Ask A Manager Self-Reported Job, Salary, and Demographic Data

ANNA SANDERS, University of Colorado Boulder, USA

The Ask a Manager advice blog runs an annual salary survey to gather data on their reader base and help demystify salaries across industries and job roles. This project cleans and tidies survey results from 2022 and 2023. There is additionally some data analysis and visualization done on the cleaned dataset. This project then attempts to reduce the number of unique Job Titles, a free text field in the survey, in order to use the field as a categorical variable in modeling. A filtered version of the final dataset, including the job title clustering results, is then used to train a model to predict salary based on various attributes and evaluated for accuracy.

Additional Key Words and Phrases: Data Mining, Data, Data Analysis, Salary Data, Jobs Data, Survey Data, Natural Language Processing, Prediction

1 INTRODUCTION

Ask a Manager [4] is a blog focusing on answering reader's advice on job related questions, especially advice with the perspective of a 'manager.' In addition to answering questions and giving advice, Ask a Manager also engages with their community by creating open forums for discussion and surveying readers. One survey that Ask a Manager releases annually is a self reported salary survey. This year's survey, 'how much money do you make?, as well as most previous years' survey, is released around the end of April. The survey itself is a Google Forms with various questions on the respondent's job, demographics, salary, and other potentially relevant information. The raw results are publicly available in a corresponding Google Sheets document.

While, in the United States, it is legal and protected by the National Labor Relations Act (NLRA), discussing salary is still a taboo subject in the workplace. As job salary and benefits are often negotiable, it is important for workers to understand the so-called 'market rate' other companies are paying for similar jobs and functional roles; however, jobs are very difficult to compare, even when just taking into account job title and years of experience. Some companies may add expectations outside of the common definition of a job title; some companies may promote individuals at different rates than other, similar companies; some companies might title their job role as an 'analyst' while similar jobs at another company might be a 'consultant.' Similarly, some companies or industries may discriminate based on factors, like age and race, which may not be present in current salary data sets and tools.

1.1 Related Work

Online websites like Glassdoor [1], Indeed [3], and Payscale [2] collect and report on salary information, but this information is often only stratified by years of experience and does not include the raw user responses.

The Ask A Manager's salary dataset differs from current pay and salary related websites. For one, the raw data is available to view and download. The survey itself is mainly categorical or ordinal variables; for instance, age instead of being an integer response is a category with different age bands. Job title is a completely free text field, while industry and functional area have selectable options as well as a free text response. As the dataset contains dimensions not usually found in salary websites, further analysis and visualizations can be done regarding those attributes. Additionally, a cleaner version of the data can be used for further analysis and modeling outside of this project.

For these reasons, it is important to have well representative, clean, and complete job and salary data available for data analysis, visualization, and modeling to aid in salary negotiations and to reduce potential wage discrimination.

Author's address: Anna Sanders, University of Colorado Boulder, 2055 Regent Dr., Boulder, Colorado, USA, 80310, anna.sanders@colorado.edu.

2 PROPOSED WORK

This project seeks to increase the amount of clean job and salary related data publicly available for analysis and modeling, as well as complete basic analysis of the survey data and attempt to create a model for predicting salary based on various dimensional attributes.

This project will be completed in Python using various data analysis, visualization, and data science packages. The data itself will be read from a downloaded version of the Google Sheets data in comma separated values (csv) format.

2.1 Scope

In order to keep this project's scope manageable for a 6 week, 1 credit program, only the 2023 and 2022 survey responses will be cleaned, analyzed, and modeled. This is done to ensure only the most recent data is being analyzed and that any potential merging issues between the individual surveys are kept to a minimum. Additionally, using two survey's data allows for analysis on year over year trends and expands the dataset by around 100%.

Overall, the dataset has over 30,000 responses with 26 available columns, which should be sufficient for data analysis, visualization, and modeling.

2.2 Cleaning

The data was loaded into separate pandas dataframes for each survey and cleaned individually to handle any unique survey issues that arose. The main packages used in the cleaning were pandas and numpy.

The data needs to be cleaned and checked for potential errors. Any NULLs or NANs for important categorical variables including age, reported salary, job title, industry, and others, will be dropped from the dataset. NULLs and NANs for less important categorical variables including, city, state, country, gender, and race, will be filled in with an 'Unknown' value. Additionally, some free-text and categorical responses will be manually cleaned to facilitate the data cleaning process. The surveys included some multiple-response questions, including allowing multiple responses for industry, race, industry, and others. In order to keep the scope of this project in line, cleaning will include keeping only the first response in these multi-response questions. The salary and bonus columns need to be cleaned, as some responses included commas and other monetary separators; both the salary and bonus will be summed to create

a new 'total salary' column, that represents the total monetary compensation for a respondent. Important string columns will be stripped of space padding before and after the text, and will be capitalized to ensure that exact match text responses will be grouped together during analysis and modeling. Lastly, ordinal categorical variables, including age and experience, will be mutated into a separate column with a numerical representation of the category matching the numerical order the category falls into.

After cleaning, the individual datasets were merged to create a combined dataset. The different surveys were specified with a 'year' column to indicate the year the survey results were from.

In addition to cleaning, for industry and functional area, any category with less than 10 responses was dropped from the dataset in order to limit the encoded columns used during modeling. Since these questions had selectable answers, the total number of categories with less than 10 responses was less than 10% of the data.

From the cleaning process, less than 400 rows were removed. The final dataset has 29,714 responses with 26 columns.

2.3 Clustering by Reported Job Title

In the dataset, there are over 14,000 unique job titles entered in the survey and over 19,000 have less than 10 responses per job title. The low number of responses per unique job title would make forecasting difficult and could potentially over or under forecast for certain jobs with a low number of responses.

Using the scapy python package, job titles were first converted into vector norms, which were then appended to the dataset. The K-Means cluster algorithm was used on the first 1,000 rows of the dataset to create 20 clusters and was then anecdotally checked for membership similarity. Additionally, the sci-kit learn package's pipeline method also included a Feature Hasher function that was able to convert strings into vectors with a hard-coded number of features, which could be run through models as inputs. This method was also tested on the K-Means cluster algorithm and seemed to perform better than the vector norm. Going forward, this method was used to vectorize the string job titles for clustering.

In the first pass of clustering, multiple attributes, including a combined job title-industry-functional area string vector, were used in addition to job title to create clusters. This proved to be problematic because unique job titles were being clustered into separate clusters, which would likely cause more issues when trying to use the clusters to predict salary.

The clustering was then rerun with only the job title as an input. In addition to the K-Means algorithm, Birch and OPTICS clustering models were used on the first 1,000 rows of the dataset and then anecdotally checked for membership similarity. Birch performed similarly to the K-Means. OPTICS performed slightly better, but as the hyperparameter was minimum samples in a cluster, some job titles were not classified and instead marked as outliers. This was not ideal for using the clusters in further modeling and prediction.

The full dataset was then clustered; there was no train test split done because there were no metrics other than anecdotal checking of individual clusters that could improve the model. Similarly, there was only one input to the model, so no analysis needed to be done to change or remove parameters. The K-Means clustering algorithm was set to find 2,000 clusters while the OPTICS model had the minimum number of samples set to 3. Both cluster results were then checked to ensure job titles were only in exactly one cluster. The K-Means cluster results had 303 clusters with only one member, which is not ideal for future training and testing of models, but is better than the over 19,000 unique job titles. The OPTICS model clustered the data into 544 clusters, but over 80% of the data was marked as outliers.

Ten random clusters were selected from the K-Means cluster to be observed. The first member of each cluster was then found in OPTICS clustering results and, if a cluster was found, compared to the K-Means cluster. Out of the 10 first job titles in the K-Means clusters, only 2 were found in OPTICS, but the OPTICS cluster was better grouped than the K-Means cluster.

2.3.1 Evaluation.

Below is cluster 1051 in the final K-Means clustering:

- MANAGER DEI CORPORATE PARTNERSHIPS
- PRE-AWARD RESEARCH ADMINISTRATOR
- PRINCIPAL ENTERPRISE PROJECT MANAGER
- RECREATION SPORTS PROGRAM MANAGER
- SENIOR CORPORATE PHILANTHROPY MANAGER

This cluster's job titles do not seem to connect with each other at all. This is likely due in part because of the inclusion of 'MANAGER' in the job title, and that the combination of words in the title are similar enough to each other that they were clustered together. Other clusters' job titles are more similar to each other, but clearly the K-Means clustering is not perfect.

Below is cluster 362 from the OPTICS clustering:

- HIMAN RESOURCES BUSINESS PARTNER
- HUMAN RESOURCES BUISNESS PARTNER

and a subset of cluster 375 from the K-Means clustering (9 total job titles):

- HIMAN RESOURCES BUSINESS PARTNER
- ...
- SENIOR HUMAN RESOURCES BUSINESS PARTNER
- SENIOR QUALITY ASSURANCE SUPERVISOR

Compared to the cluster 375 from the K-Means clustering, the OPTICS cluster picks up less than the K-Means; however, the K-Means cluster also picked up job titles that are clearly not related to human resources. The K-Means cluster is less than ideal, but is the only algorithm that clustered all job titles. From this point on, only the final K-Means cluster results were used for analysis and modeling.

2.4 Data Analysis and Visualization

Data analysis and visualizations was completed on various subsets of the cleaned and clustered data, taking into account country, currency, and survey year. Due to outliers, the visualized dataset will be limited to salaries below $600,000 when visualizing salary. The data analysis and visualization is broken into three sections: All Data, which includes all data in the cleaned dataset; USD Only, which only includes responses where the currency is USD; and 2023 USD Only, which only includes data from 2023 reported in USD. These groupings were created to ensure most if not all dimensions were

represented in a visual or analysis, but was not duplicated between groups.

All Data includes pie charts breaking down total responses by country and currency. USD Only includes pie charts for total responses by industry and functional area, a violin plot of total salary (salary + bonus) by year, bar charts by year and total salary animated by industry and functional area, and a statistical test for differences in total salary which includes a histogram of total salary by year. 2023 USD Only has pie charts for total responses by state and city, box plots of total salary by age group, work experience, field experience, work type, and gender, as well as statistical test for work type and gender across median total salaries.

There are a few key results from the data analysis and visualizations that were taken into account when building the prediction model. The first key result was the distribution of responses by currency. Because of changing exchange rates, it's difficult to compare salaries between different currencies.

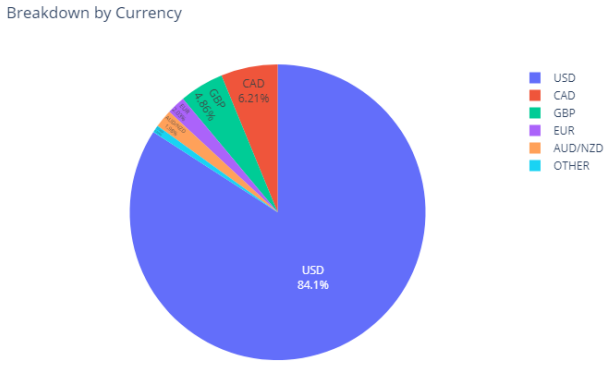

Fig. 1. Pie Chart Breakdown by Currency

As displayed in Fig 1, most of the responses are in USD, so we can limit the scope of the dataset for analysis and modeling without sacrificing many responses.

Another key result is the histogram of total salaries under $600,000 (USD Only).

We can see in Fig 2 that even when correcting for salaries, total salary is mainly distributed around $100,000, but many single outliers exist between $400,000 and $600,000. This is important to keep in mind when modeling total salary, as the predictions will likely still be skewed in some way due to the outliers.

The last key result is from statistical analysis of 2022 salaries vs. 2023 salaries. Using the Student's t-test and Mood's Median Test, when comparing the mean or median of 2022 salaries vs. 2023 salaries, the difference in sample means and medians are statistically significant (p-values > 0.05), or that 2023 salaries are statistically different than 2023 salaries.

Another interesting result can be seen when plotting total salary by gender.

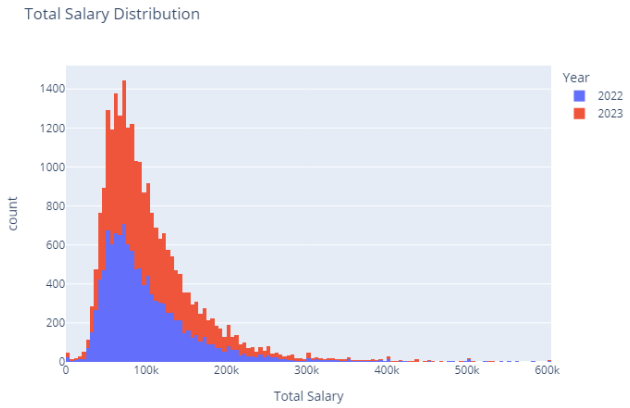

Fig. 2. Histogram of Total Salary

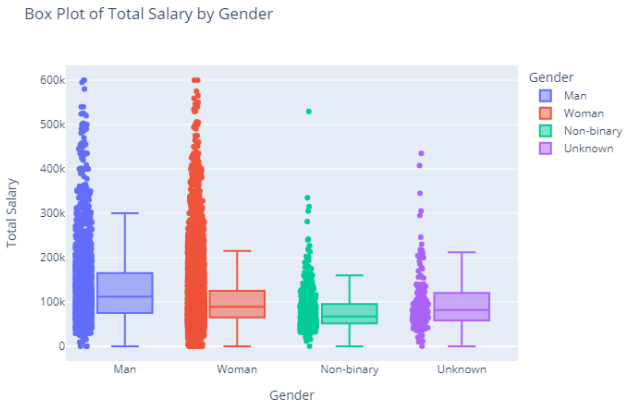

Fig. 3. Histogram of Total Salary

From Fig 3, it seems, at least visually, that men have a higher median salary and have a higher 3rd quartile salary than women and non-binary respondents. However, we cannot make any conclusions based on this visualization alone. Firstly, like any survey, there is likely a response bias; men who make smaller salaries may not have responded to the survey. Secondly, there is no differentiation between part-time and full-time jobs in the survey. Women sometimes will seek part-time jobs in order to care for their children during adolescence. Lastly, there are salary differences between functional area (e.g. computing and tech, sales, accounting), and while having unequal representation of genders in and between functional areas, the apparent lower median salaries may not be wholly based on gender.

2.5 Predicting Salary

2.5.1 Setup.

Due to outliers, currency conversion issues, and differences in salary

from 2022 to 2023, only responses in USD from 2023 under $600,000 were used in the training and testing of the algorithm. The results of the K-Means cluster from the previous section will be used in the model. A pipeline was built to standardize the ordinal variables (age, experience, field experience, and education) and encode the categorical variables (industry, functional area, work type, state, gender, and race). Unfortunately, because some of the K-Means clusters had only one member, the pipeline was run before the train-test split. This is not ideal because the training and testing data should have at least one data point from each category. The train-test split was 70:30.

The following regression models were run against the transformed data: Linear, Decision Tree, Kernel Ridge, Random Forest, GLM, Stochastic Gradient Descent, Support Vector Machine, and Gaussian Process.

From the models run, the two most promising models were the Random Forest Regressor (RFR) and the Stochastic Gradient Descent (SGD).

2.5.2 Evaluation.
The models had the following associated metrics:

Metric	**RFR**	**SGD**
R^2	0.4	0.45
Explained Variance	0.41	0.45
Mean Absolute Error	31,805.32	31,829.85
Mean Squared Error	2,455,559,425.23	2,257,590,238.32
Mean Absolute Percent Error	109.46%	120.81%

The SGD model performed better in the 'fit' metrics, or the R^2 value and Explained Variance metrics, where higher numbers indicate a better model fit. The RFR model performed better in the 'residuals' metrics, or the Mean Absolute Error, Mean Square Error, and Mean Absolute Percent Error, where smaller values indicate that the model predicted the testing data better. It is important to keep in mind that with these metrics, neither of the models are 'good'. A good model would have an R^2 value at or above 0.8, and the residuals should be as close to zero as possible.

We can investigate potentially why these models are not performing well by plotting the predicted values to the actual values.

Fig 4, which plots the RFR predicted vs. actuals, shows that at lower actual salaries, the model tends to overestimate, while at higher actual salaries, the model underestimates. At even higher values of actual salary, the model is very bad at prediction, as the points get further and further away from the line $y = x$ or where $actual = prediction$.

Fig 5, which plots the SGD predicted vs. actuals, shows similar patterns to Fig 4, but noticeably especially at higher values of actual salary is less variable from the line $y = x$ or $actual = prediction$ than the previous figure. This would explain why the SGD model had better fit metrics than the RFR model.

Both models seem to struggle as actual salaries get higher, which is unsurprising given the amount of 'outliers' or the low number of responses with a high total salary. The poor fit and high residuals indicate that there are many improvements that can be made to the model and setup, but overall validates the real challenges modeling and predicting salary, especially from this dataset, outlined in the introduction.

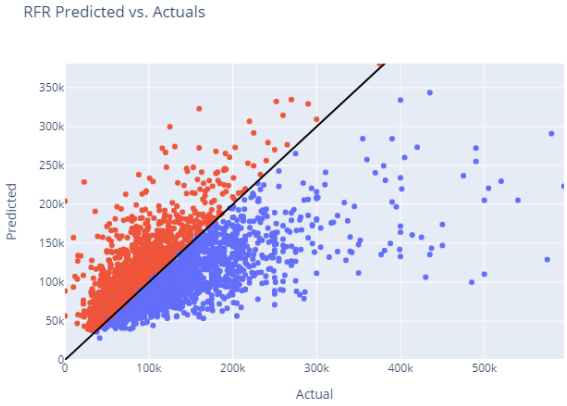

Fig. 4. RFR Predicted vs. Actual Values

Fig. 5. SGD Predicted vs. Actual Values

3 FURTHER DISCUSSION

3.1 Timeline

This project was completed in one 6 week long semester. All proposed work has been completed and documented in this proposal as well as in presentation slides and a personal write up including code. Some features and analysis were left as potential next steps in order to ensure the project is validated and completed by the end of the semester. Most of the time was devoted to cleaning and tidying the data, but significant time was also spent on clustering and modeling the data; peripherally, a significant amount of time was taken up by troubleshooting python packages and anaconda python environments. The project proposal and slide development was not taken into account in the timeline, but took less than 2

days per iteration, with three iterations in total for the proposal and presentation slides.

A the timeline was as follows:

Table 1. Project Timeline

Proposed Work	Time Estimate	Current Status
Data Cleaning	3 Days	Complete
Job Title Clustering	5 Days	Complete
Data Analysis & Visualization	5 Days	In Progress
Salary Prediction Model	7 Days	On Deck

3.2 Evaluation Plan

Success of this project depended on the completion of each section proposed above. While it would have be great if the cluster and model were both statistically significant and successful, multiple models and methods are attempted and reasoning for why the final models were selected were reasoned above, so non-significant model do not discount the success of the project.

3.2.1 Lessons Learned.

On a more personal note, this project was a great first step in completing a machine learning pipeline, starting from finding the data and generating hypotheses and questions to answer, to creating the pipeline and modeling the data. In the CU Boulder MS in Data Science, I have mainly been performing statistical analysis and modeling in R, so being able to complete this project in python has helped to increase my confidence and knowledge of completing key data tasks in machine learning.

A big lesson learned was how to troubleshoot various python errors. I use python through anaconda and jupyter notebooks. There were many points in this project, especially during the clustering work, where a promising clustering model would exist and I would forcibly download the python package, which would in turn cause the various other python packages and dependencies to fail.

3.2.2 Key Takeaways.

Key personal takeaways from this project include how to set up the sci-kit learn machine learning pipeline, how to extract features and predict models in python, and how to write up and present a full data mining pipeline project with machine learning.

Key takeaways from the project mainly revolve around the distribution of data. While there is clearly a lower limit to salary (0), there is truly almost no upper limit to salary, which makes prediction very difficult. At what point should a model not take into account high salaries? Similarly, job titles also also very unique, and in this dataset are almost too unique to use in analysis and modeling.

