Investigating the Gender Pay Gap and its Implications

By Trinity Martinez and Theo Covich

University of Washington
CSE163 WIN22 Section AG

I. Research Questions

1. **Are there occupations in which males make more than females? Are there occupations in which females make more than males? If so, what are they?**
    - This question will investigate the occupations listed in the dataset and the pay disparities between men and women. The result will demonstrate the discrepancy between gender in a specific occupation.
    - **Answer:** Software engineers and managers.
2. **How does the gender pay gap differ across degree levels achieved for each gender?**
    - This question investigates whether there might be a slight chance to close the gender pay gap by giving females access to higher education.
    - **Answer:** the gender pay gap still exists despite the educational attainment level for women. Pay is indeed lower for females than males
3. **How do attributes other than gender (such as race, ethnicity, socio-economic status) contribute to the pay gap? If so, what combination of attributes result in the largest disparity?**
    - This question will investigate whether the disparity amongst individuals in the same occupation has more to do with outside attributes other than gender. This includes race, ethnicity, geographic location within the United States, and other influential factors
    - **Answer:** Race is the most significant factor when determining pay across multiple attributes (other than sex).
4. **If a gender pay gap does exist within certain occupations, is it possible to predict the pay of individuals based on certain combinations of attributes?**
    - This question focuses more on being able to accurately predict whether or not an individual will experience a pay differential based on their characteristics.
    - **Answer:** Yes, we found that certain attributes performed much stronger on predictive models than others. In particular, we found that that degree level attainment was the leading predictor of income followed by gender, race, and region.

II. Motivation

As students who identify as ethnic minorities, and one of us identifying as female, we feel that it is important to raise awareness to this issue and bring attention to a very serious problem in our industry. Women are generally paid less than men in almost every occupation. We care about understanding the gender wage gap because it has proven to be a systematic problem that's brought about real harm to society. Salary goes hand in hand with individual livelihood and so it is something that deserves to be analyzed in great detail.Our main motivation is equal pay for equal work. As we slowly make our way into the workforce (undoubtedly in technology or other STEM field), both of us may very well experience some type of marginalization or discrimination. Whether that be social or economic like how we are investigating this project, we

aim to give voice to this controversy in hopes to eventually solve the problem within our lifetime. Knowing the answers to the research questions that we proposed, we will be one step closer to solving the issue and we may even be able to bring this knowledge into the workforce and our workplaces.

III. Dataset

The dataset that we are using was found on Kaggle and is called "The Gender Wage Gap: Extent, Trends, and Explanations." The link can be found here: <https://www.kaggle.com/fedesoriano/gender-pay-gap-dataset> This dataset is comprised of two tables, however we are choosing to focus on the PSID table because we feel that the information is more relatable to the research questions we are trying to answer. The PSID table includes additional United States data on the gender pay gap.

Additionally, we used the "Glassdoor - Analyze Gender Pay Gap" dataset to provide some supplemental information about base pay and other attributes such as gender, age, highest schooling completed, and salary bonuses.
<https://www.kaggle.com/nilimajauhari/glassdoor-analyze-gender-pay-gap>

IV. Method

To begin data analysis we will first start off by cleaning the datasets found from Kaggle. This will include using the pandas library to specify columns of interest in the PSID dataset (Panel Study of Income Dynamics), since there are over 250 different columns. An example of columns of interest include: sex, age, sch, annlabinc, degree, white, black, hisp, yrsexp, and other columns related to occupations and industries. Additionally we will drop all rows with any NaN values so as to not hinder the results of our analysis. With this new dataset that only includes information that is directly related to our research questions, we will be able to perform a more accurate analysis with a clean set of data. Our second dataset with information from Glassdoor was clean to begin with no Nan values, however we filtered only for the columns that we were interested in such as Gender, BasePay, Education, and Bonus. We omitted the columns of department and seniority. Once we've cleaned our datasets we can focus on graphing and visualizing the data.

For research question #1, we aim to focus our attention on establishing which occupations (if any) result in the largest pay gap, or have any pay gap at all. This will be achieved by representing the data visually from the Glassdoor dataset on a grouped bar graph with the columns JobTitle on the X axis, and BasePay on the Y axis. We also hope to answer this question by graphing the Department column vs Base Pay column in another grouped bar chart. Both grouped bar charts will be grouped by the column Gender because we are investigating whether there is a disparity across different genders. All graphs will be rendered using the Altair Python library and this connects to our first challenge goal where we aim to learn a new Python

library. The result of this graph will lead to the conclusion of which occupations and departments have the highest pay disparity.

