Data Job Salaries

Programming for Data Science 1007

Junwen Fang, Lyric Li, Rosy Xu

Our **primary goal** is to investigate the relationship between employers' salary and their personal information such as experiences, job titles, locations and so on. So that we are able to have an overview understanding on salaries of different data related job to make our best choice for our future job.

The dataset contains salary information from professionals all over the world in the AI, ML, Data Science space. **It contains 10 columns** with the following characteristics: work_year, experience_level, employment_type, job_title, salary, salary_currency, employee_residence, remote_ratio, company_location, company_size.

Our data is gained from Data jobs salaries - weekly updated

Dataset

NYU

Working process

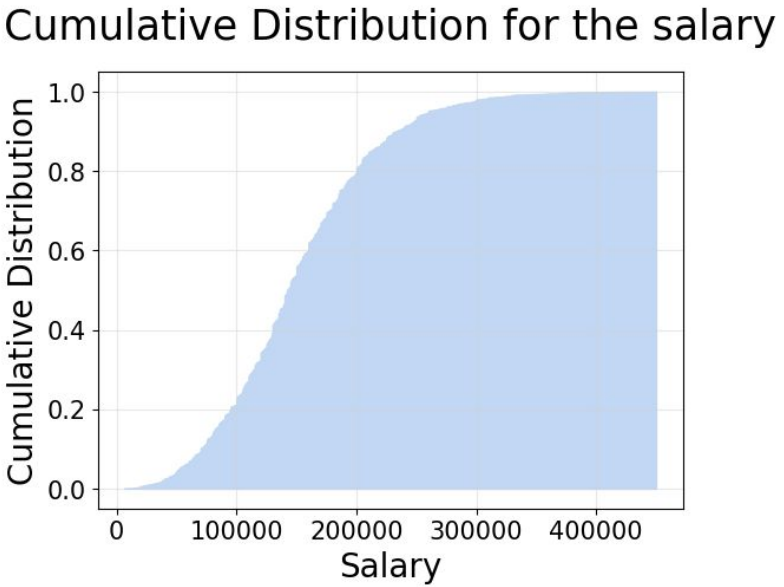
1. Data Processing
    - Clean job titles and use **pandas series.apply()** to divide jobs into 17 categories
    - Redefine remote ratio values as character objects by **pandas.replace()**
    - Standardize the unit of salaries to USD using real-time exchange rates API and **pandas.div()**
2. Exploratory Data Analysis
    - Visualizing data and answering questions by **matplotlib** (bar plot, line graph, pie graph, cumulative frequency, heatplot, histogram)
3. Machine Learning – Build a Model with the help of **numpy and sklearn**
    - Data Preparation (Train-test split and applying one-hot encoding by **dictionary and list** to transform categorical data)
    - Model Selection (Linear Regression Model, Linear Regression with regularization, Regression Tree, LGBMRegressor)
    - Model Training (train-test split)
    - Model Evaluation (Mean Squared Error, Mean Absolute Value, R squared)

Exploratory Data Analysis

	Column	Correlation to Salary
6	company_location	0.759412
4	employee_residence	0.574568
5	remote_ratio	0.020400
0	work_year	0.019457
7	company_size	0.016803
8	job_category	0.007211
1	experience_level	0.000578
2	employment_type	0.000156

Correlation test

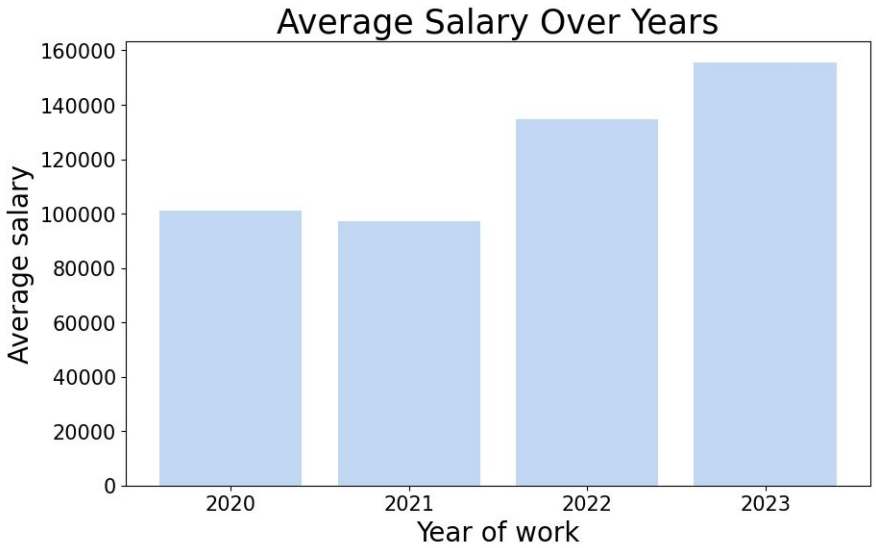
NYU



The median salary is just 200,000, with a steep initial curve suggesting a high concentration of individuals earning lower salaries and a gradual approach to the higher salary range, reflecting income inequality.

