Data Science Job Salary Predictions

Chris Colomb (V00970873)
Maika Rabenitas (V00970890)

Motivation

- Emergence of AI boom
- Relevance and appeal of data science have grown
- Demand for skilled data science professionals is evident across various industries and regions
- Disparities in compensation based on company size, job title, etc.

General Questions

- How do different factors influence job salaries?
- Are there any notable trends or patterns in salary structures?

Specific Question

If we were to become an entry-level data engineer in 2024, how much would our salary change working in Canada compared to US? Does remote work play a factor?

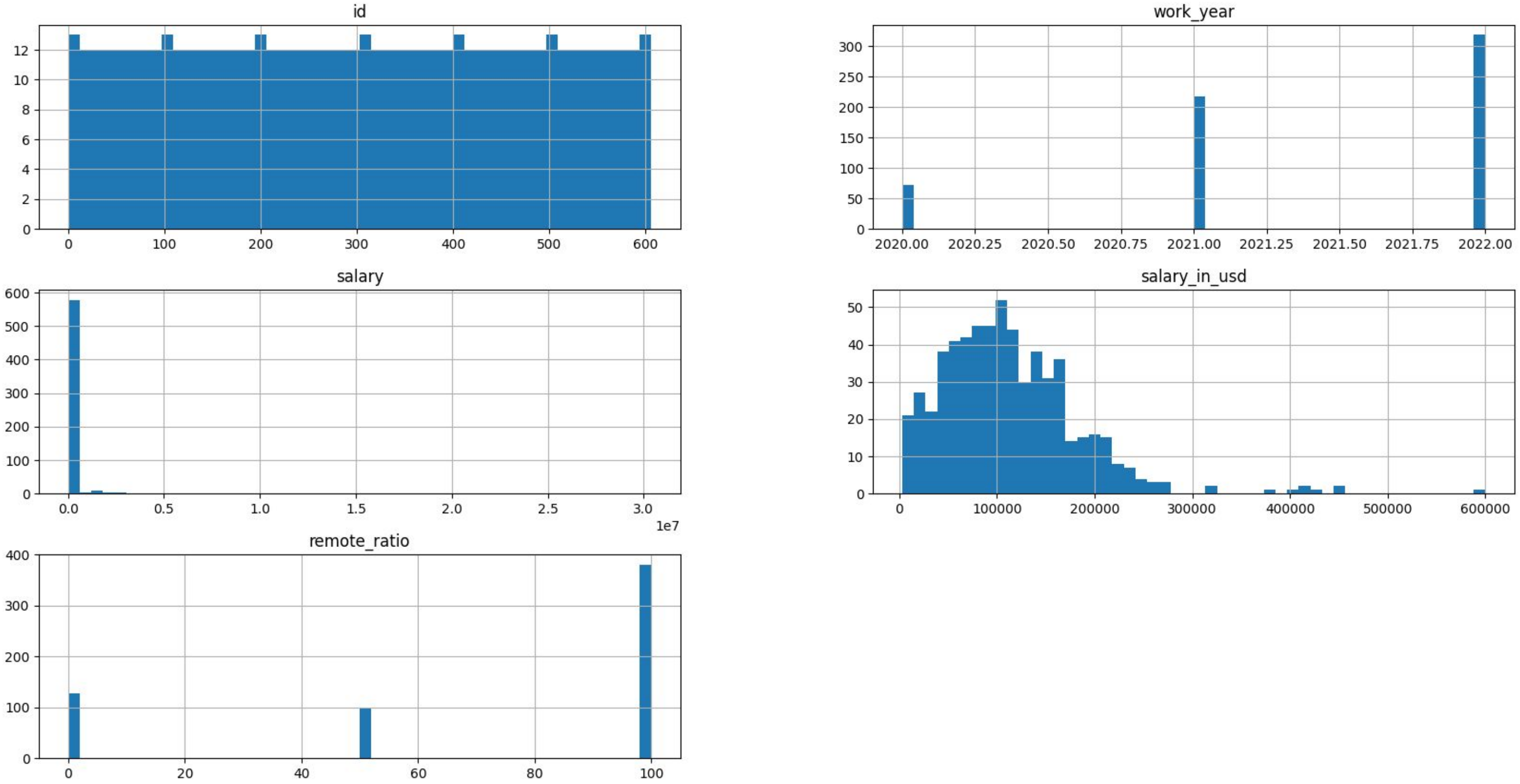
Dataset Breakdown

- id
- work_year
- experience_level
- employment_type
- job_title
- salary
- salary_currency
- salary_in_usd
- employment_residence
- remote_ratio
- company_location
- company_size