The data itself is still fairly skewed to North American respondents, as the blog is in English and writes mostly about the western workforce. Similarly, as it is a job advice blog, the overall data seems to skew more towards seemingly full-time career jobs, rather than part-time, seasonal, or temporary jobs. As the median job salary for both 2022 and 2023, $85,000 and $90,500 respectively and limited to salaries under $600,000, is well above the national median of $70,784 as of the 2021 census [5], it is likely that the blog, and thus the survey, attracted more high-earners; for this reason, regardless of the accuracy of the predictions, the model itself would likely be positively skewed when predicting salary from this dataset.

3.3 Potential Challenges

There were some potential challenges in completing this project. There are many different models to use for clustering and predicting, some of which have performance limitations. It is difficult to test all combinations of data formatting and models available, so only certain models were tested and only one data transformation was used in the models. The project was ambitious in the scope of data mining practices completed in this project, but was completed within the proposed time frame.

Backup plans included limiting the scope of the data analysis and visualization if time does not permit a full-scale analysis.

There were no major issues or setbacks that necessitated the backup plan or thrown the timeline estimates off track. Despite the lack of setbacks, the job title clustering ended up being less accurate than first anticipated, potentially because there was no easy way to include important secondary dimensions like functional area and industry without the potential of duplicate job titles between clusters. Similarly, the final prediction model also was less accurate than hoped.

4 CONCLUSION

In conclusion, this project attempted to clean and tidy data from public survey results, perform data analysis and create visualizations, as well as attempt to better cluster and predict the main response variable, salary. This project should increase the amount of usable salary data that other applications and programs can use to better estimate average wages, which in turn can be then used to help support salary negotiations and other job related negotiations. While this project's own clustering and salary prediction models may not be wholly significant, they both aim to begin to solve the inherent problems when dealing with salary data, including how best to group similar job titles and predict total salary. This project should also expand the knowledge and discourse reported job and salary data by identifying trends and potential predictors and their importance relating to overall salary.

4.1 Future Work

Decisions were made throughout the project in order to keep a limited and manageable project scope. Below are some limitations and suggested future work that could be completed after their removal:

- Add more survey data: Due to time limitations, only two years of survey data was used. There are more surveys prior to 2022 that could be included in analysis and modeling.
- Allow for multi-response items: The final dataset was tidied to only include one response if multiple responses were listed. Multiple responses could improve data quality and lead to better modeling results
- Use neural networks to classify job titles: Neural networks can be used to better group like job titles

- Further Analysis: More analysis could be done to answer if there is a significant difference in total salary depending on gender
- Predict Base Salary: Total salary (salary + bonus) was used as the response variable. It's possible that some total salaries are high because of a large bonus
- Test More Transformations and Models: Only a select few prediction models were tested and only one data pipeline was used. Future work could include further manipulation and transformation of the data and a larger set of models to train and test
- Remove Clusters with Less Than 2 Members: All data was kept in the prediction model in order to keep the cleaning data loss at a minimum. Further improvements could be made by explicitly removing the single job title clusters and the job title itself from the clustering process

ACKNOWLEDGMENTS

Thank you to Ask a Manager for running this survey and making the data publicly available. Ask a Manager is a great resource for those entering the workforce, as they often advise on salary negotiation, networking and career progression, and dealing with difficult work situations. I highly suggest checking out the blog if unfamiliar with it, as it's a great way to gain perspective entering the workforce or starting a new job or role.

REFERENCES

[1] 2023. *Glassdoor Salaries.* Retrieved October 2023 from https://www.glassdoor.com/Salaries/index.htm
[2] 2023. *Payscale.* Retrieved October 2023 from https://www.payscale.com/
[3] 2023. *Salaries.* Retrieved October 2023 from https://www.indeed.com/career/salaries
[4] Allison Green. 2023. *Ask a Manager.* Retrieved October 2023 from https://www.askamanager.org/
[5] Jessica Semega and Melissa Kollar. 2021. *Income in the United States: 2021.* Retrieved October 2023 from https://www.census.gov/library/publications/2022/demo/p60-276.html