina D., et al. *The Confidence Gap Predicts the Gender Pay Gap Among STEM Graduates*. Proceedings of the National Academy of Sciences United States of America, 16 Nov. 2020, https://www.pnas.org/doi/10.1073/pnas.2010269117.

- Viel, Theo. “Kagglers' Gender Pay Gap & Salary Prediction.” *Kaggle*, Kaggle, 5 Dec. 2018, https://www.kaggle.com/code/theoviel/kagglers-gender-pay-gap-salary-prediction/notebook.
- “Welcome to LIGHTGBM's Documentation!.” *Welcome to LightGBM's Documentation! - LightGBM 3.3.2.99 Documentation*, Microsoft Corporation, https://lightgbm.readthedocs.io/en/latest/.
- “What Is Data Visualization? Definition, Examples, and Learning Resources.” *Tableau*, https://www.tableau.com/learn/articles/data-visualization.
- Wiggers, Kyle. “Data Science Hasn't Fixed Its Huge Gender Pay Gap.” *VentureBeat*, VentureBeat, 15 Sept. 2021, https://venturebeat.com/2021/09/14/data-science-hasnt-fixed-its-huge-gender-pay-gap/.
- Young, Erin, et al. *Where Are the Women? - Alan Turing Institute*. The Alan Turing Institute, 2021, https://www.turing.ac.uk/sites/default/files/2021-03/where-are-the-women_public-policy_full-report.pdf.