Alisha Hill

Ericka Houle

Kelli Smith

December 1, 2021

Capstone Project 4

Group 2

Exploring the Gender Wage Gap

**INSPIRATION**

There is a well-known quote, “Women make $0.82on the dollar”. As an all-female group in this Data Science Capstone Project, we wanted to see if this is still true today. In the recent years there has been a lot of talk in the media and companies making an effort to close the gap in both number of females in the workplace and closing the gap in wages between females and males. We wondered if the data would show the efforts being made and if current events such as the COVID pandemic would affect the results.

**PROJECT DESIGN**

Due to our project relating to gender, we didn’t want to use the cliché baby blue and pink for our visualizations. After further research, we explored <https://blog.datawrapper.de/gendercolor/> and agreed on the Telegraph, 2018 color pattern. The male color is, light sea green, html code of #1fc3aa. The female color is, blue violent, html code of #8624f5. We wanted to utilize the same shade of black, html code #13161d for the background and dark slate gray, html code #1e2530 as an accent.

The website template was inspired by <https://themefisher.com/products/phantom-best-bootstrap-portfolio-template/> a website that provides free website templates. We felt as a team this website template is clean, unique and definitely an attention demander.

**DATA SOURCES**

We began with the thought of utilizing an existing Kaggle Dataset, “Glassdoor Gender Pay Gap” for the entire project. This data set was gathered in 2019 via a survey conducted on Glassdoor which consisted of 1,000 responses. We liked the fact that the Kaggle data consisted of gender, age, education level, seniority, department, job title and wages which would be very useful in the machine learning portion of the project. Although this was the information, we were looking we thought this data was dated and decided to see if we could find additional resources.

We found more recent data from the website of Bureau of Labor Statistics (BLS). This data consisted of occupations, gender both overall and at the state level for the years of 2015-2020 which we decided to utilize for a portion of our visualizations.

**HYPOTHESIS**

In the recent years there has been a lot of talk in the media and companies making an effort to close the gap in both number of females in the workplace and closing the gap in wages between men and women. We hypothesized the following:

* Gender wage gap has reduced over the years.
* The age gap would reduce as women age and are no longer in the childbearing years.
* Occupation does affect the wage gap.
* We can achieve at least 80% accuracy in our machine learning model.

**CLEANING DATA**

We noticed throughout literature and datasets both female/women, male/men, and pay/wage was used interchangeably. We decided to use female, male and wage across the project for consistency.

The BLS data was downloaded in individual files by year in a format that would not be conducive to merge into one data set. Since this data was not dynamic, we did the initial formatting of the data in excel since it was quicker than writing code for each year and did the final merging of the data in python. Since building a map to display our data was the plan, we merged the state data with an existing Kaggle dataset “world country and USA states latitude and longitude values” to retrieve the latitude and longitudes needed to build our map. We also dropped all nan/null values and renamed columns to import into Tableau without errors.

The Glassdoor data we decided that we would not need the PerfEval and Bonus data for the project as they were subjective based on the individual and was not a standard between individuals, so they were dropped.

**DATA TRANSFORMATION**

After cleaning the data and getting it to a point in which we could evaluate and transform, we imported the Total\_BLS\_States2 csv into Tableau for the Macro Dashboard. Our data provided the weekly salary. A calculated field was created to covert the salary from weekly to annual. We standardized both the female and male minimum/maximum salary and the average female salary as a percentage of male annually to reassure the map display a clear more accurate picture of the data. We modified the tool tips to format the information being displayed in our interactive data visualization.

In order to make the Glassdoor data work for our Linear Regression Machine learning model we needed to prepare the data set for it. We had to convert any string values to integers we did this two different was. To convert gender into an integer we wrote code to convert all Males to 0 and Females to 1. To convert Job Title, Department and Education we utilized one-hot encoding which converted each categorical value to a new column and assigning it a value of 0 or 1.