Exploratory Data Analysis

How does each feature relate to salary? - Salary Distribution



There is a consistent increase in the average salaries over the years from 2021 to 2023, with a notably steep rise in total salary in 2023 compared to previous years.

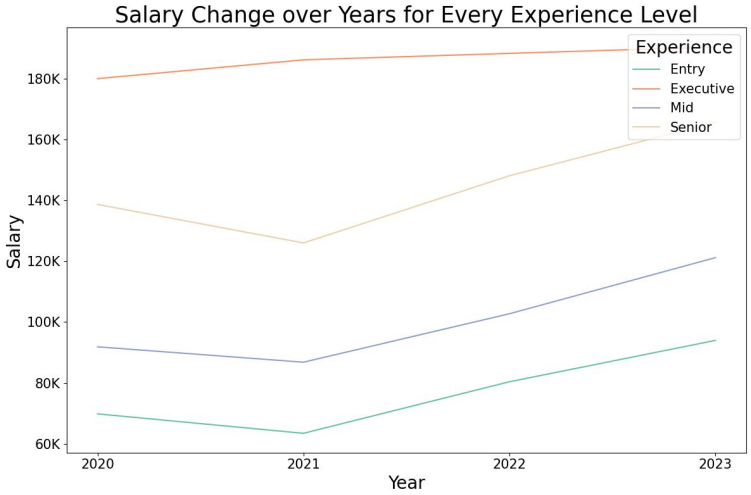
Exploratory Data Analysis

How does each feature relate to salary? - Objective Factors

NYU

Exploratory Data Analysis

How does each feature relate to salary? - Personal Factors

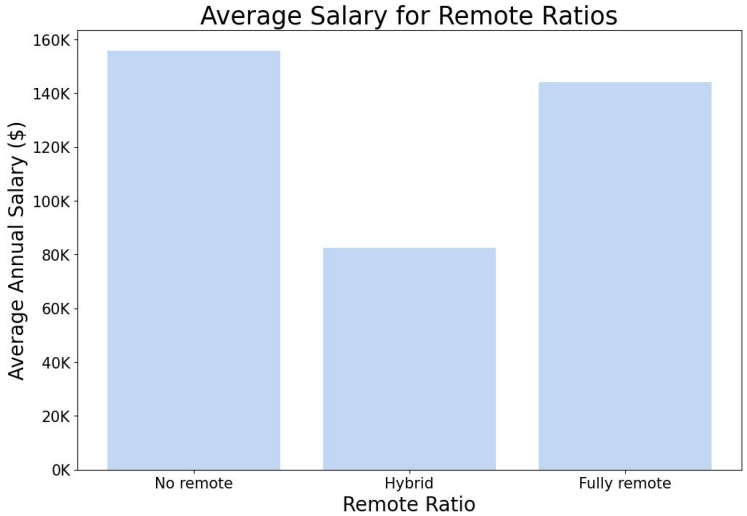


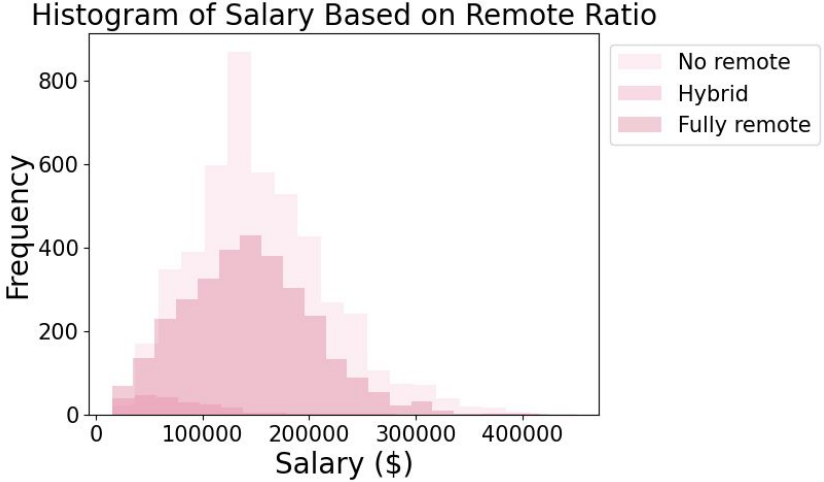
Salaries increase year-over-year across all experience levels from 2020 to 2023, with executive positions showing the highest salaries throughout and entry-level positions the lowest, yet all demonstrating upward trends.

