MGT6203 Group Project Final Report

November 28, 2023

Team 101

Alain Daccache, Daniel Avila, Jeffrey Triemer, Bruce Dale Service, Ethan Goldstein

Table of Contents

[Abstract 1](#_Toc152421489)

[Introduction 1](#_Toc152421490)

[Overview of Data 2](#_Toc152421491)

[Exploratory Data Analysis 2](#_Toc152421492)

[Data Cleaning 2](#_Toc152421493)

[Salary Distribution 3](#_Toc152421494)

[Categorical Value Distribution 3](#_Toc152421495)

[Heatmap Analysis 5](#_Toc152421496)

[Modelling 6](#_Toc152421497)

[Initial Hypothesis and Statistical Testing 6](#_Toc152421498)

[Initial Modelling Techniques 7](#_Toc152421499)

[Modelling Results and Assumptions 7](#_Toc152421500)

[Model Improvements 8](#_Toc152421501)

[Variance Inflation Factor (VIF) 8](#_Toc152421502)

[Stepwise Regression 8](#_Toc152421503)

[Addressing Cook’s Distance Outliers 8](#_Toc152421504)

[Exploring Interaction Terms 8](#_Toc152421505)

[Final Model 8](#_Toc152421506)

[Coefficients Interpretation 8](#_Toc152421507)

[Key Insights from Regression Analysis 9](#_Toc152421508)

[Statistically Significant Predictors: 9](#_Toc152421509)

[Insights 9](#_Toc152421510)

[Location Insights 9](#_Toc152421511)

[Education and Experience Insights 9](#_Toc152421512)

[Industry Insights 9](#_Toc152421513)

[Gender and Race Insights 10](#_Toc152421514)

[Regional Interactions with Education 10](#_Toc152421515)

[Conclusion 10](#_Toc152421516)

[References 10](#_Toc152421517)

[Appendix 11](#_Toc152421518)

# Abstract

The global pandemic introduced unprecedented challenges, bringing the complexities of job compensation to the forefront. This research delves into the post-pandemic changes in U.S. worker incomes, giving special attention to the interplay of demographic elements. Through a structured approach encompassing data preprocessing, in-depth exploratory analysis, hypothesis validation, and regression modeling, this study unveils subtle patterns in salary disparities.

The preliminary analysis of U.S. worker incomes reveals significant influences of demographic factors, such as gender and race, on salary disparities, with certain industries and levels of experience correlating to higher salaries. However challenges like potential outliers, model refinement through feature selection, and exploring interactions between predictors require further investigation and sophisticated modeling approaches in subsequent phases of the research.

The data, code, and associated visualizations can be found in the Team 101 GitHub repository [here](https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-101). For detailed instructions on setting up and running the code, refer to the README.md file in the repository.

# Introduction

The professional world has always been in flux, but recent global events have amplified the need to understand the nuances of job compensation. Deep-rooted issues of pay disparities, long a point of contention, have now become central to discussions about professional equity and justice. As job landscapes change and economic divides become more pronounced, there's an urgency to dissect the dynamics of worker compensation, especially for those in more vulnerable wage brackets [1]. This study aims to uncover the different elements that influence salary dynamics, with a keen interest in understanding the potential impact of pandemic-related disruptions on salary scales and the role of deep-rooted biases in these changes.

The primary objective of this project is to analyze and forecast job salaries in the context of the evolving employment landscape in the United States, while investigating the underlying drivers that influence salary level. This analysis aims to address the following key questions:

* How do factors such as occupation and age contribute to variations in job salaries, and is there some discernible demographic biases that affect those salaries?
* How can we predict an individual's salary based on such attributes so that they have more confidence in their current compensation, if applicable?

It is hypothesized that the most important predictors of salary would be certain occupation fields and years of experience.

