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

Anna Sanders
DTSA 5506 - Data Mining Project
Fall 2023

Salary Transparency

“Under the National Labor Relations Act (NLRA or the Act), employees have the right to communicate with other employees at their workplace about their wages. Wages are a vital term and condition of employment, and discussions of wages are often preliminary to organizing or other actions for mutual aid or protection. “ - NLRA

Benefits of Salary Transparency:

- Salary Negotiation
- Benefit Negotiation
- Job Searching
- Non-Discriminatory Wages

Average Wage Websites (Glassdoor, Indeed, Payscale)

Benefits

- Location specific
- Employer specific
- Experience specific
- Easy to use and navigate
- Includes bonuses and other monetary compensation

Drawbacks

- No access to raw data
- Potentially no access to additional demographic data (gender, age, etc.)
- Potentially no access to additional job data (industry)
- Logins needed to view more data, submit data

Ask A Manager Salary Data (2022 & 2023)

Raw data in csv format from the Ask A Manager Annual Salary Survey:

- Includes respondent demographics (age, race)
- Includes industry and functional area dimensions
- 30,000+ total responses

Challenges

- Self-reported responses
- Job Title field is free text
- Some fields are multi-response
- Slight variation between the surveys

Proposed Work

Data Cleaning & Tidying

- Change categorical columns to ordinal columns when appropriate
- Clean string responses by removing padding and capitalizing
- Select only first response value for multi-response columns
- Drop columns with missing data (age, experience, salary, etc.)
- Create an ‘Unknown’ category where appropriate (gender, race, etc.)
- Add total salary column (salary + bonus)
- Manually correct some free-text responses into specific categories
- Merge survey results into one dataframe

Clustering Algorithm for Job Title

Job Titles to Vectors

- Scapy: Job Titles → Vector Norms
- Sci-kit Learn: Job Titles → Vectors

Testing

- Subset of the first 1,000 rows
- Cluster and check results

Models

- K-Means
- Birch
- OPTICS

Full Dataset

- Cluster all data
- Check for duplicates
- Check a random subset of clusters

Clustering Algorithm - Lessons Learned

First Try

Included other dimensions, including combined job title-industry-functional area vector, salary, etc.

- **Pros:** promising job clusters
- **Cons:** Caused unique job titles to exist in multiple clusters

Final Process

Use only job title. No other dimensions were included

- **Pros:** forces unique job titles to exist in only one cluster
- **Cons:** groupings only reliant on job title vector or vector norm, some confusing groupings

Clustering Algorithm - Results

K-Means

- 2,000 clusters
- 303 clusters with only 1 member
- Clusters could be more specific

OPTICS

- Minimum membership of 3
- 544 clusters
- 82.22% of data labeled as outliers

```
Cluster: 1051 ['MANAGER DEI CORPORATE PARTNERSHIPS' 'PRE-AWARD RESEARCH ADMINISTRATOR'
 'PRINCIPAL ENTERPRISE PROJECT MANAGER'
 'RECREATION SPORTS PROGRAM MANAGER'
 'SENIOR CORPORATE PHILANTHROPY MANAGER']
```

```
Cluster: 544 ['DIRECTOR OF EQUITY, DIVERSITY & INCLUSION'
 'DIVERSITY, EQUITY AND INCLUSION DIRECTOR'
 'HEAD OF DIVERSITY, EQUITY, AND INCLUSION']
```