NYU

Exploratory Data Analysis

How does each feature relate to salary? - Job Factors

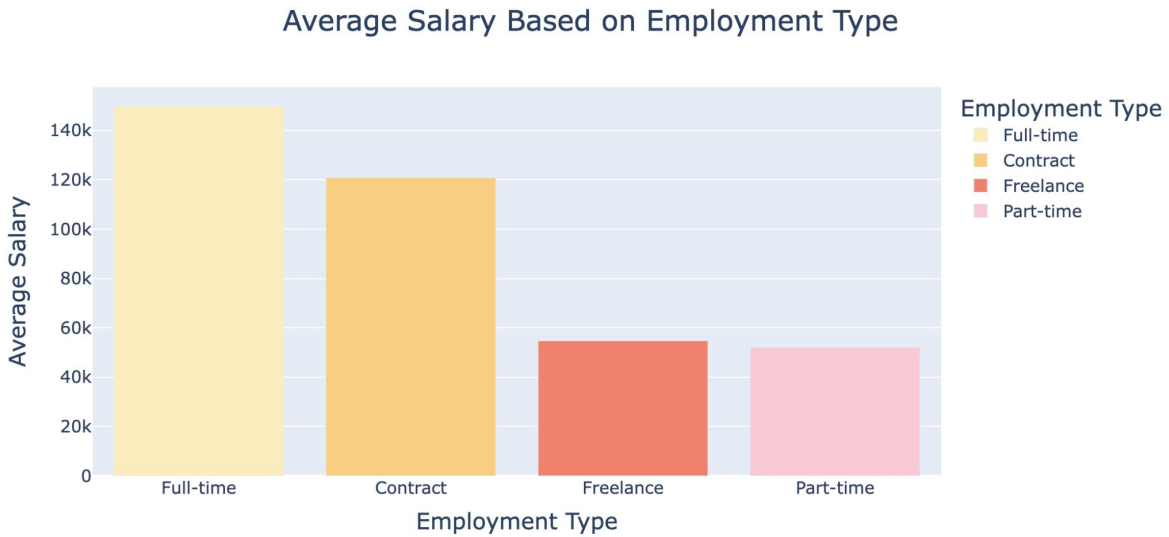




Employees with no remote work option have the highest average salary, followed closely by those who are fully remote, while employees in hybrid work arrangements have a notably lower average salary.

Exploratory Data Analysis

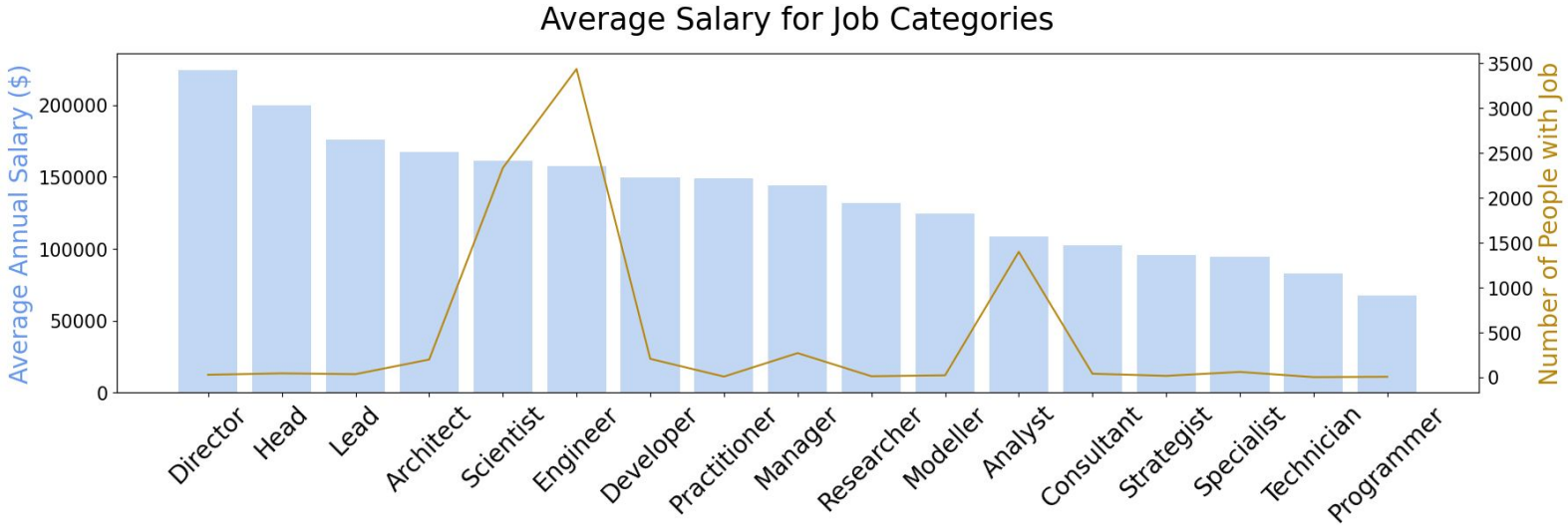
How does each feature relate to salary? - Job Factors



Full-time employees have the highest average salary, followed by contract workers, with freelancers and part-time employees earning less on average.