For research question #2, we will try to answer the question of how degree levels achieved affect salary based on gender. For this question we will use the PSID dataset and focus our attention on the sex, annlabinc, and degree columns. Once the data is properly filtered down to these three specific columns, we'll use the altair Python library to visualize the data on a grouped bar chart. Annual Income (annlabinc column) will be plotted on the Y axis, and Degree level (degree) will be plotted on the X axis. The graph will then be grouped by Gender (sex column). Additionally we will use the Glassdoor dataset to render another grouped bar chart for degrees and annual income. We focus our attention on the Education, Basepay, and Gender columns.

Question # 3 leads us to one of the most important questions to be asked when examining the gender pay gap. Here we are curious to uncover the combination of demographic attributes that lead to the most disparity in pay. First we will use the PSID dataset and focus on the columns: sex, annlabinc, white, black, hisp, yrsexp, and age. The first visualization we will produce will focus on combining the Race attributes. For each race we will plot a bar chart grouped on Gender with Annual Income on the Y axis and Race on the X. Next, we will plot a grouped bar chart based on gender to investigate if disparities exist based on the age of individuals. Annual income will continue to be on the Y axis and Age will be plotted on the X axis. Lastly, we will take into account Annual income and years of experience where annual income is plotted on the Y axis, and gender will be plotted on the X axis. For this visualization we will be using a heat map where the colors will be based on the amount of experience an individual has.

Our last research question is based around machine learning models that were created to try and predict income based on a variety of factors. We used two separate regression models when fitting the dataset and reported each model's effectiveness against that certain task. The two models used were DecisionTreeRegressor and RandomForestRegressor. The methods that we included print the mean error of predicted income to actual income based on specified variables and include the model used to compute it. This will show us how well the individual model performs on that data subset, as well as the weight of each variable on income data. The variables tested as predictors for income were gender, degree level, race, and region. In order to ensure that we did not underfit or overfit over data, we used 5-fold cross validation so that our model can perform training and testing amongst all of the data, instead of a train-test-split.

V. Results

A. Background Context

The base of our investigation was whether or not a gender pay gap exists. To provide some background context of our problem space, we created the following visualizations using Altair. Our first figure 1.1 shows a pie chart displaying the annual income per gender, where male individuals are represented in the color blue, and female individuals are represented in the

color orange. From figure 1.1 we can clearly see that male individuals are earning more than female individuals.

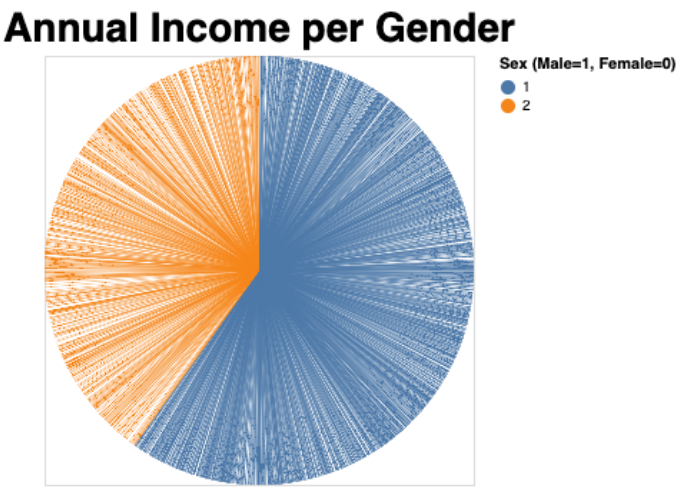


Figure 1.1: Annual Income per Gender

To provide more context we also included a visualization to show Bonus and Base pay per gender. Figure 1.2 shows a clear disparity in pay with the base pay income for males being higher than females by about $20,000 USD. Figure 1.3 shows a slight disparity in bonus pay, being of higher value for males than females. Despite the slightly higher value for males, there still exists a prominent and noticeable gap in both base pay and salary bonus amount.