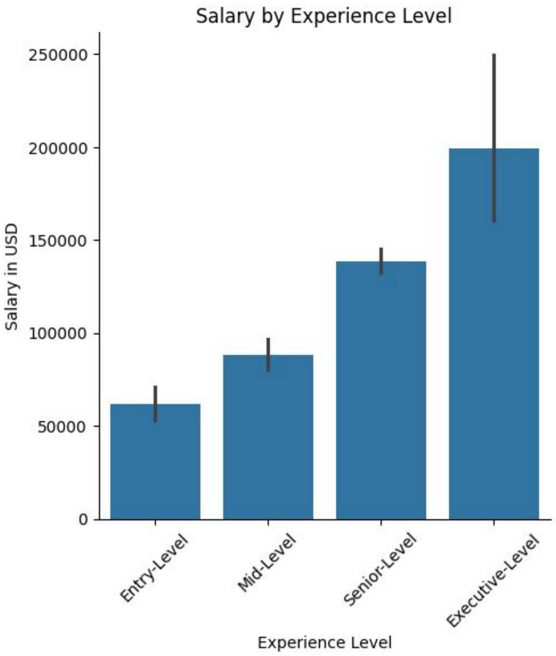
ds_salary.info()

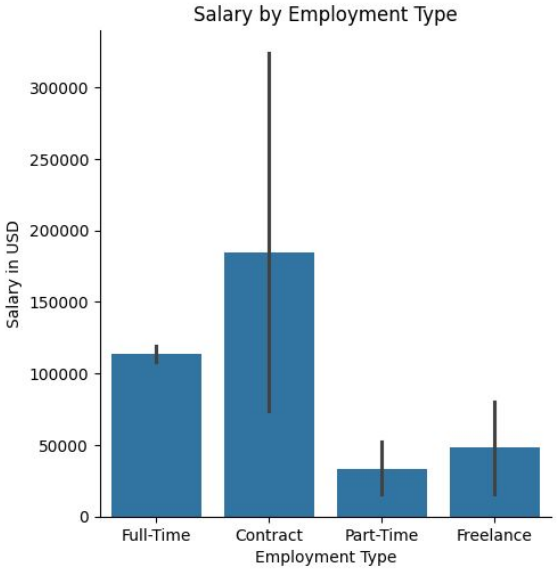
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   id                  607 non-null    int64
 1   work_year           607 non-null    int64
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
memory usage: 57.0+ KB
```

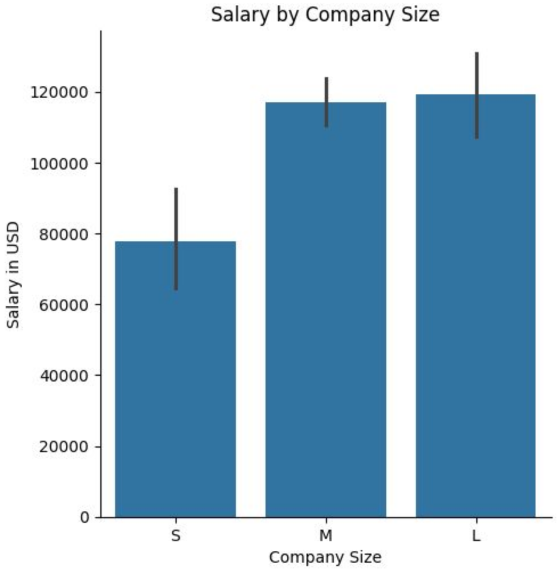
ds_salary.hist()



Plotting Against Salary



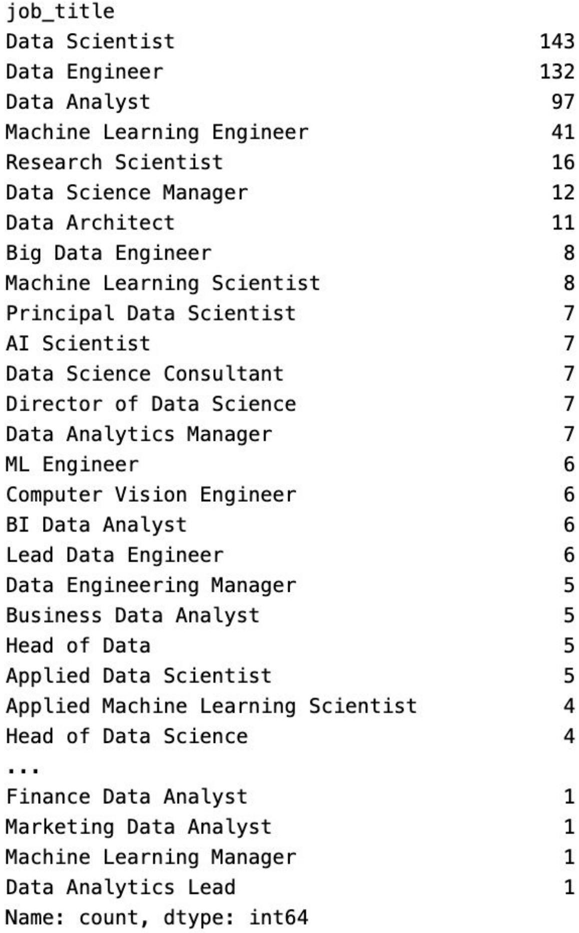




Problems

- Seven categorical attributes
- Small dataset with more than half of our data based in the US
- Outliers in most of our categorical data

Example