An in-depth methodology with an emphasis on data cleaning, preparation and exploratory data analysis forms the basis of this analysis. Single-variable and multivariate analyzes are conducted, enhanced by statistical methods to confirm the identified patterns. Other steps in feature engineering refine the presentation of the data, such as the use of dummy variables, and improve the accuracy of the model, for example through logarithmic transformations.

The results presented in this report stand out due to its holistic view, blending both demographic and professional metrics to understand salary determinants. These metrics have been examined through their collective impact on salary as interactions between these terms yield greater insights than examining them on their own, especially in a post-pandemic world. By focusing on often-neglected lower-wage workers, our study provides insights that are both fresh and broadly relevant.

# Overview of Data

*Link to the human resources dataset: https://www.kaggle.com/datasets/rhuebner/human-resources-data-set*

The human resources dataset comprises respondents' professional and demographic details, including age, industry, job title, annual salary and bonus, state of employment, years of experience, education level, gender, and race. Let us provide a concise walkthrough of the data preprocessing and cleaning steps:

* **Column Renaming:** Columns have been renamed for clarity and standardization.
* **Data Type Conversion:** Character columns were cleaned and converted to factors.
* **Data Standardization:** The content of certain columns was standardized.
  + For example, ‘Industry’ was a freeform text response, and many unique responses were provided in the initial dataset. A regex pattern matching algorithm was implemented to standardize those values. Any values that could not be sorted into a defined group were discarded.
* **Missing Values Treatment:** Missing continuous values were assumed to be zero, while survey responses with missing or multiple responses were removed from analysis.
* Since our only continuous variables are annual\_salary and annual\_bonus, we will simplify our analysis by adding up those two columns.

Approximately 96% of available data was maintained through the cleaning and standardization process. This process facilitated easier data handling and analysis and ensured consistency in reporting and modelling.

# Exploratory Data Analysis

Exploratory data analysis is the foundational analytical step, revealing underlying patterns, potential anomalies, and relationships among variables. The dataset encompasses various demographic and professional attributes of respondents, including age, industry, job title, annual salary, state of employment, years of overall and field-specific experience, education level, gender, and race. This paper explores the distributions of individual variables via univariate analysis and their relationships through multivariate analysis, with the aim of unveiling the nuanced factors that might influence an individual's salary.

It is important to note a sampling bias in the dataset, as all responses were voluntary. As such, salaries for those industries and individuals outside the sample data cannot be generalized due to the inherent non-randomness of this survey’s sample.

## Data Cleaning

The dataset comprises several categorical variables, posing the potential risk of the curse of dimensionality if not managed appropriately. To reduce the dataset’s dimensionality, the 51 unique states (+ the District of Columbia) were grouped into three broad regions - West, Central, and East. To further streamline the analysis gender classifications other than male and female were removed from the dataset.

## Salary Distribution

|  |  |
| --- | --- |
| A graph of a salary  Description automatically generated | Salary, the only continuous variable, is the target variable in this dataset. The salary ranges from $0 to $3,600,000, with a median $83,000, average $103,524, first quartile $59,751 and third quartile $121,000. |

## Categorical Value Distribution

Most respondents hold a college degree, followed by a master’s degree. The other categories (high school, PhD, professional degree) each have less than 1000 responses. In terms of gender, most respondents are female (14209), followed by men (3423). The states with the most respondents are California (2130), followed by New York (1783). Other categories, including industry and race, are illustrated in the plots below.

A screenshot of a graph

Description automatically generated

Most respondents are white, women, hold a college or master’s degree, are between the ages of 25 to 44, hold 11 to 20 yrs. of work experiences, work in computing, education, nonprofits, and accounting or banking or finance and reside in California or New York. This information provides essential insights into the dataset’s representativeness and is essential for understanding any potential biases in subsequent analyses.

Further disparities can be discovered by analyzing salaries across gender, race, and age, as shown in the visualizations below.

A graph of a graph with text

Description automatically generated with medium confidenceA graph with blue squares and black dots

Description automatically generatedA graph with blue and black squares

Description automatically generated

The race with the most salary variability is White. However, the race with highest median salary is Asian/Asian American, at around $100,000, followed by Middle Eastern or North African, at $94,000, while the race with lowest median salary is Hispanic, at approximately $80,000, followed by white, at $81,000.

In terms of gender, we observe that men have the highest median salary, at $112,000, while women’s stand at $79,000. These insights give us a preliminary bias into salary differences amongst races and genders.

In terms of industries, Computing or Tech have the largest median salary, followed by Law, Business or Consulting, and Banking & Finance. Further analysis on the interaction between industry and race / gender was completed to identify any cofounders. As there is an imbalance of white and female responders, relative metrices are examined (as opposed to absolute values).

As shown bar plot below, there are no significant differences in the industry distribution across Asian Americans and Hispanics, which can lead to the hypothesis that there is a further selection bias in the dataset. While there are more people who identify as Hispanic working in industries this analysis shows are lower paying, such as HR and Social Work, that percentage is capped at 3% and is not significant.

|  |  |  |
| --- | --- | --- |
| Percentage of Each Gender in Different Industries | Percentage of Each Race in Different Industries | A list of different colored squares  Description automatically generated |

When it comes to gender, by visualizing the bar plot below, we notice an overwhelming majority of men work in Computing or Tech (~39% of our dataset), while only third of that number of women work in that industry (~13%). Conversely, almost double the percentage of women work in Education (11% women; 6% men), triple the percentage of women work in Non-profits (~11.4% women; 4.6% men).

Additionally, when examining years of work experience distributions to explain pay gap between male and female, we realize the distributions are similar in our graph on the left. However, at each level of years of experience, we realize men earn (median-wise) more than women.

A graph of a gender

Description automatically generated with medium confidenceA graph of different colored squares

Description automatically generated

Finally, when comparing differences in races across states, we notice a large proportion of Asian Americans and Middle Eastern in a traditionally high-paying state (New York at 15% each), while the largest proportion (16%) of Native Americans live in Oklahoma, a state with the third lower median salary, at $53000. We can see that races do have preferences per state, and states can therefore provide an explanation as to why salaries can differ among races.

### Heatmap Analysis

Comparing annual salary against industry, state, gender, and race reveal that primary/secondary educators seem to be well-paid, which is surprising as this industry is commonly acknowledged for its lower pay.

A graph of a salary

Description automatically generated with medium confidenceA chart of a salary

Description automatically generated with medium confidence

A graph of a salary against gender

Description automatically generated with medium confidenceA colorful squares with black lines and white text

Description automatically generated

Companies provide approximately equal compensations to men and women throughout the pay-scale range. It is surprising because conventional wisdom suggests that women tend to be low-paid for most jobs while men are typically well-paid for the same work. People who identify as Hispanic currently tend to be well-paid for high-end jobs, while those who identify as white historically have used to be well-paid for.

# Modelling

The predictors used in this analysis are exclusively categorical, which influenced our modeling decisions. For binary predictors, like gender, the linearity assumption is inherently satisfied. In terms of nominal categorical predictors with multiple categories, such as age or education level, there's a discernible trend in the data as illustrated in the graphs below.

A graph showing the age of a person

Description automatically generated with medium confidenceA graph showing a number of different levels

Description automatically generated with medium confidence

However, it's essential to approach ordinality with caution. For instance, while a master’s degree is academically a step above a college degree the median salary doesn’t reflect a substantial difference between the two. Age presents an intriguing pattern: post the age bracket 35 to 44 the number of upper-tail outliers and but average salaries dwindle.

### Initial Hypothesis and Statistical Testing

From the exploratory data analysis, it can be hypothesized that pay gaps between males and females may be partly explained by the traditionally higher percentage of such genders in particular domains, while gaps between races can be explained partly by their geographic preferences, which in turn impact salary. Running an ANOVA test to evaluate the effect of differences in industry, state, gender, and race, each on annual salary, return a p-value less than .05, telling us that there is a statistically significant difference amongst such attributes. Running a Chi-squared test across combinations of such attributes also makes the p-value less than .05, indicating that there are differences between classes.

## Initial Modelling Techniques

Multiple linear, semi-log, and logistic regression models were iterated through to accurately model the data and predict salaries. The quality of the factors were evaluated using common statistical tools, such as p-value, confidence interval, and t-stat analysis. The overall models were evaluated on their R-squared, adjusted R-squared, and RMSE values. F-score was also examined for linear regression models, while metrics like Precision, Recall, AUC-ROC were analyzed to evaluate the logistic regression models.

Furthermore, the model’s performance in terms of variance, independence, normality, and heteroscedasticity, were evaluated on using diagnostic plots, leading to multiple rounds of data preprocessing, feature engineering, and feature selection steps above.

When interpreting the models, extra care was taken to account for the impact of variables while keeping others constant. and examine the significance of coefficients i.e. keeping all else constant, and on average, one unit or % increase of one variable could result in how much of an increase / decrease of the response variable (salary). Interpreting indicator variables w.r.t. base case, and same with the interaction terms.

The final model was optimized using Ridge, LASSO, Elastic Net, and cross-validation. These measures were implemented to address overfitting and optimize the model’s hyperparameters. Specifically, alpha, the penalty term for amount of shrinkage and the classification threshold, w.r.t. to the quality metrics mentioned previously, to select the best models.

## Modelling Results and Assumptions

The initial model led to an adjusted R-squared equal to .3009. A subsequent regression with ordinality yielded similar performance, but coefficients become harder to interpret since they succumb to quadratic and cubic transformations. A log transformation on the response variable improved the adjusted R-squared to .3056 and improved the residual distribution. The summary table can be found in the Appendix. The residuals shown on the Residuals vs Fitted plot (below) indicate that our model might be underestimating salaries, especially at higher ranges, which could be due to right-tailed outliers.

A graph of a graph showing a number of values

Description automatically generated with medium confidence A graph of a graph showing the amount of salary

Description automatically generated with medium confidence

Using Cook’s Distance cutoff of 4/*n*, where *n* is the number of data points (17204), 505 influential records (or 2.9% of the dataset) were identified. These outliers are addressed in the next section of this report.

However, challenges remain as the standardized residuals do not follow a normal distribution (Q-Q Residuals Plot under Residuals vs Fitted plot). Notably, other advanced regression techniques might offer robustness against outliers and capture non-linearity more effectively, as highlighted in Kibekbaev’s and Duman’s *Benchmarking Regression Algorithms for Income Prediction Modeling* [2].

## Model Improvements

### Variance Inflation Factor (VIF)

Factors like Age, overall experience and field experience had high multicollinearity; therefore, only Field experience was included in the final model as it appears to have be the most relevant in predicting Salary. This change adversely impacts our adjusted R-squared, reducing to 0.3313, but the multicollinearity effecting the model has been eliminated.

### Stepwise Regression

A stepwise regression algorithm was implemented to reduce the number of input parameters and simplify / strengthen the model. This approach did not remove any features, suggesting that further removal was not conducive to a better model performance and each factor is important in predicting salary.

### Addressing Cook’s Distance Outliers

Outliers were identified using Cook’s Distance to ensure the model was not overfit. This results in an immediate improvement to the adjusted R-squared value, which increased to 0.51. However, removing outliers may induce bias, so median values we imputed to replace the outliers and improve model performance.

Cross-validation and LASSO regulation were performed on both versions of the dataset, one with only the outliers removed and one with the outliers replaced with the median results. Imputation slightly decreases the overall adjusted r squared to ~ 0.49 but improves the model’s RSME, while ensuring that a reasonable amount of bias is maintained due to the imputation technique.

### Exploring Interaction Terms

One model was fitted per pairwise interaction to determine statistically significant interaction terms. This approach shows that industry and region, industry and field experience, industry and education level, industry and gender, industry and race, and region and education level are the interaction terms that improve statistical significance and adjusted R squared. The final model was adjusted to ensure these interaction terms were maintained as inputs.

### Final Model

With model improvements in place, the final model was built using feature selection, regularization, and cross validation to address overfitting. This outputs a final adjusted R squared of 0.864 and an RSME test of 0.184.

This model is likely overfit and may benefit from further PCA. The new adjusted model’s results can be seen in the “Feature Engineering & Modelling” notebook, located in the Team 101 GitHub repository.

Coefficients Interpretation

With the new model we can display updated estimates to Salary as expected given all else constant from our model. Any variations from this baseline can be interpreted using the rightmost column of our table on the left. For instance, a category with a +10% change implies a 10% salary increase over the base. To illustrate, given a survey respondent has 41 years of experience and works in hospitality, our model would expect an increase in salary by ~20.3% compared to the baseline reference categories (which have an intercept value of $43,052.05).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor**  **(Positive Salary Correlation)** | **Percent Change** |  | **Predictor**  **(Negative Salary Correlation)** | **Percent Change** |
| industry\_field\_experience\_yearsHospitality & Events.41 years or more | 20.26% |  | industry\_raceAgriculture or Forestry.Hispanic | -10.26% |
| field\_experience\_years41 years or more | 18.17% |  | industry\_regionAgriculture or Forestry.Central | -10.31% |
| field\_experience\_years21 - 30 years | 17.06% |  | industry\_raceSocial Work.Native American or Alaska Native | -10.33% |
| field\_experience\_years31 - 40 years | 16.06% |  | industry\_field\_experience\_yearsGovernment and Public Administration.1 year or less | -10.34% |
| industry\_education\_levelEntertainment.Professional degree | 15.60% |  | industry\_raceAgriculture or Forestry.Native American or Alaska Native | -10.36% |
| field\_experience\_years11 - 20 years | 13.88% |  | industry\_education\_levelLeisure, Sport & Tourism.High School | -11.28% |
| industry\_education\_levelEntertainment.PhD | 13.76% |  | industry\_regionHospitality & Events.Central | -11.30% |
| industry\_education\_levelMarketing, Advertising & PR.High School | 10.55% |  | industry\_field\_experience\_yearsEducation (Higher Education).1 year or less | -11.31% |
| industry\_raceHospitality & Events.Asian or Asian American | 10.18% |  | industry\_field\_experience\_yearsNonprofits.1 year or less | -11.54% |
| field\_experience\_years8 - 10 years | 10.17% |  | industry\_education\_levelSocial Work.College degree | -11.88% |
| industry\_raceLeisure, Sport & Tourism.Black or African American | 10.03% |  | industry\_field\_experience\_yearsLaw Enforcement & Security.1 year or less | -13.05% |
| industry\_field\_experience\_yearsRecruitment or HR.31 - 40 years | 9.85% |  | industry\_education\_levelNonprofits.High School | -13.30% |
| industry\_raceLaw.Native American or Alaska Native | 9.39% |  | industry\_education\_levelEducation (Primary/Secondary).College degree | -13.72% |
| education\_levelPhD | 9.34% |  | industry\_field\_experience\_yearsArt & Design.31 - 40 years | -14.64% |
| industry\_education\_levelComputing or Tech.High School | 9.06% |  | industry\_raceEducation (Primary/Secondary).Native American or Alaska Native | -15.73% |
| industry\_education\_levelMedia & Digital.High School | 9.05% |  | industry\_field\_experience\_yearsHospitality & Events.1 year or less | -16.60% |
| industry\_raceHealth care.Native American or Alaska Native | 9.04% |  | industry\_education\_levelRetail.High School | -16.85% |
| industry\_raceEducation (Primary/Secondary).Black or African American | 8.39% |  | industry\_raceTransport or Logistics.Native American or Alaska Native | -18.17% |
| stateDistrict of Columbia | 8.33% |  | industry\_field\_experience\_yearsRetail.1 year or less | -18.90% |
| industry\_field\_experience\_yearsUtilities & Telecommunications.5 - 7 years | 8.21% |  | industry\_education\_levelEducation (Primary/Secondary).High School | -24.35% |

## Key Insights from Regression Analysis

Statistically Significant Predictors:

Gender, regions (Central, East, West), race, education level, and field-specific experience all demonstrated statistical significance with p-values less than .05. Age's significance varied across its brackets, while the overall\_experience\_years variable showed no significant distinction, indicating its potential removal in future model iterations.

### Location Insights

High cost of living states, including California, Massachusetts, New York, and the District of Columbia all have a positive influence on salary while Arkansas, Kentucky, Louisiana, Maine, Montana, Oklahoma, South Carolina, and West Virginia all have a negative influence on salary.

### Education and Experience Insights

Higher post-secondary education (professional, masters, and doctorate degrees) have the largest positive impact on salary. Those with only a high school diploma or some college education can expect a lower-than-average salary, especially in the Eastern region and in the Business/Consulting and Health Care industries.

### Industry Insights

Working in the Nonprofit, Entertainment and Hospitality, Social Work, Higher Education industries has an overall negative effect on salary and is often compounded by region, education, and less working experience.

Engineering/Manufacturing careers have an overall positive influence on salary, especially in Central regions. Positive interactions in the Marketing, Advertising & PR, Media & Digital industry were also noted in specific regional interactions.

The Accounting, Banking & Finance, Computing or Tech, and Health care industries did not have a significant impact on salary when examined on their own, however, these industries had a notable influence on salary when combined with state and one other factor.

### Gender and Race Insights

Identifying as Black or African American generally had a mixed impact on salary, those who work in industries like Entertainment and Sales may expect a positive impact to salary, while those working in Accounting, Banking & Finance and Higher Education may expect a negative impact.

People who identify as Middle Eastern or North African and Native American or Alaska Native and work in law have a higher salary, on average. However, working in Government and Public Administration/Health Care or Education (Primary/Secondary) / social work may see a lower average salary, respectively.

Salary of those who identify as Asian or Asian American can expect a positive impact in specific industries like Hospitality & Events, Marketing, Advertising & PR, and Media & Digital. Identifying as Hispanic has a negative impact on salary in the Agriculture or Forestry industry and a positive impact in Government and Public Administration and Media & Digital.

Across all races, women are likely encountering negative salary impacts in specific industries like Computing or Tech, Higher Education, and Nonprofits. People of all genders who identify as white likely see a negative impact on their salary in Education (Primary/Secondary) and Retail but a positive in Insurance.

### Regional Interactions with Education

Individuals who work in the Central region will generally have a lower salary with only a high school diploma. Professional degrees and doctorates are predicted to have positive impacts on salary.

Similar results apply to those who work in the Eastern region, as holding a high school degree and doctorate have a negative and positive impact on salary, respectively. Finally, holding a professional degree in the West region is associated with a positive impact on salary.

# Conclusion

This analysis has provided several insights into the factors influencing salaries. Demographic factors such as gender or race were shown in exploratory analysis and regression modelling to play a role in determining salaries, thus indicating concerning bias.

Sophisticated model selection and hyper-parameter optimization form the basis of this analysis. Efforts to incorporate imputation of outliers, remove multicollinearity, and add significant interaction terms led to large model improvements, allowing for the prediction of salary given survey respondent factors with improved accuracy and model fit.

# References

[1] Johnson, Elizabeth R., and Ashley V. Whillans. "The Impact of the COVID-19 Pandemic on the Satisfaction of Workers in Low-Wage Jobs." Harvard Business School Working Paper, No. 23-001, July 2022.

[2] Kibekbaev A. and Duman E. “Benchmarking Regression Algorithms for Income Prediction Modeling”. 2015 International Conference on Computational Science and Computational Intelligence

# Appendix

1. Initial Regression Summary for our Log-Linear Model

A black and white text with white text

Description automatically generated