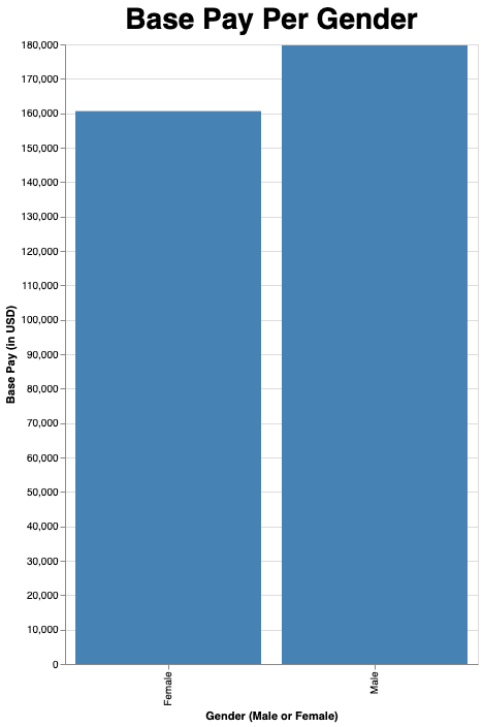


Figure 1.2: Base Pay per Gender

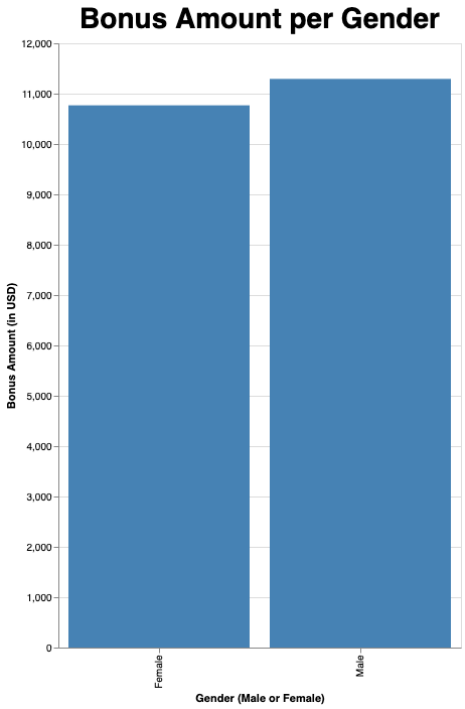


Figure 1.3: Bonus Amount per Gender

B. Are there occupations in which males make more than females? Are there occupations in which females make more than males? If so, what are they?

From using Altair to produce data visualizations we were able to plot different occupational data of interest for male and female individuals and relate them to pay. Here we first focused on plotting the department that the occupation falls under: Administration, Engineering, Management, Operations, and Sales. Figure 2.1 shows the department for each gender and how much they make.

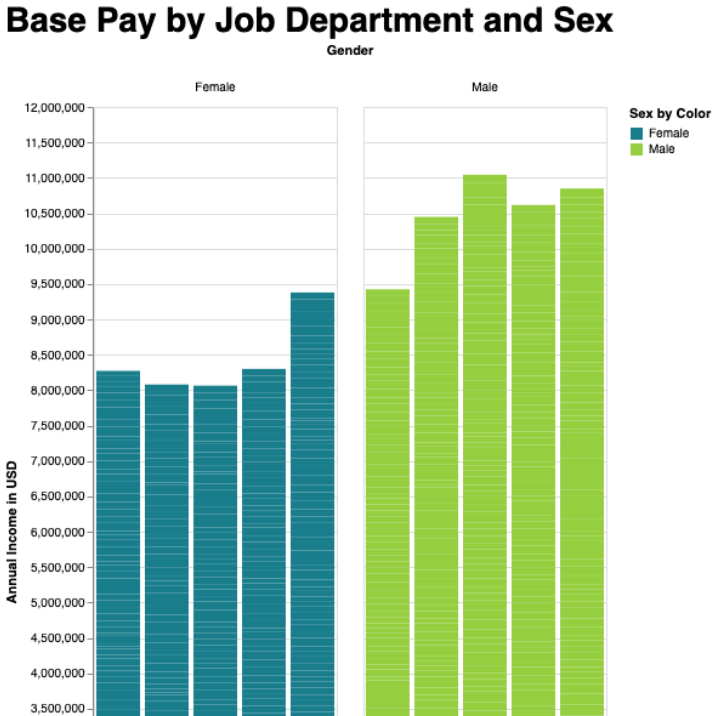


Figure 2.1: Base Pay by Job Department and Sex

From the graph it is clear that males (represented in lime green) are paid more than females (represented in teal) for all departments. Regardless of department type, males are always paid higher than females. One might argue that the definition of the job department is much too broad to compare pay, however Figure 2.2 displays the job category and dives deeper into which specific categories have the highest discrepancy.