Exploratory Data Analysis

How does each feature relate to salary? - Job Factors

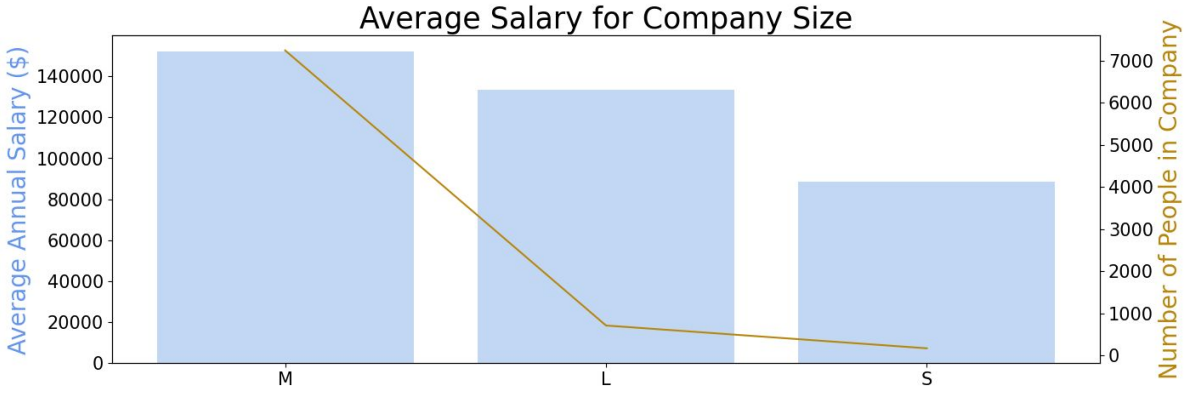


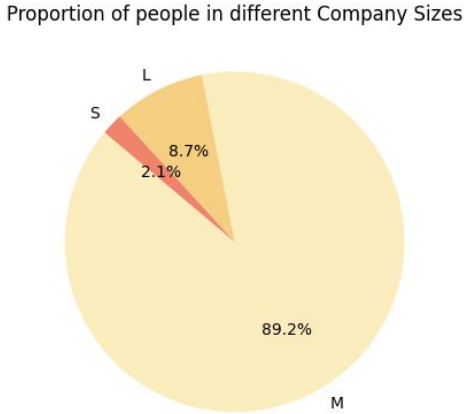
The dual-axis chart shows that roles like 'Director', 'Lead', and 'Architect' not only have higher average salaries but also a lower number of individuals in those roles, suggesting a potential scarcity of skilled professionals in higher-paying positions.

NYU

Exploratory Data Analysis

How does each feature relate to salary? - Company Factors





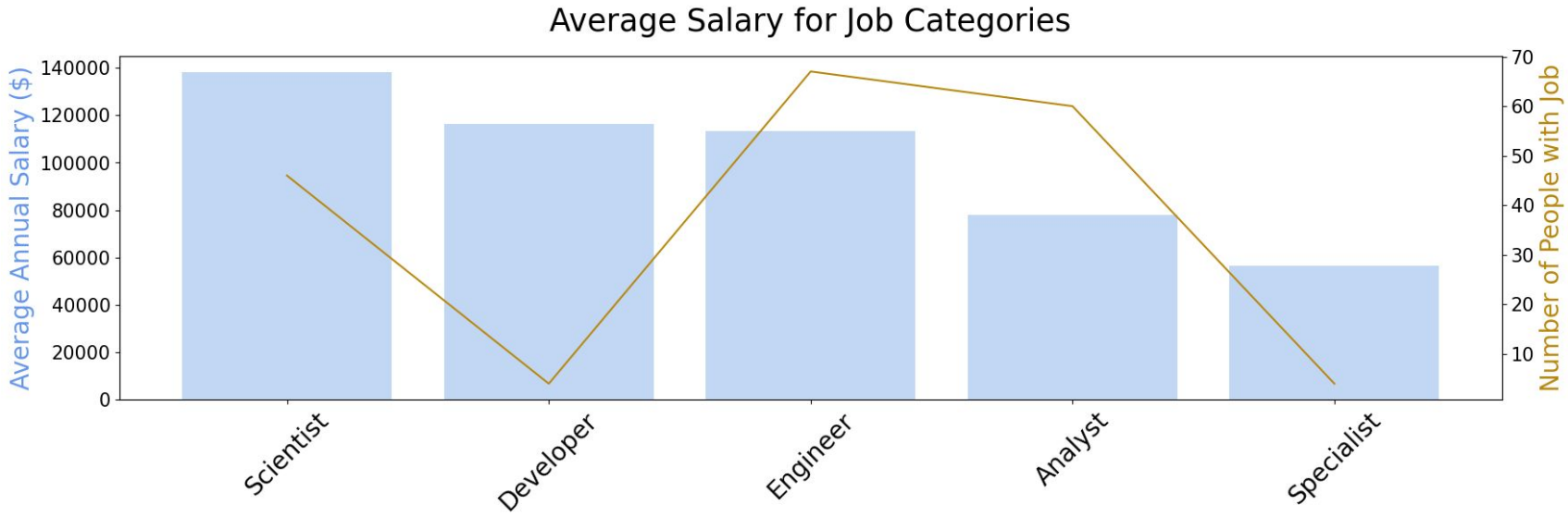
The two charts, suggests that most people work in medium-sized companies, which also offer the highest average annual salary. Large companies employ a smaller percentage of people and offer a slightly lower average salary than medium-sized companies. Small companies, despite having the smallest workforce, offer the lowest average salary.