**MACHINE LEARNING**

We chose to perform a Linear Regression Model to predict someone’s wage because wage is a continuous target. For our initial model we did not think that age was relevant or would affect the model, so we decided to remove this value from the data set before splitting the data set up into test and train sets. Our regression score was only 52.2% and would show female wages higher than those of males. The feature correlations indicated there was a slight correlation between wage, seniority and education. This model also indicated that females would make more than men in all predictions, which did not seem consistent with what we had thought. Further analysis of the data set and determined that over all females make less than males, so we knew we had an issue with the model. This caused us to see if we could make modifications to the model and see if changing the feature combination would make a difference. See below for Model 1 correlation map and associated code.

Initial Model – Features: Gender, Education, Seniority, Job Title, Department

Chart

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

We reevaluated the data along with the correlation chart to see under the relationship between the features. This resulted a testing session to see which combination of features were optimal and adding age back in as a feature. After evaluating several combinations, our optimal model required us to remove education and utilize the following features: gender, age, seniority, job title and department. This provided a regression score of 82.9% and its predictions supported the data with males’ salary being greater than females’ salary. The correlation map also shows more feature correlations than the previous models. See below for final chosen Model correlation map and associated code.

Final Model – Features: Gender, Age, Seniority, Job Title, Department

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated

Once we had decided on the final model to utilize, we ran through multiple number of predictions and noticed that the gap between female and males were the same no matter the variation of inputs. This caused us to look at the data in a more detailed manor and determined that since our data set was based on a survey it might not be exactly accurate. For instance, some of the data entries indicated that some 19-year-olds had a PhD, which is highly unlikely. We determined that this might not have been the best data set to use for our machine learning model and if there was more time, we would seek a better data set.

**CONCLUSONS**  
The data showed that the national wage gap average remained relatively consistent meaning women make on average $0.82 for every one-dollar males make between the years 2015-2020. But this does not actually tell the whole story of what is really going on. We found that it often depends on the occupation and even where you live that will determine what the gap in wages are. For instance, when it comes to different STEM professions there are not many women. BLS excludes wage data when there are fewer than 50,000 females reported in that occupation and therefor can’t perform a true comparison for all occupations.

Chart, bar chart

Description automatically generated

When looking at the data at the state level we saw that there was a fluctuation from year to year within the same state. Even with the fluctuations trends in states such as California, Florida, New York and the DC surrounding states consistently outperforming others such as Wyoming, Utah and Idaho. As seen in the maps below:

2015

Map

Description automatically generated

2020

Map

Description automatically generated

Once we determined that not all states are created equal, we investigated the individual occupations to see if we could find any trends. We found that there were 19 occupations where males made a wage higher than $75K and females made up less than 50 percent of the work force. When looking at the opposite condition females wage over $75K and male workers less than 50 percent there were only 4 occupations. The occupations that had the lowest ratio of female workers to male were mainly in the STEM fields.

Graphical user interface

Description automatically generated

**LIMITATIONS**

Most of our limitations steamed from our datasets. The Glassdoor data was relatively small and some of the data was questionable when evaluated deeper. We would want to look for a larger more recent dataset and one that was not heavily dependent on surveys.

**FUTURE WORK**

To further build on our work there are a few things that we would want to take a deeper dive into what makes some states have a lower wage gap than others. We would like to identify any industry trends and lessons learned to see if they could be passed on to other industries throughout the nation.

To further take a deeper dive into the wage gape we wanted to explore the if the impact based on age. If we had more time and the correct dataset we would like to see if the wage gap fluctuates based on the age of employee. We would take those trends and see if there is correlation between the age and what’s considered typical life events such as female child bearing years.

When it comes to the machine learning model, we would like to gather a larger dataset that we could train and test on to get more accurate. It would be interesting to see if we could achieve different results by changing the type of model performed such as a Nero network.

With unlimited resources we would create a website that can toggle between light and dark mode according to user preference.

**WORKS CITED**

<https://www.kaggle.com/nilimajauhari/glassdoor-analyze-gender-pay-gap>

<https://www.glassdoor.com/research/how-to-analyze-gender-pay-gap-employers-guide/>

<https://www.bls.gov/>

<https://themefisher.com/products/phantom-best-bootstrap-portfolio-template/>

<https://blog.datawrapper.de/gendercolor/>

<https://www.kaggle.com/paultimothymooney/latitude-and-longitude-for-every-country-and-state>