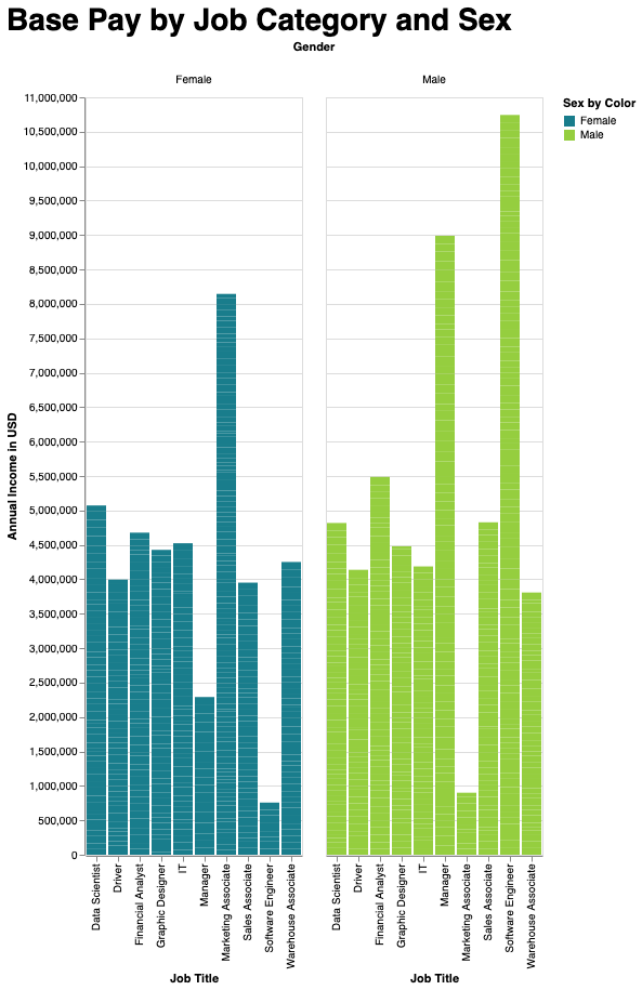


Figure 2.2: Base Pay by Job Category and Gender

To recap, males earn more than females in every job department, and almost every Job Category. One key takeaway here is to compare the female and male pay for the job category of Software Engineer. Here we see the *largest* gap in pay differential. Despite the fact that male and female individuals are performing the same job and in the same career, males are paid 45% higher than women. Another key takeaway from the data is that the job category of Manager produces another extremely high discrepancy. On the converse, it seems that women are paid more than men in the job category of Data Scientist which is an intriguing takeaway since Software

Engineer and Data Scientist could be thought of to be in the same industry (STEM fields). Despite female and male individuals being in the same job category, department, and performing similar duties in their careers, females are being paid less than males.

C. How does the gender pay gap differ across degree levels achieved for each gender?

Moving onto our next research question, we investigate how the gender pay gap differs across degree levels achieved. If in fact there is another reason behind the pay gap being so large within occupational categories, surely education plays an important role. Figure 3.1 illustrates the Annual Income of individuals based on Degree level obtained (either no college degree, bachelor's degree, or an advanced degree).

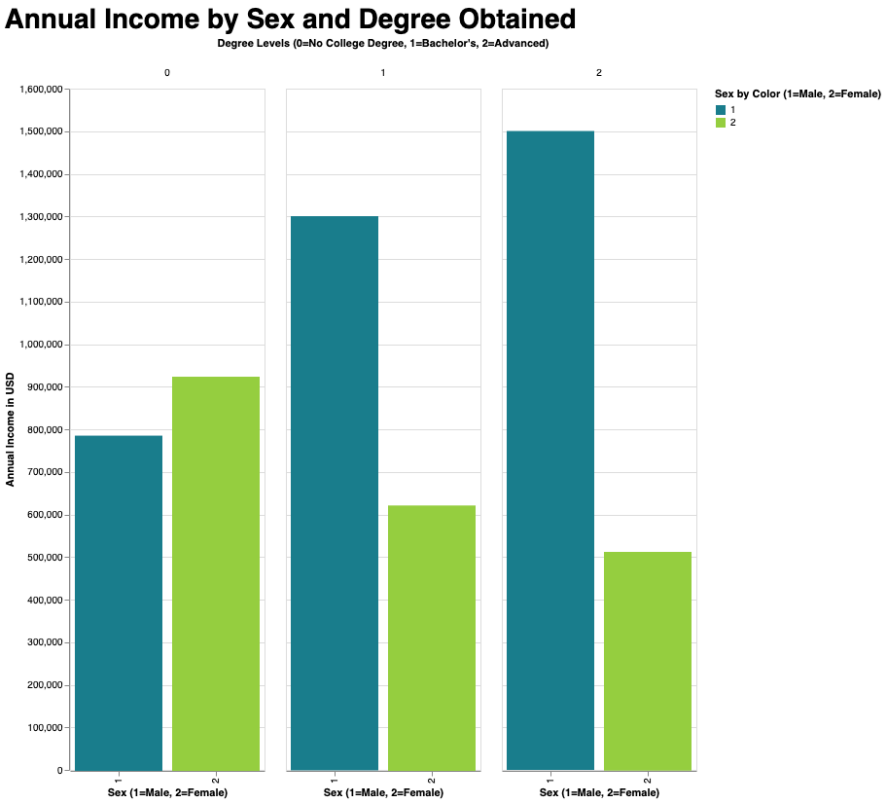


Figure 3.1: Annual Income by Sex and Degree Level Obtained

For columns based on Bachelor's and Advanced Degrees, Males are being paid exponentially higher than females despite attaining the same education level. An interesting notion presented in this chart is how females are actually paid more than males who have no college degree. Figure 3.2 focuses on College, High school, Masters, and PhD degrees. The data here supports our claim again that males earn more than females despite attaining the same higher education degrees. The largest pay gap here is in the Master's column. Women are now attaining higher

levels of education in order to ‘keep up’ with their male counterparts to earn a competitive wage. However, the Master’s degree column is one of the largest disparities in pay. This shows that despite the education level of women, they will still be underpaid compared to men.

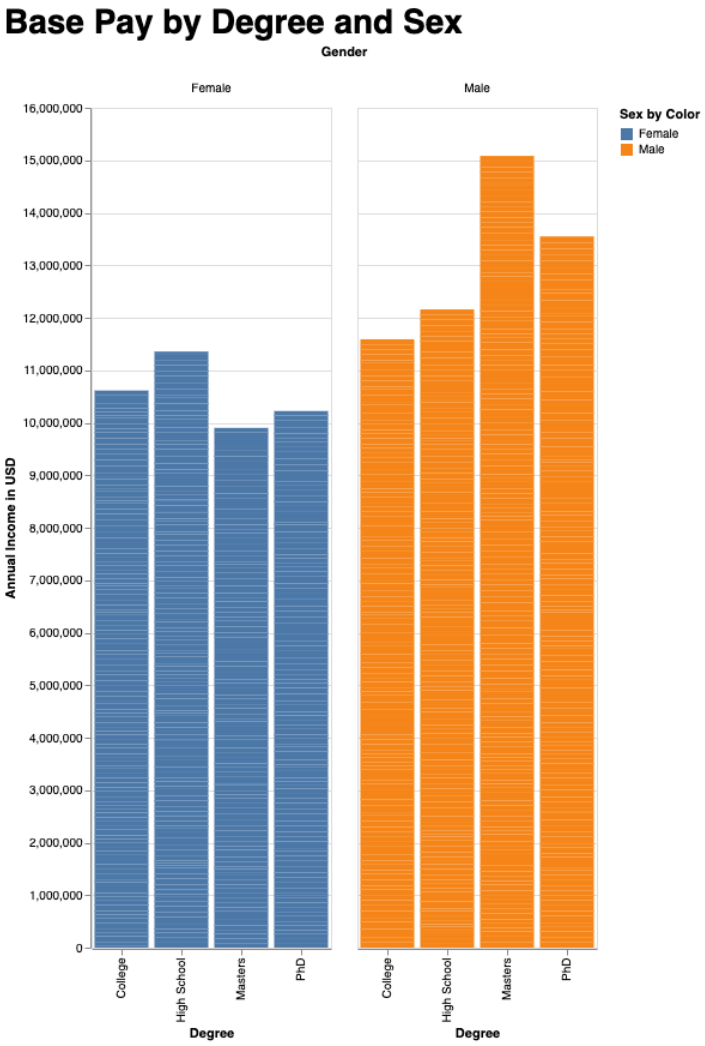


Figure 3.2: Base Pay by Degree and Sex

D. How do attributes other than gender (such as race, ethnicity, socio-economic status) contribute to the pay gap? If so, what combination of attributes result in the largest disparity?

Lastly, in order to discover which attributes and combination of demographic data result in the worst pay gap, we chose to focus on the following features: Race, Age, and Years of Experience in a specific career. We first compare Race (Hispanic, Black, and White). Figure 4.1.1 uncovers the fact that if an individual is classified as Hispanic, they are severely underpaid as compared to

non-Hispanic identifying individuals. Something intriguing from this graph is that Hispanic females are actually paid more than Hispanic males. This might be due to the fact that there are discrepancies in the dataset. If we look at Black and Annual Income we also see both male and female individuals identifying as Black are also severely underpaid compared to white males and females. If we compare Black and Hispanic wages, Hispanic people perform better. So to come to a conclusion, an individual identifying as Black and female will make about ⅙ of what a white male will earn. This is a huge disparity and one of our key major findings from this project.