Data Analysis & Visualization

All Data

- Percent of responses by country
- Percent of responses by currency

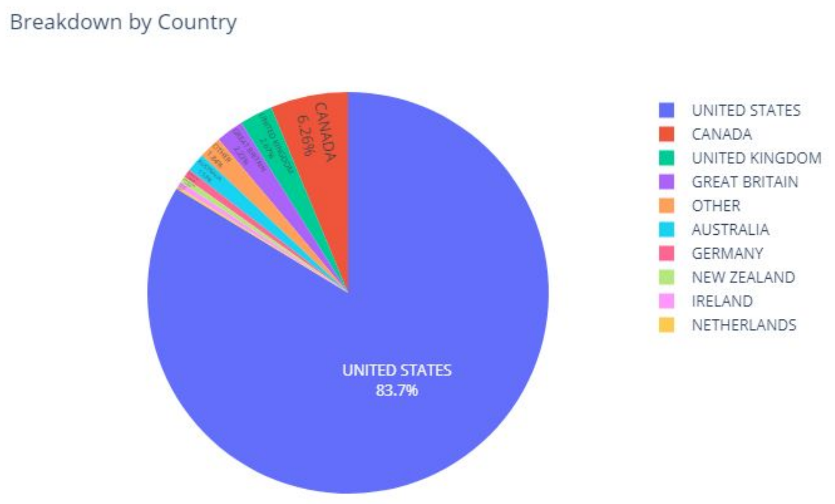
USD Only

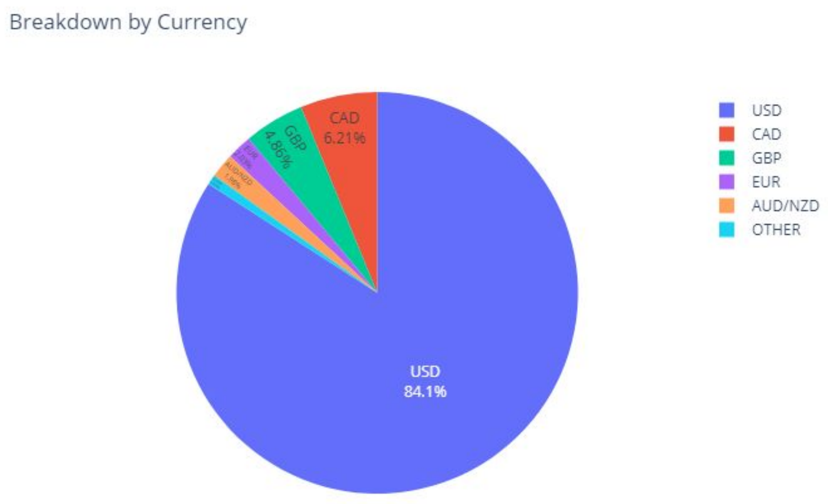
- Percent of responses by industry
- Percent of responses by functional area
- Total salary in 2022 vs. 2023

USD 2023 Only

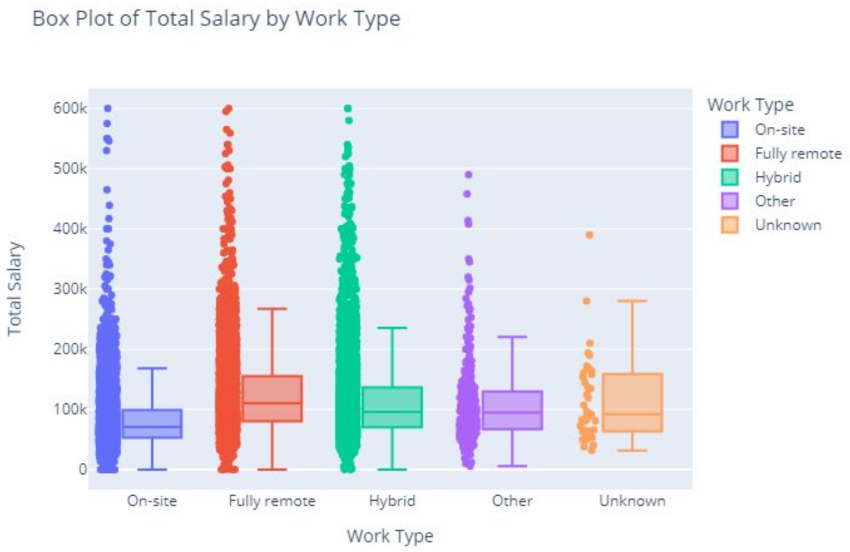
- Percent of responses by state
- Percent of responses by city
- Breakdown of total salary by age
- Breakdown of total salary by experience
- Breakdown of total salary by gender

