GROUP 26

DATA SCIENCE SALARY PREDICTIONS

Jasmine Chen, Vedant Goyal, Usman Moazzam,
Kanisha Shah, Lin Zhu

VISION

We aim to understand salary trends and factors influencing compensation for data science-related roles.

Goal: develop a machine learning model capable of predicting salaries for professionals in these data-centric roles

READ MORE

DATA CLEANING

- Selected 21 out of 39 questions
- Included Qs about demographics, company info, high-level professional background, and ML related techniques/skills
- Used ordinal and one-hot encoding, also transformed some encoded columns into binary columns
- Removed columns with no salary

READ MORE

Size

20,036 rows, 355 columns

Data

Survey responses with 39 questions

Concerns

High dimensionality, limited # responses, sparse columns

COLUMN SELECTION

- Questions 1 - 7, 20 and 21 cover basic demographics, company information as well as high-level professional background of the respondent
- Questions 15, 17 and 22 cover machine-learning related techniques/skills.
- Questions 30, 32 and 33 all focus on technologies that the respondents use on a daily basis
- Removed columns that were too specific, redundant/codependent, etc

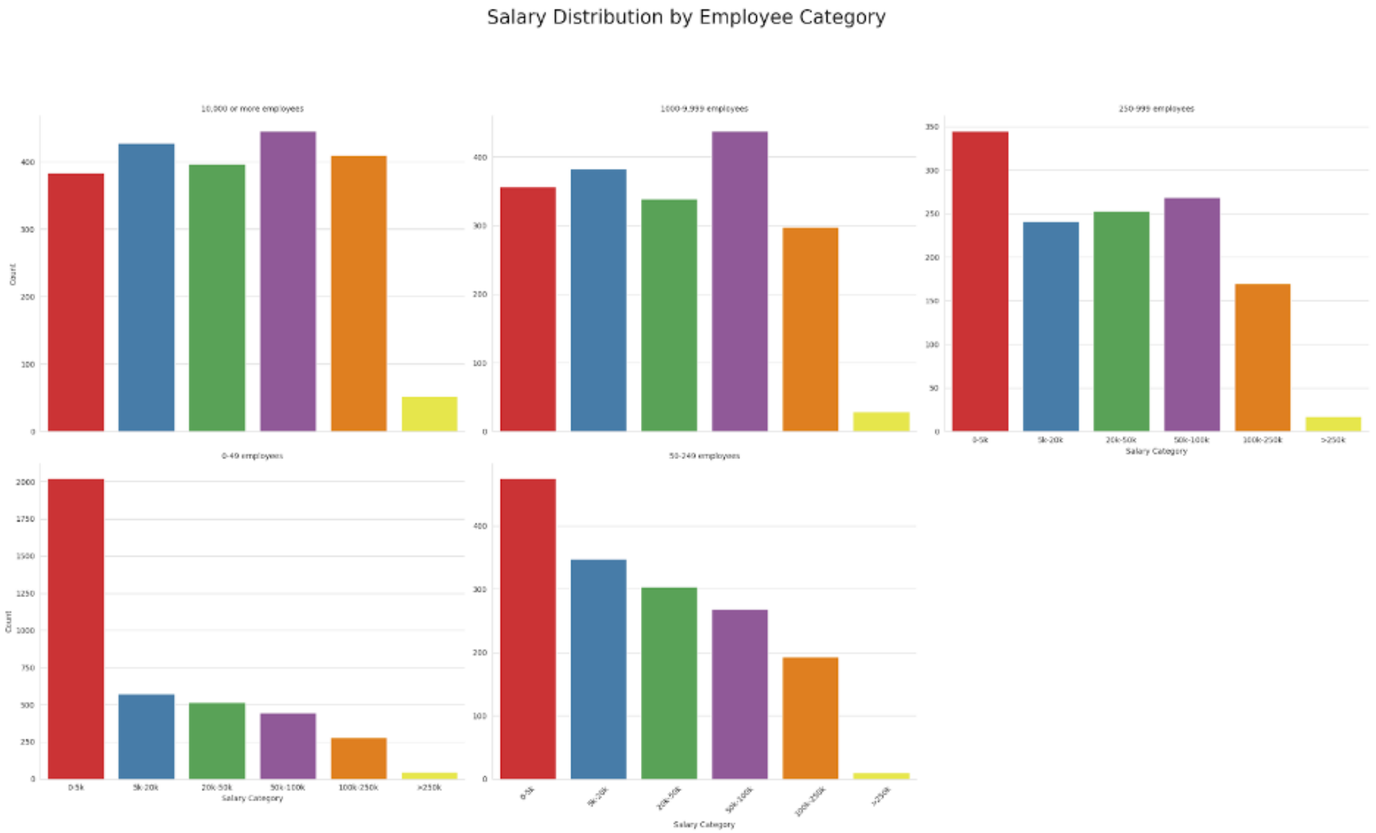
READ MORE

DATA EXPLORATION:

Exploring Company Size vs Salary

This bar chart shows the distribution of salaries faceted by the number of employees in the company.

This chart suggests that companies with fewer number of employees follow a trend in the salary, i.e. the no. of employees in a particular salary range is inversely proportional to the actual salary range. Another clear suggestion is that companies with higher numbers of employees pay a greater salary compared to the companies with lesser employees.