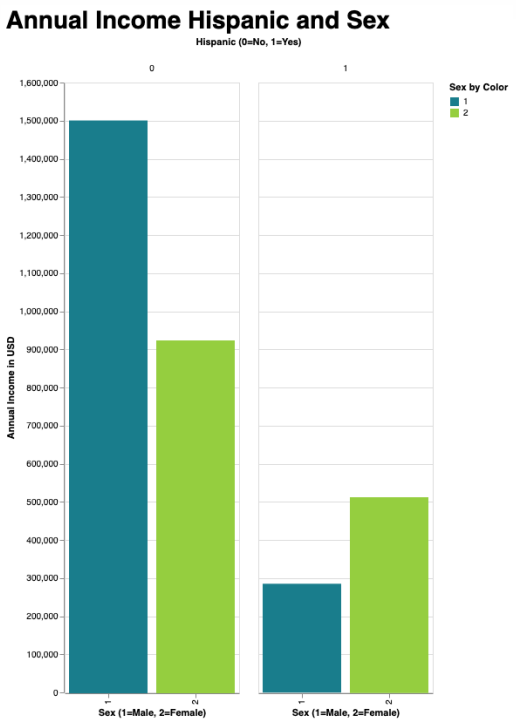


Figure 4.1.1: Annual Income Hispanic

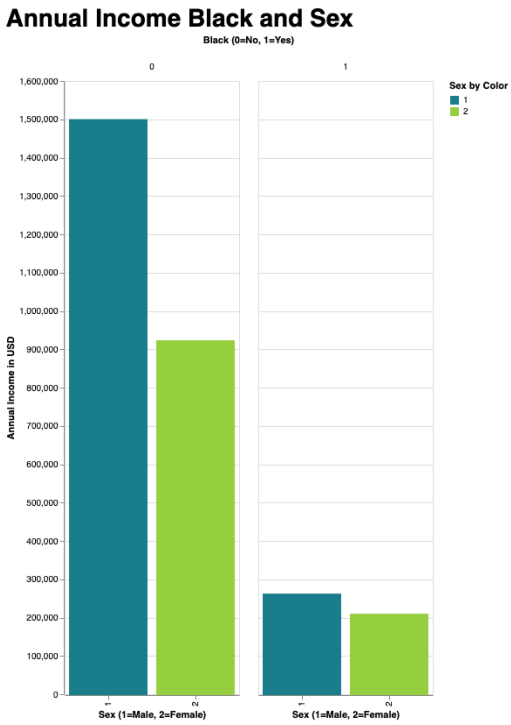
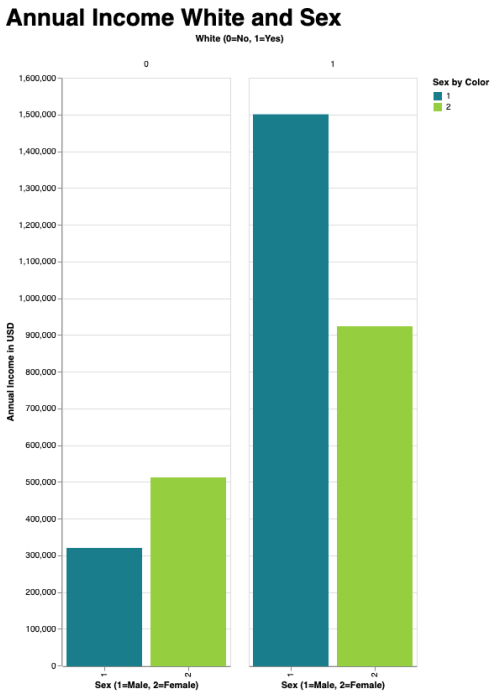


