Introduction

In 2021, a TikTok survey aimed to enhance salary transparency and improve employees' negotiating power. Participants were requested to fill out their salary and qualifications on a shared Google Sheets document, which resulted in over 80,000 entries. The survey data was disorganized, allowing me to utilize my expertise and expand my knowledge. I used SQL and ETL packages to clean up the information, but this notebook still needs more. The subsequent steps involve more in-depth analysis and the creation of machine learning models to predict salary based on factors such as gender, industry, and years of experience, which will be detailed in the "Sal_Surv_Analysis" notebook.

Data Collection and Cleaning Process

Data collection involved soliciting salary and qualification information through a TikTok campaign, resulting in unstructured data. Data files were first processed and wrangled through SSIS packages and SQL tables, but significant cleaning and preprocessing were still necessary. The dataset's open-ended nature led to non-standardized entries, necessitating extensive cleaning.

The salary survey data underwent comprehensive cleaning, addressing missing values across columns such as 'Industry' and 'Job Title'. An initial analysis revealed varied distributions in salaries, annual bonuses, and sign-on bonuses, indicating potential factors influencing these distributions. Outliers were detected using the Interquartile Range method, identifying a significant number primarily in bonus columns. Care was taken to handle missing values, with caution exercised to avoid biased results. Rows with missing values were dropped to ensure data integrity.

The salary survey data cleaning process involved several steps to further standardize and categorize the dataset for better uniformity and analysis.

Standardization of Categorical Columns:

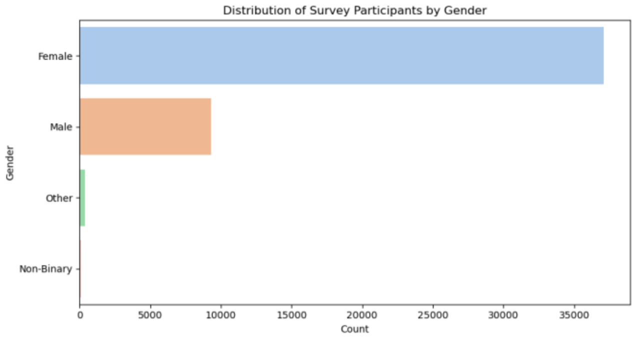
1. The column 'Highest level of Education Received' was renamed 'Education'. Any misspellings in the education levels were corrected. Based on the content of the entries, the data was grouped into broader categories such as 'No Schooling', 'Trade School', 'Some College', 'High School Diploma', 'Associate Degree', 'Bachelor's Degree', 'Master's Degree', and 'Doctoral Degree'.
2. The 'Industry' column was standardized by first scraping a comprehensive list of industries from the Bureau of Labor Statistics website. This list was then used to map the survey's industry data to these standardized titles using fuzzy matching, ensuring high accuracy by setting an 80% match threshold.
3. The 'Job Title' column underwent a similar process as the 'Industry' column, with job titles scraped from the Bureau of Labor Statistics website. Fuzzy matching was then employed to standardize and categorize the job titles. Due to the many unique job titles, parallel processing and batch processing were used to manage the process efficiently.
4. Misspellings were corrected for the 'Country' column, and similar country names were grouped to standardize the entries.
5. Salary columns were converted into USD using country-specific currency codes, conversion rates, and exchange rates obtained through web scraping. This ensured that the salary data was accurate and consistent across different countries. This systematic approach ensures consistency and accuracy in the dataset's salary predictions across various countries.

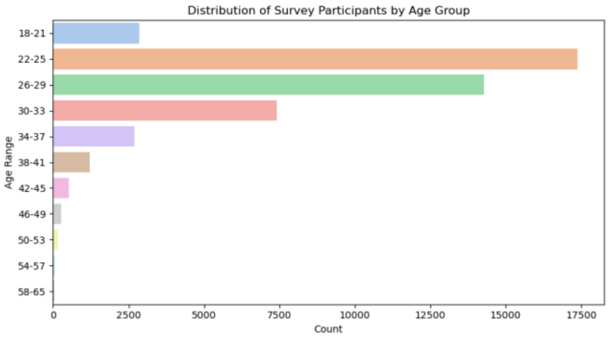
6. The 'Gender' column was grouped to categorize the data into uniform classes.

The dataset's length and structure were examined to ensure that it was correctly processed and that the data frame's structure remained intact after the cleaning steps. The cleaned file was then saved as one CSV file to continue with the analysis.

