**Introduction**

Our group of three geniuses invented a “no fail” project that is now ready to hit the market. In order for the product launch to be successful, we needed an established company and a good workforce. As previously mentioned, the group is not skilled in running a company and thus needed some help. We are, however, data savvy and spent the last 11 weeks utilizing our big data analytics skills to comb through a publicly available salary data set to accomplish the task of establishing a successful company.

In this report we will discuss how we used the salary data set to learn about employment best practices. Our aim is to make sure the positions we fill are hired at the right salary rate based on education and experience. Our aim is to avoid gender, age, and race discrimination in our hiring practices as we are an equal opportunity employer. We will also make sure our employees are being paid based on location. We will also attempt to forecast salary ranges for our employees as they progress in their careers with us.

**Data**

The dataset consists of a comprehensive collection of salary and demographic information based on a survey conducted with around 6,700 real employees. It offers a valuable resource for studying the relationship between income and various socio-demographic factors. The demographic attributes include age, gender, education, country, and race, providing a diverse range of variables for analysis, shown below.

The data is merged with the Chicago salary average to ensure the workforce will be appropriately compensated based on location.

Below is a .head() of the dataset being used for the remainder of the salary compensation objective:

A screenshot of a computer screen

Description automatically generated

**Data Cleaning**

Before any analysis could be done, the group had to ensure that the dataset was ready to be worked with. At a quick glance the data seemed to be ready for analysis:

A screenshot of a computer program

Description automatically generated

Df.dropna() was used to remove rows with missing or NaN values from the dataframe. As the above image showed; “Age”, “Experience”, “Salary” and, “Chicago\_avg” are all floats and therefore had to be transformed into integers. All columns were grouped into arrays to make sure no redundancies were observed in the Dataset. Finally, all abbreviations in occupation titles were corrected to ensure uniformity in the data.

**Initial Discovery**

Based on the dataset, the job market is dominated by Whites and Asians. We decided to leave African American and Blacks in separate categories due to race being self-reported. Therefore, it would be unethical to group people together in a way they do not identify with.

A graph with red and blue bars

Description automatically generated

We also observed a majority of the workforce being within the age of 24 – 45 as shown below:

A graph of number of responders

Description automatically generated

We also observed bachelor’s education being the highest educational achievement for most, followed by Master’s, Ph.D. then High School.

A graph of a number of people

Description automatically generated with medium confidence

Males Lead the workforce, and there are a small portion of non-binary survey respondents:

A graph of a number of responders

Description automatically generated

Number of respondents by country showed an even distribution as shown below:

A graph of number of responders

Description automatically generated

We observed a normal distribution within the salary histogram where 90% of the salary range fell within $50,000 - $200,000.

A graph of blue columns

Description automatically generated

Finally, we observed that the majority of experiences fell between 0 – 20 years.

A graph of a number of individuals

Description automatically generated

**Analysis**

A graph of a salary

Description automatically generated

The above bar plot was unsurprising, as one will expect salary and age to correlate to one another. As age goes up, salary will go up in most cases as well.

Since we want to avoid discrimination between race and salary and were thus curious about industry practices, we took a look at the plot between race and salary.

A graph of blue and white bars

Description automatically generated

Salary by education level showed what we will typically expect, the higher the education, the more money you make and that is evident in the image below:

A graph of a salary

Description automatically generated with medium confidence

We decided to look at three of the main variables we have been exploring. Age, experience, and salary. Experience and salary correlates in this dataset. So does age. As we can see, as age goes up, so does experience and in turn, the workforce is rewarded with a higher salary.

A graph showing a number of blue dots

Description automatically generated



**Modeling**

We performed a linear regression model. We wrote a code to specify and fit a linear regression model to our salary dataset. Salary was the dependent variable in this model with age, gender, education, experience , race, and country being the independent variables.

Adjusted R-squared being 71.7% indicated that 71.7% of the variance in the salary is explained by the variables used in the linear model.

P>|t| are 0 for Gender, Education, Age, and Experience leading us to conclude that these predictors are statistically significant. We can also conclude that race and country are not significant in predicting salary as they are higher than 0.05.

A screenshot of a computer screen

Description automatically generated

We ran the model again based on the discovery that country and race were not statistically significant and similar results were observed with the R-squared and adjusted r-squared. We could conclude from the model that gender is not statistically significant as well.