Figure 4.1.2: Annual Income Black



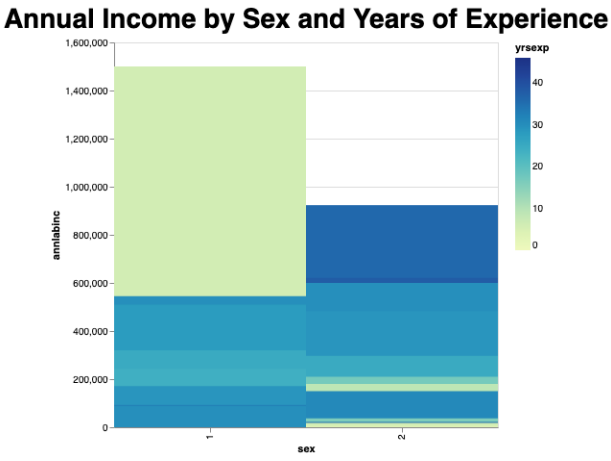


Figure 4.1.4: Annual Income and Years of Experience

Figure 4.1.3: Annual Income White

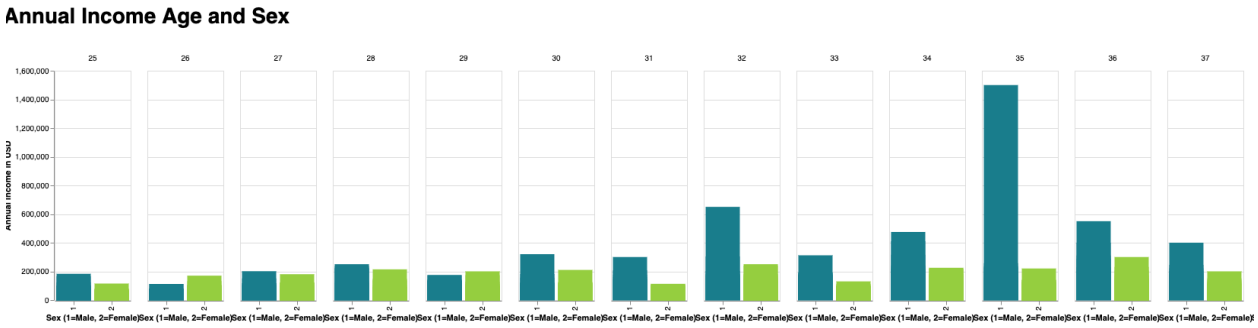


Figure 4.1.5: Annual Income Age and Sex

In addition to investigating race as a potential contribution to a worsening pay gap, we also chose to focus on income based on years of experience. Surely there won't be any pay differential if both male and female individuals have the same number of years of experience. However our assumption was disproved. Figure 4.1.4 displays a heat map of gender and annual income color coded based on years of experience. This was one of the most interesting plots that we produced because it visualizes what is called "the glass ceiling." The darkest blue hue represents the max number of years of experience, and light green represents the low number of years of experience. It is evident that despite the fact that women hold more years of experience than men, and men hold a lower number of years of experience, male individuals are still being paid more. "The phrase 'glass ceiling' was first used in 1984 by Gay Bryant, who was the editor of Working Woman magazine. In that profile, she was quoted as saying, 'Women have reached a certain point—I call it the glass ceiling … in the top of middle management and they're stopping and getting stuck.'" (Boyd). As this notion started to become more prominent, th Department of Labor constructed a working definition, stating that, "a glass ceiling is made up of 'artificial barriers based on attitudinal or organizational bias that prevent qualified individuals from advancing upward in their organization into management-level positions.'" (Boyd). This is exactly what we are seeing in the heat map. There seems to be a cutoff point for women with 40+ years of experience of earnings just under one million USD. As compared to their male counterparts, men are earning upwards of 1.5 million USD despite having very little experience. The glass ceiling is still very much real for women in the workforce, so organizations and companies can focus their efforts on establishing equitable pay for equal work. Lastly we chose to focus on age demographics and displayed ages of male and female individuals 25-37. This graph shows a trend of higher annual earnings as age increases for *male* individuals only. If we

take a closer look at female individuals we see an almost stagnant, fairly stable trend of pay. The biggest disparity comes from the 35 age group where men are earning almost seven times of what females are earning. There could be other factors related to this such as women taking maternity leave to raise their children, which would put a pause on any annual income growth. Whereas males do not have to take maternity leave so their annual earnings would continue to grow based on their occupation and overall work life.
Putting all of the demographic factors together, we can conclude that the combination of attributes that result in the worst pay differential would be an individual who identifies as female, black, little to zero years of work experience, and who are 25 or 31 years of age (although the pay across age seems to be steady so we can assume than age doesn't have a huge impact on pay for women).

E. If a gender pay gap does exist within certain occupations, is it possible to predict the pay of individuals based on certain combinations of attributes?

After conducting our own research on some of the most well known predictors of income, we chose to include the gender, race, degree, and region data when making our machine learning models. In order to gauge the weight of each of these variables on predicted income, we used regression models on each of these variables separately. In addition to measuring the weight of the variables, we also wanted to measure the effectiveness of the models used (DecisionTreeRegressor and RandomForestRegressor) to see which afforded more accurate predictions over that data subset.

We found that there were certain variables that did end up being better predictors of income than others. The scores below illustrate the mean error from predicted income according to different variables passed and model used.

Variable: Gender
Model 1: DecisionTreeRegressor - 39065.58754379205
Model 2: RandomForestRegressor - 39063.643475386336

Variable: Race
Model 1: DecisionTreeRegressor - 39933.22167031672
Model 2: RandomForestRegressor - 39932.11780842886

Variable: Degree Level Attained (no college degree, bachelors, masters)
Model 1: DecisionTreeRegressor - 38296.88625725292
Model 2: RandomForestRegressor - 38294.49388332534

Variable: Region
Model 1: DecisionTreeRegressor - 40084.63190511616
Model 2: RandomForestRegressor - 40089.51212905871

Interpreting these numbers is the most important part in this section of the analysis. For one, the difference between the two model scores shows its individual effectiveness and should be interpreted as the lower score meaning a more effective model. We can observe here that the RandomForestRegressor model performed better on gender, race, and degree level attainment, while the DecisionTreeRegressor performed better on region data.

The output number here shows how far off the models predictions were on average from the actual income amounts. We can observe that the range of values here is around 38,000 to 40,000, which shows variance among the different variables. In order to interpret the weight of each variable on its predictive ability, we take the numbers in order from lowest to highest, which once again shows higher accuracy at lower levels. We see that degree level obtained results in the lowest values, followed by gender, race, and then region.