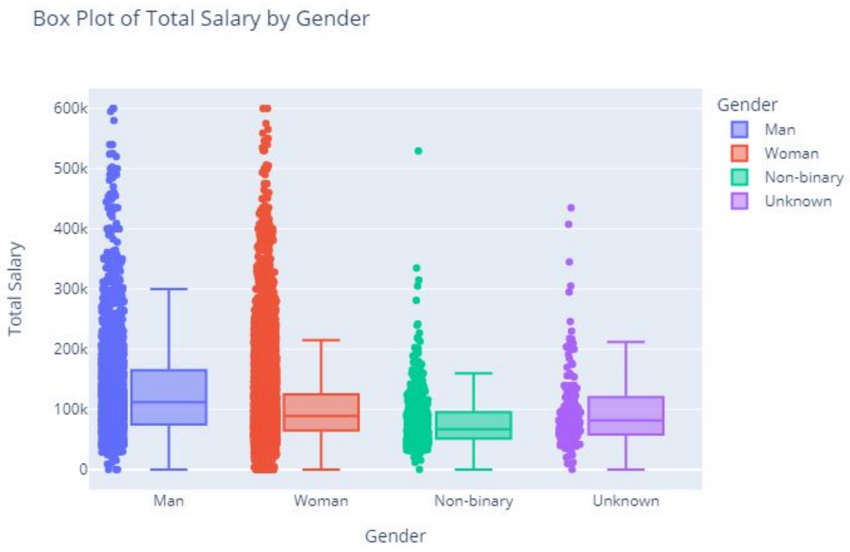
Breakdown by Country and Currency





Total Salary by Work Type and Gender





*total salary under $600,000

Total Salary Prediction Model

Setup

- Pipeline Transformation
  - Standard Scalar for Ordinal Variables
  - OneHotEncoder for Categorical Variables
- Train and test split (70:30)
- Run over multiple models
- Calculate metrics for all models
- Select the best model

Regression Models

- Linear
- Decision Tree
- Kernel Ridge
- Random Forest
- General Linear Model
- Stochastic Gradient Descent
- Support Vector Machine
- Gaussian Process

Total Salary Prediction Model - Results

	Random Forest Regressor	**Stochastic Gradient Descent**
R^2	0.40	0.45
Explained Variance	0.41	0.45
Mean Absolute Error	31,805	31,829
Means Squared Error	2,455,559,425	2,257,590,238
Mean Absolute Percent Error	109.46%	120.81%

Total Salary Prediction Model - Results

Total Salary Prediction Model - No Job Cluster

	Random Forest Regressor	Stochastic Gradient Descent
R^2	0.36	0.39
Explained Variance	0.36	0.40
Mean Absolute Error	33,047	33,018
Means Squared Error	2,606,408,982	2,469,916,673
Mean Absolute Percent Error	109.91%	114.84%

Evaluation

Timeline

Project Start: October 9th

Finish Data Cleaning (October 12th) - 2 days	**Status:** Done!
Finish Job Title Clustering (October 16th) - 5 days	**Status:** Done!
Finish Data Analysis & Visualization (October 23rd) - 5 days	**Status:** Done!
Finish Salary Prediction Model (October 30th) - 7 days	**Status:** Done!

Evaluation Plan

In general, a successful project will have completed all proposed work and included reasoning for decisions and potential downstream consequences, and will thoroughly document all work done, including:

- Data cleaning procedures and methodology
- Creation of visualizations
- Explanation of data analysis and hypothesis testing
- Testing and evaluating multiple models for the clustering and predicting processes
- Complete write up and presentation slides updated with high level processes and findings

Evaluation Plan - Models

Job Title Clustering: Cluster results viewed and assessed on a heuristic basis

Salary Prediction Model: Models evaluated with residual and fit based metrics, potentially unique to each model used

Prediction Model Metrics:

- R^2 - higher is better
- Explained Variance - higher is better
- Mean Absolute Error - lower is better
- Mean Squared Error - lower is better
- Mean Absolute Percent Error - lower is better

Evaluation Plan - Assessment

Overall, the project was successful because:

- All proposed work was completed
- All work was documented in the project proposal, presentation slides, and secondary write-up
- Data was cleaned and tidied and would be usable in other projects and analysis
- Multiple visualizations and analyses were completed
- Multiple models were run for clustering and prediction

Evaluation Plan - Reflection

Lessons Learned

- Machine learning pipeline in python
- Generating hypotheses
- Statistical analysis in python
- Troubleshooting python errors

Key Takeaways

- Lots of outliers in the data
- Unique job titles
- Data skewed:
    - Higher earners
    - North America

Future Work

- Add more survey data
- Allow for multi-response items
- Use Neural Networks to classify job titles
- Further analysis and visualization
- Predict base salary only
- Test more transformations and models
- Find a way to remove clusters with less than 2 members