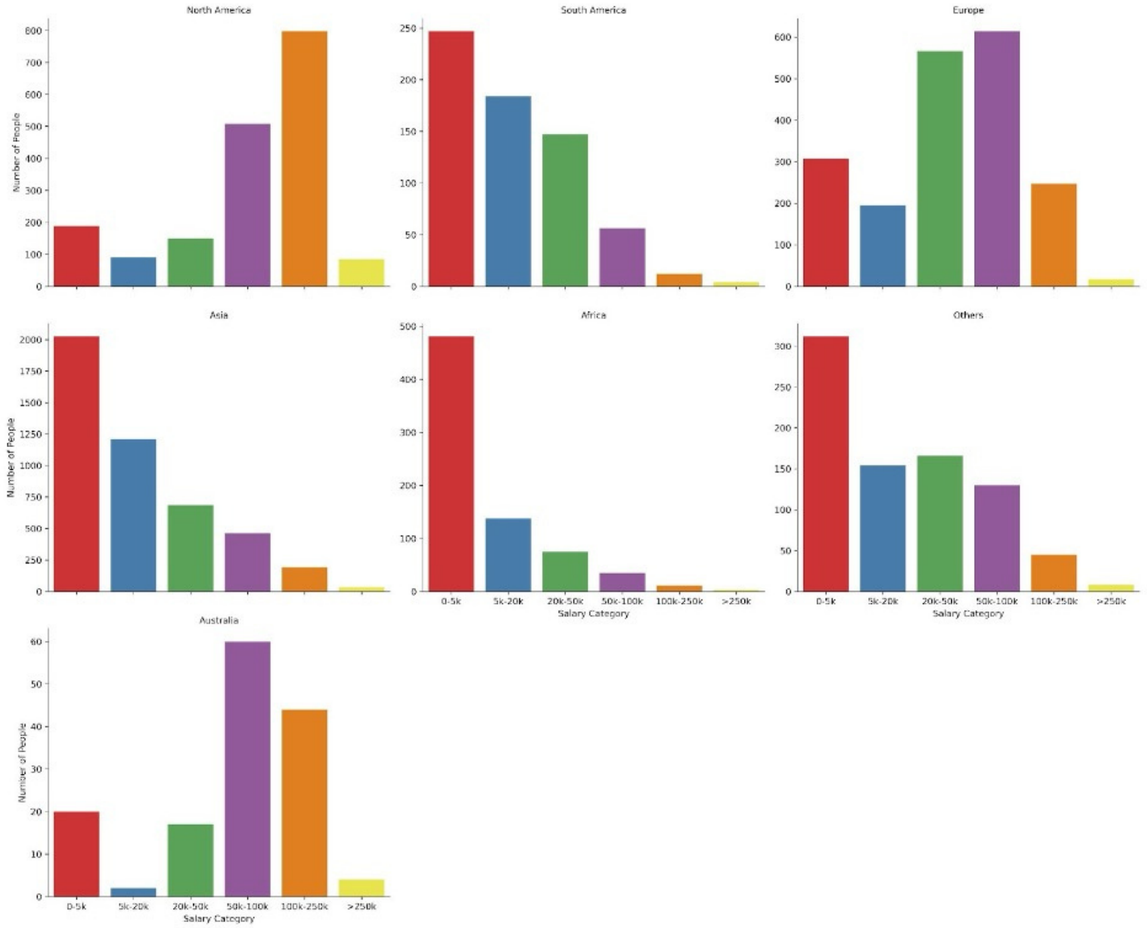
Created in CoLab

DATA EXPLORATION:

Exploring Continent vs Salary

This bar chart shows the distribution of salaries faceted by the continents.

There is a clear suggestion that the pay scale is significantly higher in North America, Europe and Australia. For Asia and