If we were to find a job this year (2023),

as an entry level and US employee residence,

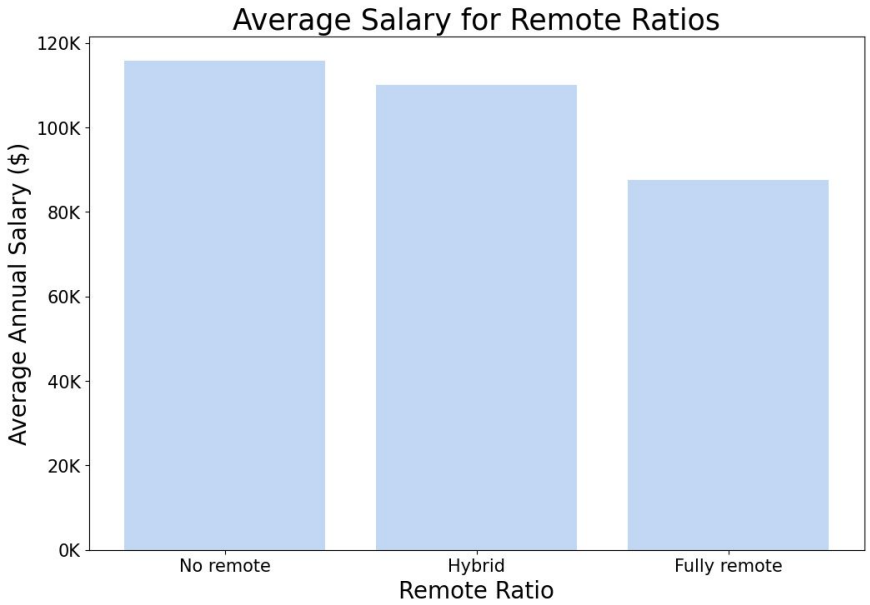
what kind of job should we looking for?

What kind of job should we looking for?



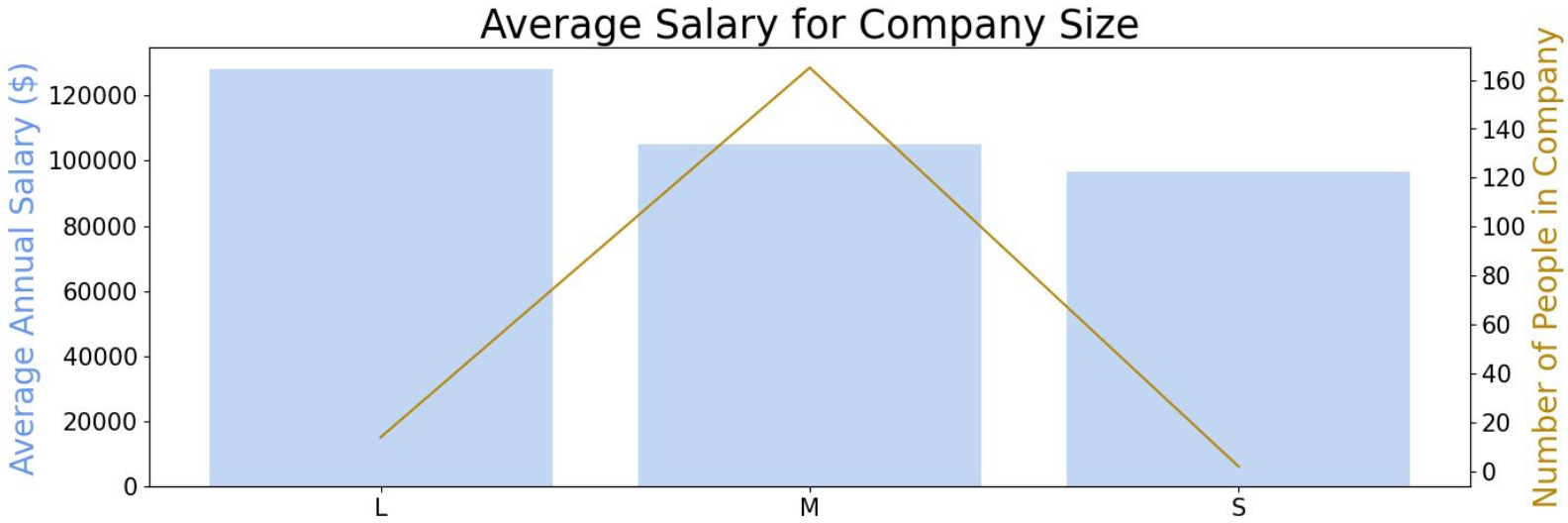
Scientists and Developers command the highest average annual salaries among the job categories listed, making them attractive options for those prioritizing income.