Summary of Analysis and Visualizations

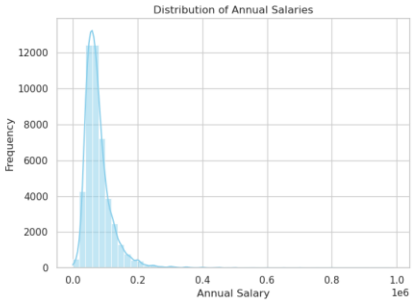
Participants Demographic

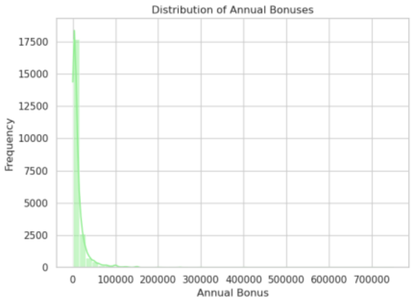


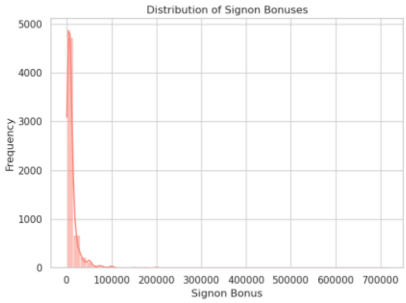


The dataset predominantly comprises Female participants, with Males forming a smaller yet substantial group. Non-Binary or Other identifications are notably scarce. Age-wise, younger demographics, particularly 22-25 and 26-29 age groups, dominate the survey, while older age brackets show lower representation, possibly skewing results towards younger individuals' experiences. Moreover, the dataset exhibits a significant skew towards U.S.-based respondents, with countries like Canada and Australia following suit, while some nations, like Singapore and Lithuania, are underrepresented. These findings suggest potential biases that should be considered in analysis and modeling, such as addressing class imbalance in gender and the demographic skew towards younger age groups.

Salary Distributions

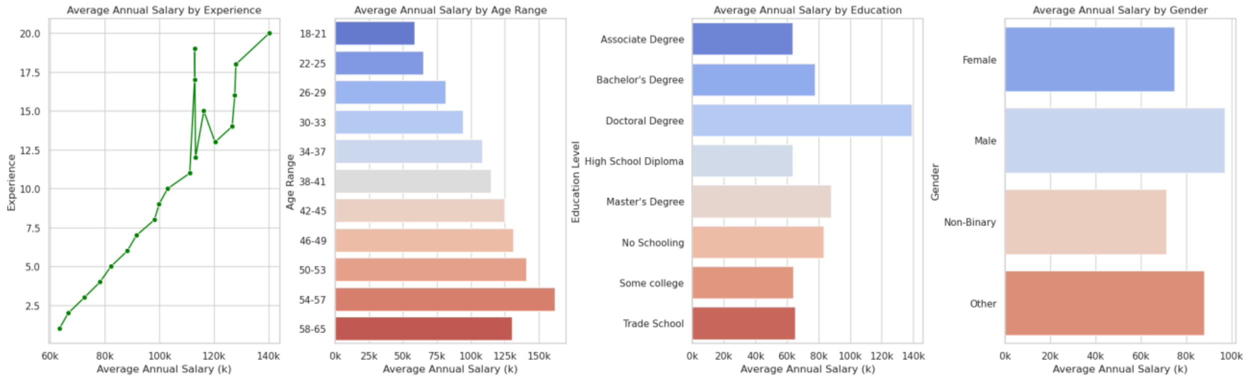






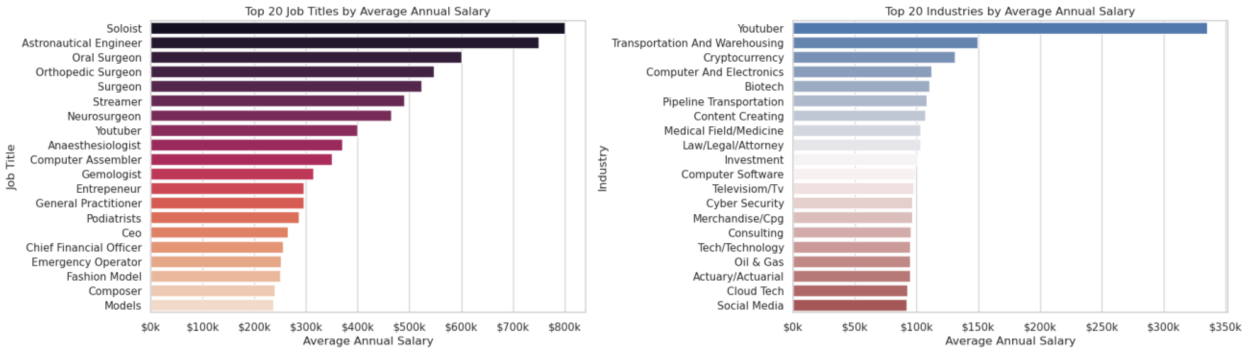
The histograms illustrate right-skewed distributions for Annual Salaries, Annual Bonuses, and Signon Bonuses, indicating that most participants receive lower compensation, with fewer individuals earning higher amounts. This pattern reflects typical socio-economic dynamics in salary data. Additionally, bonuses across various ranges suggest variability in compensation structures. In contrast, the wide range of sign-on bonus amounts may be influenced by job level, industry, or negotiation skills.