Africa, one may observe that the proportion of employees in a salary bracket is inversely proportional to the actual salary.



Created in CoLab

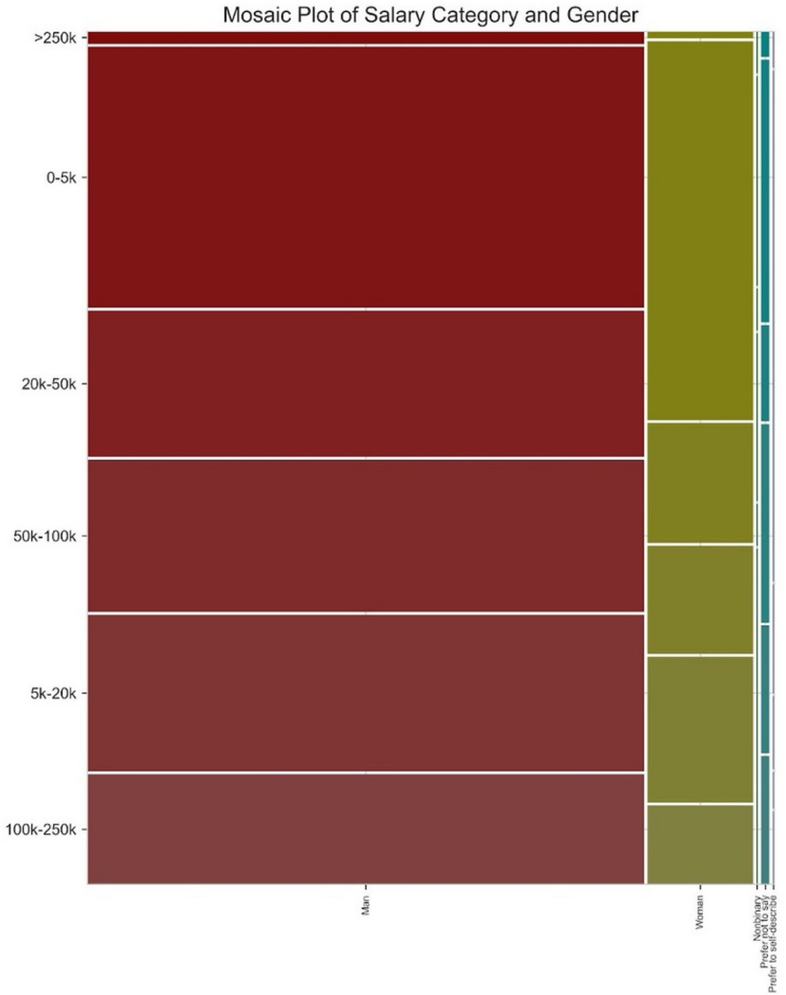
DATA EXPLORATION:

Exploring Gender vs Salary

This mosaic plot shows the distribution of salary categories and gender in a company. