Data Profiling of a Salary Survey Dataset

Adriano Machado
Faculty of Sciences, University of Porto
Porto, Portugal
up202105352@up.pt

1 INTRODUCTION

This report presents a data profiling analysis of a salary survey dataset , examining salary trends across various industries, job titles, and countries. Through detailed data quality assessment, three critical issues were identified and analyzed: data standardization challenges in the country feature, salary data formatting inconsistencies, and demographic representation imbalances. The analysis employs various profiling techniques, statistical methods, and visualization approaches to characterize these issues and their potential impact on subsequent analysis.

2 DATA PROFILING SUMMARY

The dataset, sourced from an AskAManager.org salary survey [2], comprises 28,085 records across 18 features, capturing demographic information, job details, compensation data, and geographical locations. Initial profiling revealed several noteworthy characteristics: The dataset exhibits a moderate level of incompleteness, with 17.2% missing values concentrated primarily in optional fields of the survey such as job context, additional compensation, and state information. The features span multiple data types, with a predominance of categorical variables (e.g., gender, race, country) and numerical fields (annual salary, additional compensation). Several fields accept free-text responses, introducing variability in data format and content. The dataset contains no duplicate entries.

To further understand the relationships between variables, we conducted an association analysis using Cramér's V[4], which measures the strength of association between categorical variables. The resulting association matrix is presented in Figure 1.

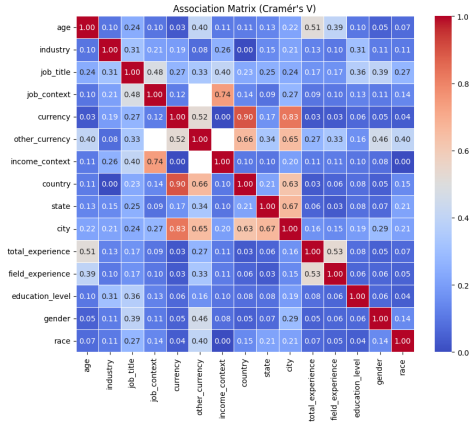

Figure 1: Association matrix of categorical variables using Cramér's V

The association analysis reveals the following key findings:

- Strong correlations (> 0.7):
  - Currency & Country (0.90), Job context & Income context (0.74), Currency & City (0.83).
- Moderate correlations (0.4-0.7):
  - Total experience & Field experience (0.53), State & City (0.67).
- Interesting patterns:
  - Demographic variables (gender, race) show weak correlations with professional variables, while geographic variables show moderate to strong correlations with each other.

3 KEY ISSUES AND ANALYSIS

Our data profiling analysis revealed three critical issues that significantly impact the quality and interpretability of the salary survey dataset. This section provides an in-depth discussion of each issue, its implications, and potential solutions.

3.1 Data Standardization Challenges in Categorical Variables

The dataset exhibits significant inconsistencies in categorical variables, particularly in the 'country' field. For instance, the 'country' column contains multiple variations for representing the United States (e.g., "USA," "United States," "US"), leading to fragmented data and potential underrepresentation of certain countries in analyses.

This lack of standardization complicates geographic analysis and can lead to misleading conclusions about salary distributions across countries. In our initial analysis, the United States had 9,337 entries, but after basic standardization, this number increased to 23,039 - a dramatic difference that could significantly alter any country-based comparisons.

To address this issue, we propose the following corrective measures:

- Text normalization techniques to standardize country names
- Creation of a mapping dictionary for common variations of country names
- Use of external reference data (e.g., country codes/names) to validate and standardize entries

3.2 Salary Data Formatting Inconsistencies

The 'annual_salary' field contains inconsistent formatting, with some entries using commas as thousand separators, and some having no separators at all. This inconsistency prevents straightforward conversion to a numeric data type for analysis. To solve this problem we could develop a preprocessing function that:

- Removes non-numeric characters from salary entries
- Converts the cleaned entries to a numeric data type

- Implements validation checks to ensure data integrity after conversion

3.3 Demographic Representation Imbalances

By analyzing the bar charts depicting gender and race distributions, we observed significant imbalances in demographic representation within the dataset. These visualizations provide a clearer understanding of the disparities present.