A screenshot of a computer

Description automatically generated

**Employee 1: Data Analyst**

In the next phase of our analysis, we set to identify specific employment opportunities within our company that we needed to fill.

We first looked at the “Data Analyst” role.

A graph with colored dots

Description automatically generated

As the graph shows, as the analysts grow within their career fields, so did their salary. However, there could be several factors for this. We first did a linear model of salary using age, gender, education, experience, race, and country. Age and experience were found to be statistically significant. In the case of the data analyst role, a Master’s degree appeared to correlate with lower income which could be a result of various factors not captured in the data. Due to a lower income with higher education being unrealistic, we dropped it from the modeling.

A screenshot of a computer

Description automatically generated

With that information, we decided to predict the salary using only experience. We dropped the age variable since we are equal opportunity employers and wanted to avoid discrimination or any potential lawsuit.

A screenshot of a computer

Description automatically generated

Data analyst modeling based on experience showed an adjusted r-squared of 41.8% from 74.4% since we dropped a few variables. We think because Data Analyst is an entry level role, most people even with a master's degree but no years of experience must start at the entry pay before graduating into a senior data analyst role.

**Employee 2 – 4:**

We used the same technique for the Financial Analyst, Customer Support, and Software Developer role.

**Financial Analyst:**

A screenshot of a computer

Description automatically generated

**Customer Support:**

A screenshot of a computer

Description automatically generated

**Software Developer:**

A screenshot of a computer

Description automatically generated

The Finance Analyst, Customer Support and Software Development roles all boasted adjusted R-squared's of 0.94, 0.87 and 0.92 respectively as well as significant p-values which dictates that our models fit well.

**Salary Prediction**

We set out to predict salaries we would pay the 4 roles we needed to fill.

**Data Analyst:**

A screenshot of a computer

Description automatically generated

We predicted the starting total compensation of a Data Analyst with 2 years of experience to be $107,254.36 based on living in Chicago.

**Financial Analyst:**

A close-up of a sign

Description automatically generated

The predicted starting total compensation for the Financial Analyst with 2 years of experience is $63,448.32.

**Customer Support:**

Our customer support specialist was the only role where education ended up being a relevant variable. In order to capture the three scenarios (High School, Bachelor's, and Master's), the function below was written. This will be used to create a pay range, where the potential candidates will be paid based on their educational level and experience.

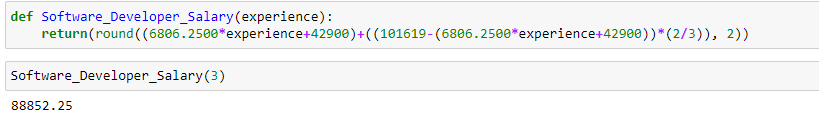
A screenshot of a computer program

Description automatically generated

The predicted starting Customer Support salary for an employee with a High School diploma and 3 years of experience is $40,764.33.

**Software Developer:**

We noticed that the salary predicted by the model was significantly lower than what a Software Developer would be receiving in the United States, as well as Chicago specifically. To boost the pay, there has been a buffer implemented where the new hire will receive two-thirds of the difference between the calculated pay and the average Chicago pay. This pay buffer will be on top of the hiring bonus provided to the new employee.



The predicted starting salary for a Software Developer with 3 years of experience is $88,852.25.

We compared the average salary from the survey we fit our model to the average salary for those same roles in Chicago, according to Glassdoor.com. It shows that our models will tend to be overpaying Data Analysts while underpaying Software Developers. We have addressed the underpaying of Software Developers in our salary function, but the overpaying of Data Analysts will have to be addressed based on the role’s specifications and company needs.

A graph of a survey and salary comparison

Description automatically generated

**Conclusion**

We have learned that there is a correlation between age, years of experience, and salary. Our linear model also emphasized the statistical significance of those variables and simultaneously confirmed that gender, race, and country is not statistically significant in the salary of the workforce. Using insights from our models, we were able to establish salary predictions and ranges for the four roles that we needed to fill within our company. We were also able to avoid any gender, age, and race discrimination in hiring our new employees. More importantly, we were able to accurately determine starting salaries for employees that will make them happy and productive.

We wanted to forecast their salaries over time, however we determined that wasn’t necessary and it required changing and acquiring all new data to accomplish that goal. We decided to be a unique company so we disregarded the forecast and will increase pay based on productivity and profits as the company grows.