Average Salary by demographic



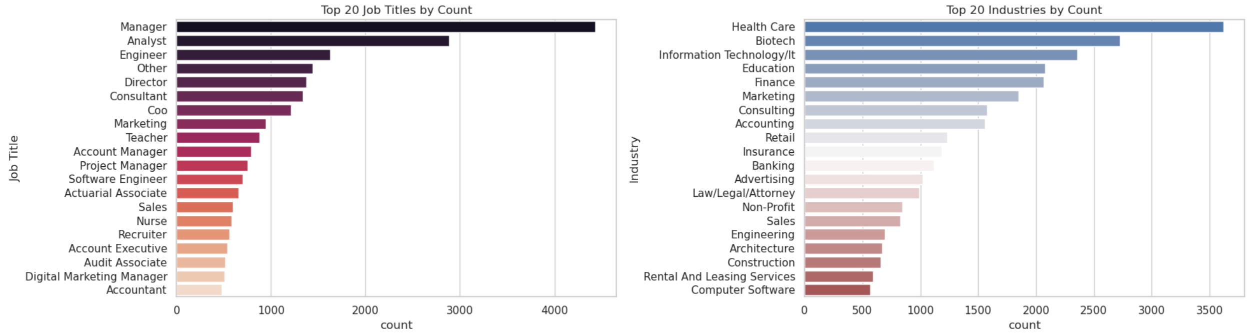
The visualizations provide valuable insights into the average annual salaries across various demographics. The line plot depicting average salary by years of experience reveals a positive correlation, suggesting that more experienced individuals tend to earn higher wages. However, other factors may also influence earnings. Similarly, the bar plot illustrating average salary by age range demonstrates an increasing trend, reflecting the typical career progression and accumulation of experience over time. Additionally, the plot showcasing average salary by education level highlights the impact of higher education on earning potential, albeit with variations based on specific fields of study or certifications. Notably, the bar plot depicting average salary by gender reveals differences, which could be influenced by numerous factors such as industry dynamics and historical socio-economic factors.

Top 20 Industries and Job Titles by Average Annual Salary



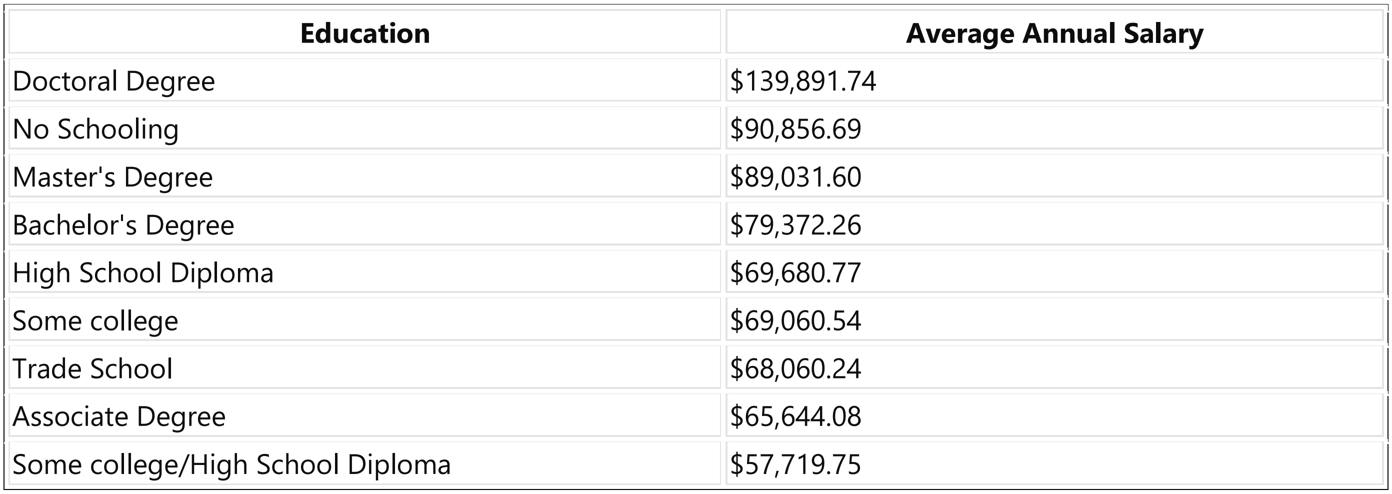
The top 20 job titles encompass specialized fields like healthcare, technology, and business leadership, with roles such as surgeons and CEOs ranking high. Creative professions like YouTubers also make the list, reflecting digital career viability. Meanwhile, the top-paying industries include tech-centric fields and traditional sectors like Law and Medicine, with transportation and warehousing showing high-value niches.

Most Common Job Titles and Industries amongst all Demographics



The survey highlights various job titles and industries, with managerial and analytical roles being the most common, indicating their prevalence in the surveyed population. Technical positions like "Engineer" and "Software Engineer" also show significant representation, reflecting the prominence of the tech sector. Industries such as "Computer Software" and "Health Care" lead in the respondent count, suggesting the dataset's focus on these sectors. The survey provides insights into various professions and economic sectors, possibly reflecting current job market trends and demographic reach.

Summary of Average salary by Education level:



Education	Average Annual Salary
Doctoral Degree	$139,891.74
No Schooling	$90,856.69
Master's Degree	$89,031.60
Bachelor's Degree	$79,372.26
High School Diploma	$69,680.77
Some college	$69,060.54
Trade School	$68,060.24
Associate Degree	$65,644.08
Some college/High School Diploma	$57,719.75

Average salary by education level reveals trends, with Doctoral Degree holders earning the highest at approximately $139,891.74. "No Schooling" ranks second at $90,856.69, suggesting alternative pathways to high earnings. Higher education generally correlates with higher salaries, but exceptions exist for those with a High School Diploma, some college or Trade School education. When examining salary by education level and experience, salary tends to increase with experience across education levels, but anomalies like significant salary spikes require further investigation.

Insights from Visualizations

Based on the analysis:

1. **High-Paying Careers:** Careers in specialized fields like healthcare, technology, and business leadership tend to offer competitive salaries. For instance, roles such as surgeons, anesthesiologists, CEOs, and CFOs rank among the highest-paying positions. Additionally,

creative professions like YouTubers and streamers have emerged as lucrative career paths in the digital age.

2. **In-Demand Degrees:** Doctoral degree holders typically command the highest average salaries, followed by individuals with master's and bachelor's degrees. However, unconventional education paths like trade school or no formal schooling can also lead to competitive wages, as evidenced by some outliers in the data.
3. **Geographical Considerations:** The United States is the dominant country in the dataset, with the highest number of respondents and potentially higher average salaries than other countries. However, countries like Canada, Australia, and the United Kingdom also offer competitive salary opportunities, albeit with relatively more minor representation in the dataset.
4. **Gender Disparities:** There are significant gender disparities in average salaries, with 'Male' and 'Female' participants typically earning higher average wages compared to 'Non-Binary' and 'Other' genders. This highlights the importance of advocating for pay equity and addressing systemic biases in compensation structures across industries and job roles.
5. **Career Trajectory:** Salary growth tends to correlate with years of experience, with individuals typically experiencing salary increases as they gain more seniority in their careers. However, it's essential to consider potential plateauing or slight declines in salary after a specific age range, which may necessitate strategic career planning to maximize earning potential over time.

Considering these insights, individuals exploring high-paying careers or educational paths may benefit from focusing on specialized fields, pursuing advanced degrees, evaluating salary prospects in different geographic regions, advocating for gender equity in compensation, and strategically planning their career trajectory to align with long-term financial goals.

Model Building:

Machine learning models are used for two primary purposes: gender prediction and salary prediction. Here’s an overview of the models used, their results, and critical insights:

1. X Gradient Boosting Model for Salary Predictions:

The prediction model employs an XGBoost to predict salary based on age range, years of experience, industry, job title, highest level of education, country, annual salary, annual bonus, and sign-on bonus. The results from the XGBoost model for predicting salaries using cross-validation (CV) and a hold-out test set provide a reasonable basis for evaluating the model's performance. Here's an analysis of these results:

Cross-Validation (CV) MSE Scores

- **CV MSE Scores**: The MSE scores from cross-validation are relatively consistent, ranging from approximately 446 million to 481 million. This indicates that the model is somewhat stable across different subsets of the training data.
- **Mean CV MSE**: The average MSE across all CV folds is about 468 million. This value gives a general idea of the model's prediction error across the CV process.

- **Standard Deviation of CV MSE**: The standard deviation of the CV MSE scores is approximately 12.98 million, which is relatively small compared to the MSE values' magnitude. This low standard deviation suggests that the model's performance is consistent across different folds, indicating a stable model that does not suffer too much from variability due to the randomness in the data splitting.

Hold-out Test Set Performance

- **Hold-out Test Set MSE**: The MSE on the hold-out test set is approximately 446 million, which is slightly better (lower) than the mean CV MSE. This suggests that the model generalizes well to unseen data regarding the MSE metric.
- **Hold-out Test Set R-Square (R^2)**: The R^2 value of 0.602 indicates that the model explains about 60.22% of the salary variance. This is a decent level of predictive power, suggesting the model has learned meaningful patterns from the features contributing to salary prediction.
- **Model Performance**: The gradient boosting model shows a good balance between bias and variance, as indicated by the consistency of CV MSE scores and the reasonable R^2 value on the hold-out test set. An R^2 of over 0.60 indicates a model that captures a significant portion of the variance in the target variable.

Potential Areas for Improvement:

Improving model performance through feature engineering, such as creating interaction terms, deriving new predictive features, or selecting features that most impact salary, may enhance predictive accuracy. Exploring advanced hyperparameter tuning techniques like Bayesian optimization or AutoML frameworks could also optimize model performance. Ensuring high data quality, potentially gathering more data, particularly for underrepresented groups, and further categorizing the Industry and Job Titles columns may improve the model's ability to generalize and make more accurate predictions.

2. Random Classification Model for gender prediction

This algorithm was selected due to its robustness and effectiveness in handling bias and variance through ensemble learning. Constructing multiple decision trees and voting for the most popular output class is generally good for classification problems with complex data structures and relationships. Random forests also handle overfitting better than many other algorithms and can deal with unbalanced datasets, which is advantageous given the gender imbalance typically found in such surveys.

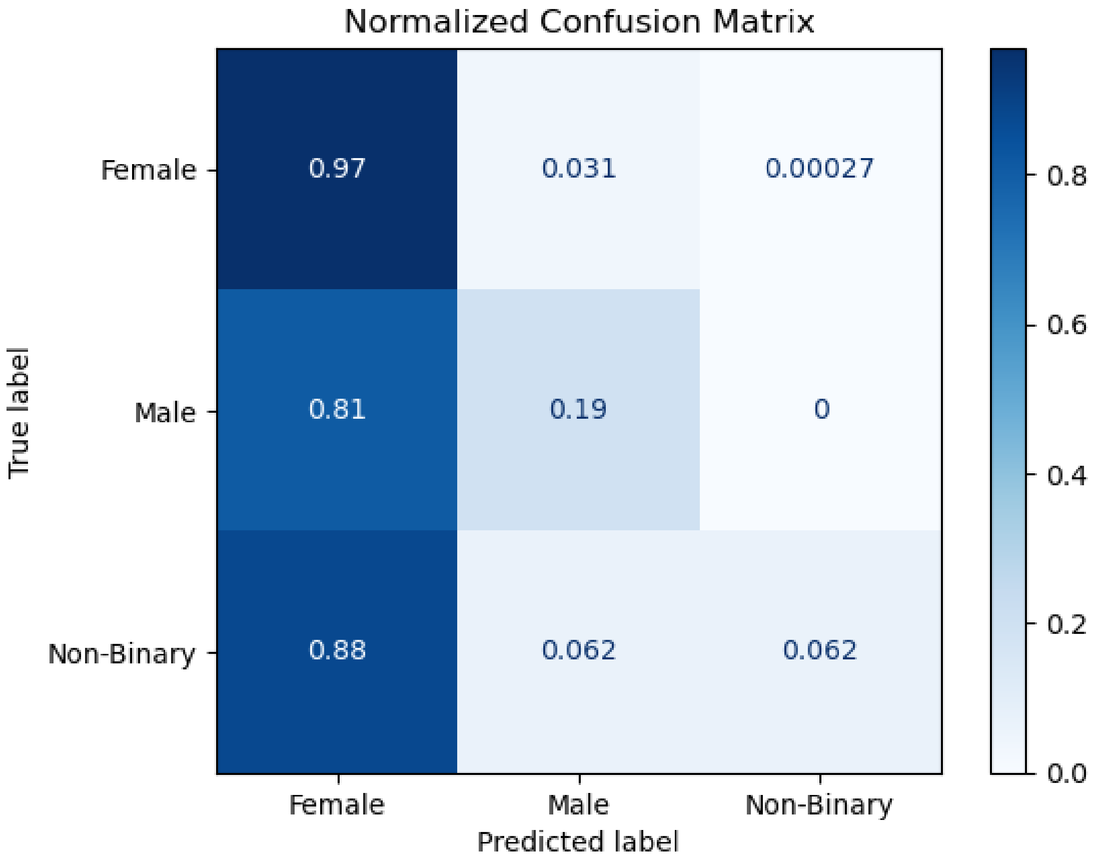
The prediction model employs a Random Forest classifier to predict gender based on age range, years of experience, industry, job title, highest level of education, country, annual salary, annual bonus, and sign-on bonus.