What kind of job should we looking for?

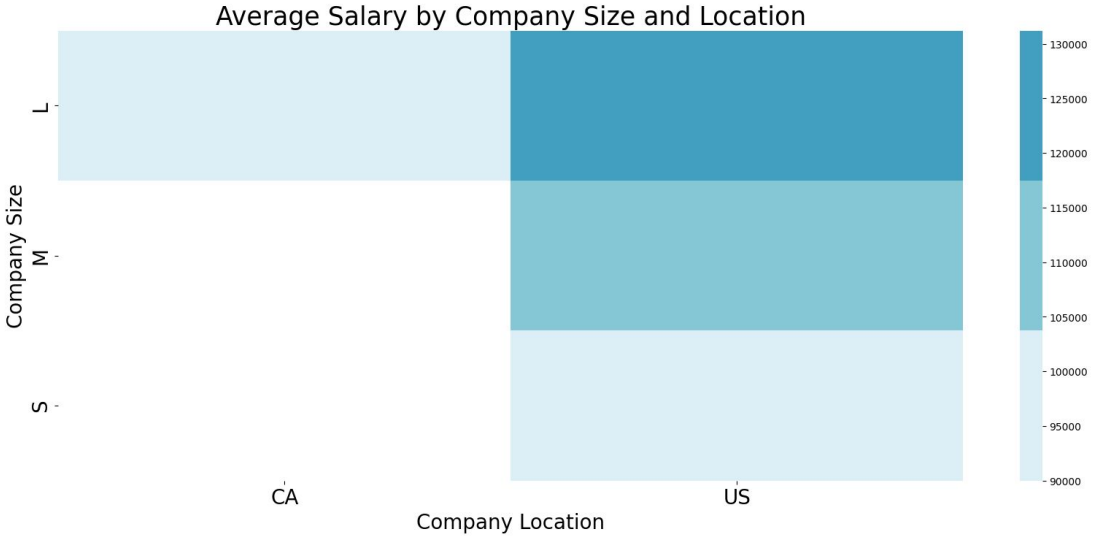


The histogram suggests a wider range of salaries for fully remote positions, indicating variability in pay that could be influenced by the job role, industry, or individual negotiation skills.

What kind of company should we looking for?