```
job_title
Data Scientist                              143
Data Engineer                               132
Data Analyst                                 97
Machine Learning Engineer                    41
Research Scientist                           16
Data Science Manager                         12
Data Architect                               11
Big Data Engineer                             8
Machine Learning Scientist                    8
Principal Data Scientist                      7
AI Scientist                                  7
Data Science Consultant                       7
Director of Data Science                      7
Data Analytics Manager                        7
ML Engineer                                   6
Computer Vision Engineer                      6
BI Data Analyst                               6
Lead Data Engineer                            6
Data Engineering Manager                      5
Business Data Analyst                         5
Head of Data                                  5
Applied Data Scientist                        5
Applied Machine Learning Scientist            4
Head of Data Science                          4
...
Finance Data Analyst                          1
Marketing Data Analyst                        1
Machine Learning Manager                      1
Data Analytics Lead                           1
Name: count, dtype: int64
```

Feature Engineering

- Removed unnecessary columns (`id` and `salary`)

- Converted categorical data -> Numeric data
  - Entry-level -> 1, Mid-level -> 2 and so on..

- Bucketed the outliers
  - For the job title, if the count is less than 10 grouped into `Other`

- Created a new column `same_country`
  - If `employee_residence == company_location` then 1 otherwise `0`

- Created three new columns `salary_mean`, `salary_min` and `salary_max` for each of the salaries grouped by clusters
  - Clusters: `work_year`, `experience_level`, `employment_type`, `job_title`, `remote_ratio`, `company_size` and `same_location`

Data Preprocessing

- Numerical: `MinMaxScaler()`
- Categorical: `OneHotEncoder()`

Modeling for Regression

- `Ridge()`
- `SGDRegressor()`
- `RandomForestRegressor()`
- `DecisionTreeRegressor()`

Results

Mean Squared Error (MSE)

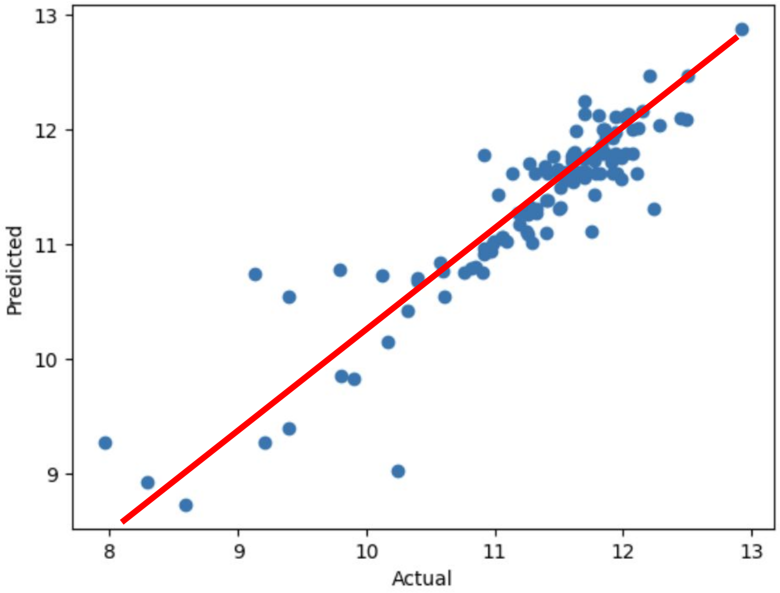
Ridge	0.19725377418055814
SGD	0.4231541951213471
Random Forest	0.12491690719113324
Decision Tree	0.14350962674677212

R2 Score

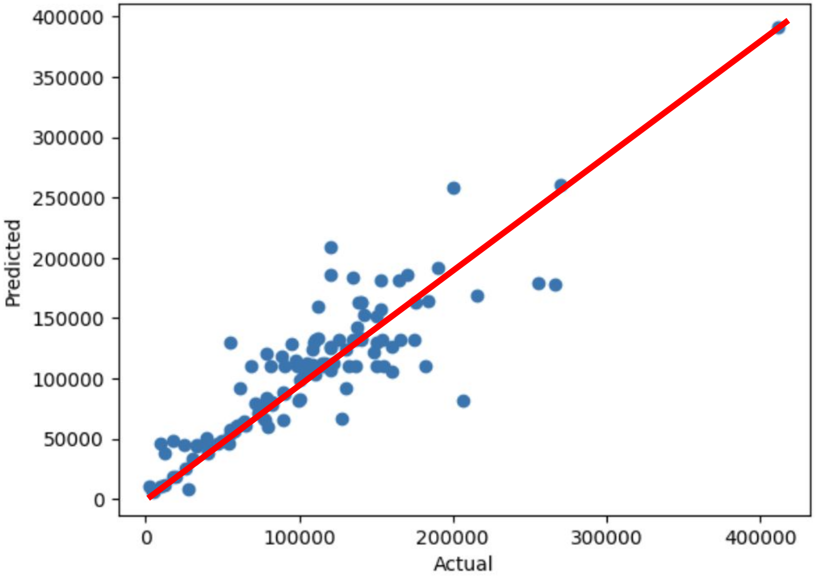
Ridge	0.7195537582301159
SGD	0.39837904646471056
Random Forest	0.8223989512960932
Decision Tree	0.7959646873874725

Actual vs Predicted Salaries

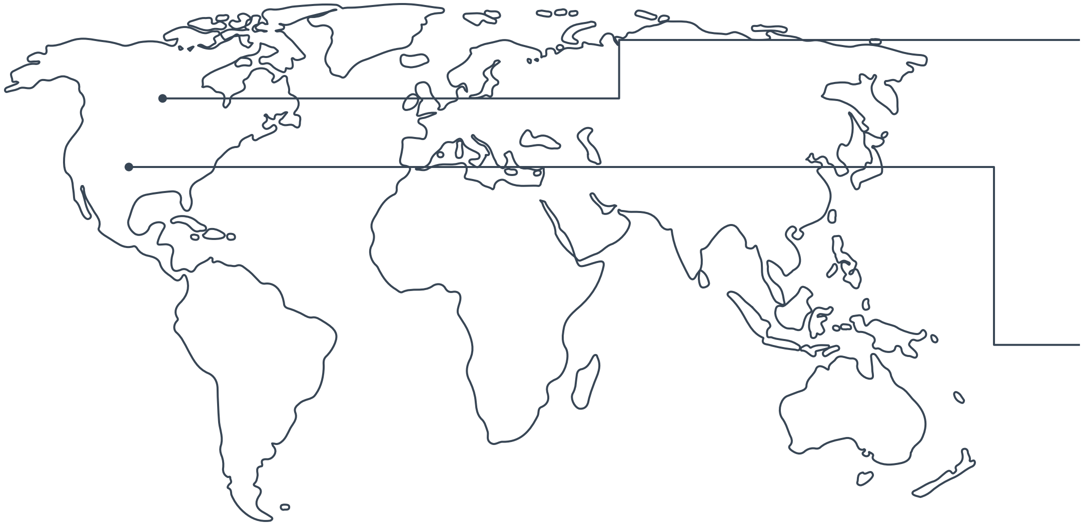
Log Transformed



Exponential Transformed



Salary Prediction Results (USD)



Canada

Remote: $93,333
Non-Remote: $67,000

United States

Remote: $161,366
Non-Remote: $98,666

Thank you