1. Summary of Key Findings: The model achieves an overall accuracy of 81%, indicating a reasonable capability to predict gender based on the provided features. However, the performance varies across different gender classes:

- **Female:** High precision (83%) and recall (97%), suggesting excellent model performance for this class.

- **Male:** Moderate precision (61%) but low recall (19%), indicating the model struggles to identify this class correctly.
- **Non-Binary:** Moderate precision (33%) but low recall (6%), showing the model's difficulty in identifying this class.

2. Visualizing the confusion matrix to elucidate where the model performs well and where it struggles



While the model performs well for the Female class, it exhibits challenges in accurately predicting the Male and Non-Binary classes. This is probably caused by the disproportionate distribution of gender in the dataset. With a significantly larger sample size for females (37,088) compared to males (9,320) and non-binary (81) individuals, the model is biased towards predicting the majority class (females) more accurately while struggling with minority classes (males and non-binary individuals) due to the limited amount of data available for training on these groups. One significant challenge was managing the imbalanced dataset for gender prediction. This was addressed by experimenting with different techniques, such as adjusting class weights and under-sampling. SMOTE techniques and other models were applied to mitigate this bias. Still, it worsened the model's overall performance, causing lower predictions for all classes except for the Non-Binary class.

3. Model's Practical Impact: The model shows promise in applications where accurately identifying the female gender is crucial. However, its limited ability to recognize other genders accurately could restrict its usability in scenarios requiring equitable treatment or understanding of all gender identities.

4. Recommendations for Improvement:

- **Balancing the Dataset:** Beyond SMOTE, I plan to explore other oversampling techniques for minority classes, undersampling for the majority class, or synthetically generating data for underrepresented genders could help improve model fairness.
- **Feature Engineering:** Exploring additional features or reevaluating the current feature set might uncover new insights and improve model accuracy across all classes.
- **Bias Mitigation:** The objective is to develop a model capable of accurately predicting genders—Female, Male, and Non-Binary—with high precision. Strategies will be implemented to detect and address bias, ensuring the model performs equitably across all gender identities.

Challenges and Learnings:

Throughout the project, various challenges emerged, including handling misspellings, categorizing industry and job title columns, calculating salary and country codes, and addressing class imbalance. Despite encountering setbacks, each obstacle presented an opportunity for growth and learning, emphasizing the iterative nature of data science projects. Creative solutions, such as keyword matching using lemmatization, fuzzy string matching, and NLP techniques, were employed to deal with misspellings and inconsistencies in the data. Despite efforts to automate the process, manual intervention was often necessary, underscoring the importance of data quality in the initial collection phase. Additionally, class imbalance posed difficulties in building predictive models accurately capturing the complexities of salary and gender, mainly due to the large number of jobs and industries and the disproportionate representation of female participants. Future optimization efforts will prioritize leveraging NLP and machine learning to address class imbalance and enhance industry and job title categorization, building on the lessons learned from this endeavor. Future iterations of this project can offer even greater insights into salary dynamics.

Conclusion:

While the project achieved its primary data cleaning and analysis goal, predictive modeling accuracy can be enhanced. By refining methodologies and exploring advanced techniques, such as NLP and machine learning, addressing the class imbalance, and refining the standardization of categorical variables, future iterations of this project can offer even greater insights into salary dynamics and higher accuracy and precision in predictive modeling.