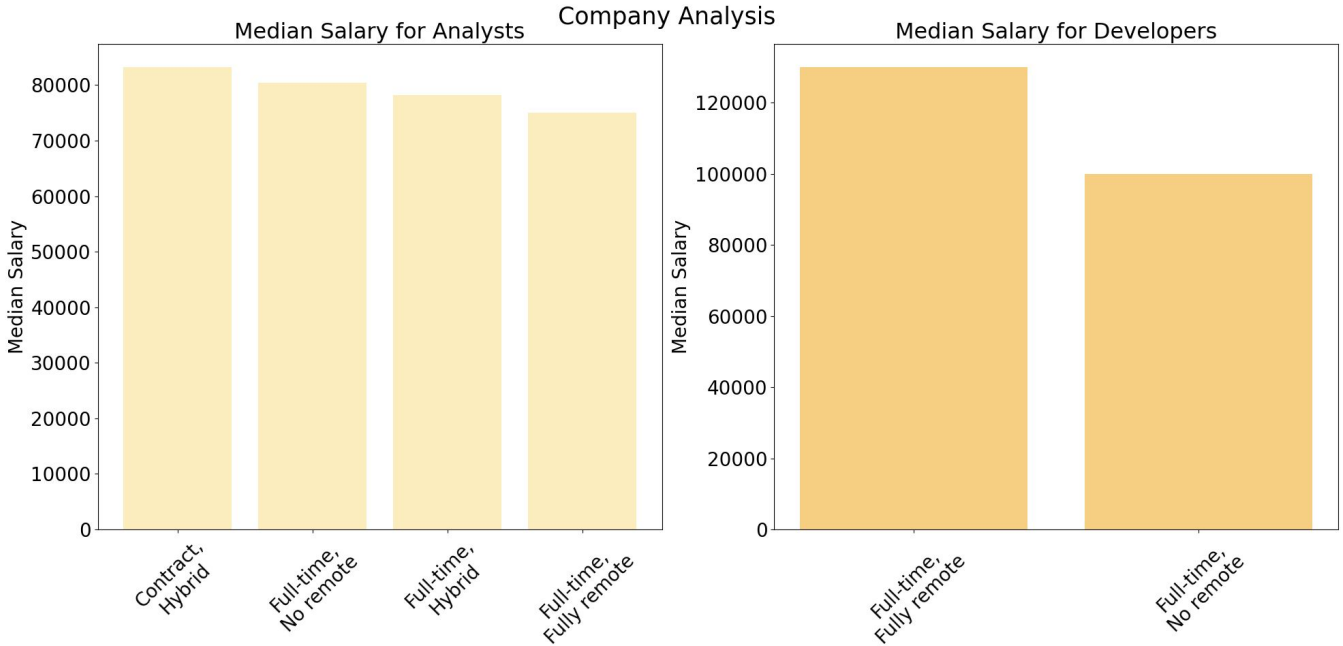


What kind of company should we looking for?



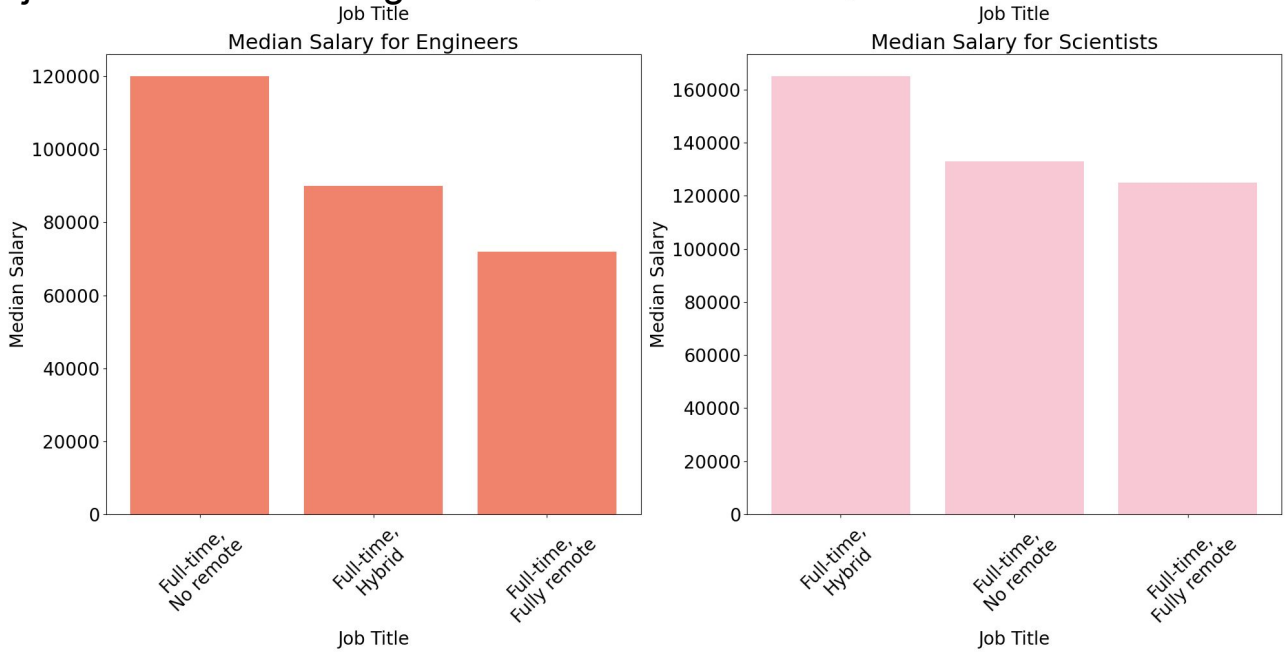
Large-sized companies in the US seem to offer the highest average salaries, and small companies have the lowest average salaries. Meanwhile, large companies in Canada(CA) only offer competitive salaries compared with those of small companies in the US.

What kind of job should we looking for?



PART 03

What kind of job should we looking for?



Across all job titles, full-time positions tend to have higher median salaries compared to remote or part-time options, with Engineers showing the most significant discrepancy based on employment type.

NYU

If we were to find a job this year (2023),

as an entry level and US employee residence,

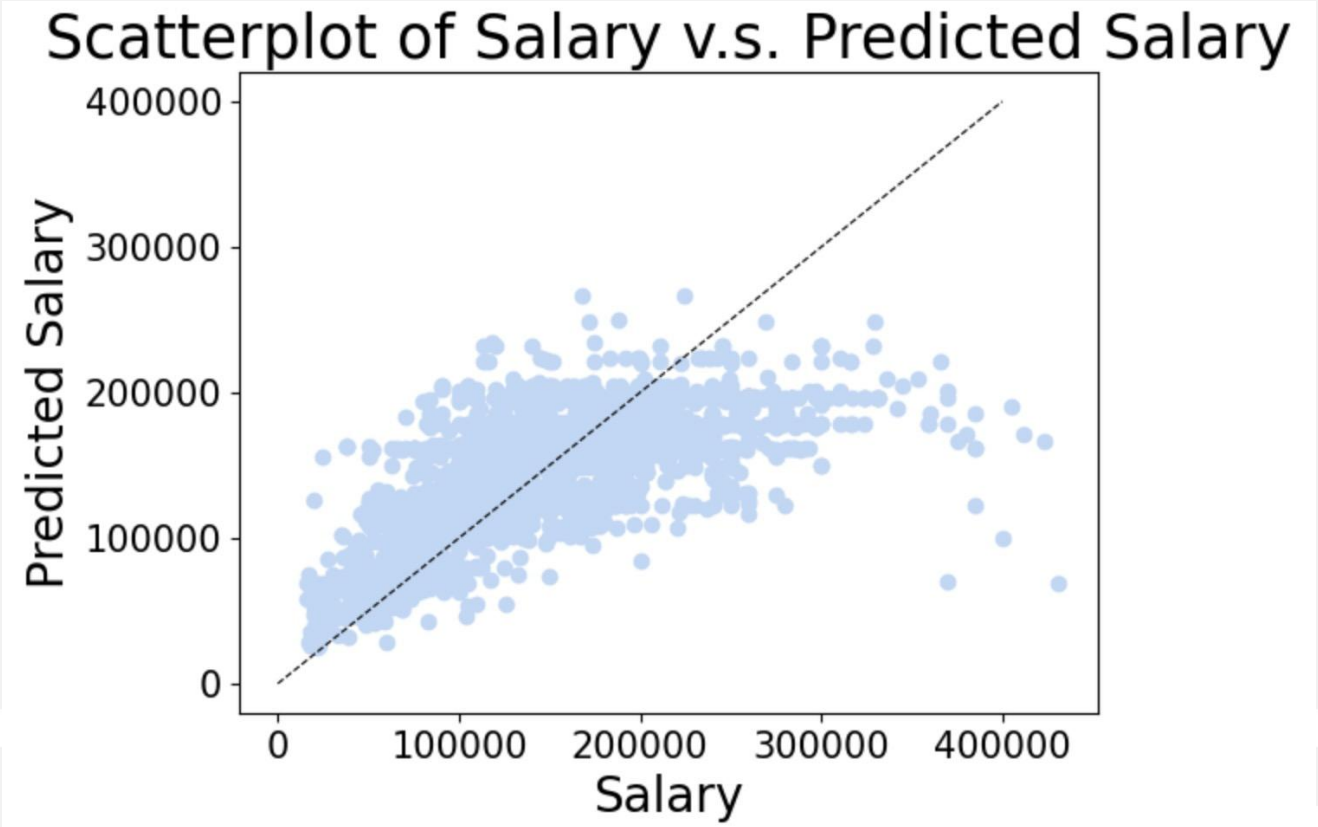
what kind of job should we looking for?

Machine Learning – Predict Salary

1. Linear Regression
2. Regularization Linear Regression
   - Lasso
   - Ridge
3. Regression Tree
4. Light GBM

	Mean Squared Error		Mean Absolute Error		R Squared	
Model	Train	Test	Train	Test	Train	Test
Linear	2.27×10^9	1.54×10^{25}	36380.47	2.34×10^{11}	0.44	-3.69×10^{15}
Lasso	2.32×10^9	2.61×10^9	37037.17	3.84×10^4	0.43	0.37
Ridge	2.28×10^9	2.62×10^9	36653.22	3.85×10^4	0.44	0.37
DecisionTree	1.95×10^9	2.80×10^9	32145.72	3.91×10^4	0.52	0.33
Light GBM	2.27×10^9	2.56×10^9	36275.78	3.79×10^4	0.44	0.39

LightGBM


Scatterplot of Salary v.s. Predicted Salary
Predicted Salary
400000
300000
200000
100000
0
0
100000
200000
300000
400000
Salary

Future Work for Model Improvement

If we had more time, the model may be future improved…

1. Feature Engineering
    - For now, we apply all the features we have while training the model. If we have time, we may do feature engineering according to the correlation before training.
2. Try more Models
    - XGB Regression
    - Deep Learning