The first bar chart (Figure 2) illustrates the gender distribution in the dataset. It reveals that women account for a substantial majority of the entries, with 21,000 records compared to only 5,500 for men. This overrepresentation can skew analyses and may lead to conclusions that do not accurately reflect the experiences of men and non-binary genders in the workforce.

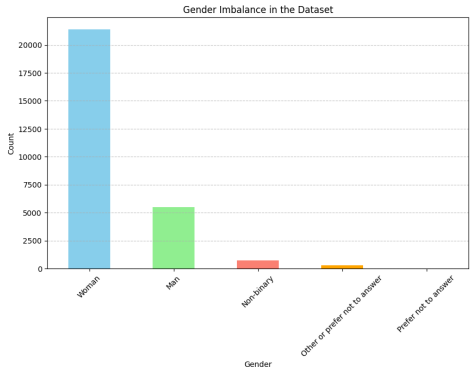

Figure 2: Gender Distribution in the Dataset

Similarly, the second bar chart (Figure 3) highlights significant racial representation imbalances. The data shows a predominant overrepresentation of White individuals, who make up 24,000 of the entries, while other racial groups are vastly underrepresented. This imbalance poses challenges for accurately analyzing salary trends, as the experiences of White individuals, particularly White women, may overshadow those of underrepresented groups.

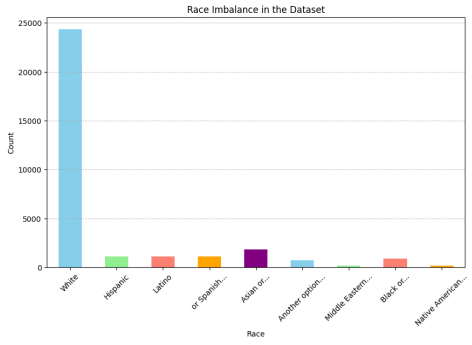

Figure 3: Race Distribution in the Dataset

These imbalances can lead to biased analyses and conclusions that may not accurately reflect the broader workforce. For instance, salary trends derived from this dataset might disproportionately represent the experiences of white women, potentially masking disparities faced by underrepresented groups.

To address this issue we could:

- Implement weighted sampling techniques to balance demographic representation in analyses
- Use stratified sampling methods when creating subsets for specific analyses
- Clearly communicate the demographic composition of the dataset in all reports and analyses
- Consider supplementing the dataset with additional data sources to improve representation of underrepresented groups

4 CONCLUSIONS AND RECOMMENDATIONS

Our analysis of the AskAManager.org salary survey dataset has uncovered several key issues that affect how we can use and interpret this information. The problems we found with data consistency and uneven representation of different groups mean we need to be careful about the conclusions we draw without some serious data cleaning and statistical adjustments.

Moving forward, we recommend adopting a data-centric approach. This means focusing first on improving the quality of the data itself, rather than jumping straight to analysis. We should start by developing robust methods to clean and categorize text inputs for job titles, locations, and other open-ended fields. Additionally, addressing the demographic imbalances in the dataset is crucial. This could involve techniques such as weighted sampling or supplementing the data with other sources to ensure a more representative picture of the workforce. By tackling both the data quality and representation issues, we can enhance the reliability and fairness of insights drawn from this survey, leading to more accurate understanding of salary trends across diverse job roles and demographic groups.[3]

REFERENCES

[1] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452. https://doi.org/10.1002/widm.1452 arXiv:2110.00530v3 [cs.LG] 21 Jan 2022.

[2] Ask A Manager. 2021. *How much money do you make?* https://www.askamanager.org/2021/04/how-much-money-do-you-make-4.html Online. Available: https://www.askamanager.org/2021/04/how-much-money-do-you-make-4.html.

[3] Molahajloo Shahla. 2019. *Imbalanced Data Visualization and Random Forest.* https://medium.com/@smollaha/imbalanced-data-visualization-and-random-forest-25cbff51f711 Online. Available: https://medium.com/@smollaha/imbalanced-data-visualization-and-random-forest-25cbff51f711.

[4] Maninder Singh. 2023. *Understanding Categorical Correlations with Chi-Square Test and Cramer's V.* https://medium.com/@manindersingh120996/understanding-categorical-correlations-with-chi-square-test-and-cramers-v-a54fe153b1d6 Online. Available: https://medium.com/@manindersingh120996/understanding-categorical-correlations-with-chi-square-test-and-cramers-v-a54fe153b1d6.