VI. Impact and Limitations

From our findings and investigation, it is clear that a gender pay gap does exist and that there are certain combinations of attributes that earn a significantly lesser amount than men. Our target audience who might benefit from our analysis are women in the workforce because our findings show a powerful discrepancy between male and female income. Not only is this unfair for women, but it should be a more prominently discussed topic in the industry. Concepts such as affirmative action do exist to eliminate discrimination during the hiring process, but what's to be said once the employee starts to work and earn income? Are there any concepts to prevent inequitable pay amongst males and females? These are some limitations of our study. Our analysis displays competent evidence that the gender pay gap exists, however we did not investigate tools and methods that companies can use in order to close this gap between their employees. Additionally, there may be potential biases in our research such as biases with data collection. Neither one of us actually conducted the survey and gathered the raw data. Instead we found the data off of Kaggle, a data science website that hosts thousands of different dataset topics. We did not do substantive research into each dataset we found. We could assume that the research contains biases such as selection bias, where some participants were excluded (ie. focusing on surveying more women because prior knowledge already assumes that a gender pay gap exists and researchers want to cater to women experiences). Relatably, data collection bias could be prominent, and this occurs when members of the population are excluded from the data during research because of data collection methods. An example of this would be asking individuals who don't have access to the internet to take an online survey. As stated before, there was limited research that went into discovering where the data we used came from which could be an important factor for determining credibility of our analysis.

VII. Challenge Goals

For our project we are selecting the following challenge goals: 1) Machine Learning and 2) Learning a new interactive data visualization library. Our first challenge goal is arguably the most important facet of our project because it relates to one of our research questions. We want to be able to predict whether or not an individual will experience a gender pay gap based on their demographic characteristics. This may include whether an individual identifies as female, ethnic minority, highest level of education, geographical location, occupation, and other factors. With our second challenge goal, we want to create a few data visualizations to further illustrate the narrative of how a gender pay gap has existed in the past and what factors can influence it. For example, we might create an interactive graph that showcases two lines (male and female) and the rate of achieving higher education over time. Other graphs might look like the rate of pay between genders over time, or a bar graph to show occupation and ethnicity between genders. These graphs will also help to answer our research questions and demonstrate to the audience the very serious issue of gender pay disparity. This second challenge goal was scaled back to just include learning a new data visualization library. We didn't take into account how we would present interactive plots in our report so we decided to just focus on making regular plots.

VIII. Work Plan Evaluation

Task 1: Complete Research Questions

- ~ 30 mins per question x 4 questions = ~ 2 hours

Task 2: Complete Challenge 1: Create Machine Learning Model

- ~ 2-3 hours

Task 3: Complete Challenge 2: Learn and Implement Altair Library

- ~ 3-4 hours

Task	Estimate	Evaluation
1. Answer Research Questions	30 minutes per question x 4 questions = 2 hours	This took much longer than 2 hours total for all questions. We had to learn Altair first and implement the library, then brainstorm data visualization plot ideas that were related to the columns of interest in the data we selected. After that we had to organize which plots answer which questions, perform a written analysis as well as code up the plots. Our estimate for this task was far from reality because we didn't know exactly what we

		wanted to plot to answer each question. WE had to evaluate the data set and filter through over 200 columns to specify the exact data we wanted to use.
2. Complete challenge 1. Learn and implement new machine learning concepts.	3 hours	This took a bit longer than the 3 hours that we were expecting. While we were able to complete this by using the RandomForestRegressor model and implementing the cross validating concept on our analyses, we found that it took a long time to understand the application of these concepts on our actual project and dataset. This was partially due to the depth of the concepts in conjunction with trying to analyze our data.
3. Complete challenge 2. Learn and implement Altair library.	4 hours	Learning and implementing the Altair library was much more difficult than anticipated but we were still able to meet our work plan estimate. We were close to our estimate but still went over time because we did not take into account setting up a data science environment using VS Code, installing python, downloading packages, installing libraries, creating a GitHub repo, reading Altair documentation extensively, etc. These tasks were something we did not take into account when aiming for this challenge goal.

IX. Testing

In terms of testing our results, the main sections that we needed to test were in the machine learning portion. Some of the computations performed had to do with finding mean and square root values so we needed to verify that these values were correct. We used the built in square root and mean functions in our code but verified them by hand using a calculator to make sure that these functions were indeed returning the correct values.

X. Collaboration:

Boyd, Karen S. “Glass Ceiling - Sage Publications Inc.” *Encyclopedia of Race, Ethnicity, and Society*, Thousand Oaks, CA: SAGE, 2008, 30 Jan. 2012, https://edge.sagepub.com/system/files/15_GlassCeiling.pdf.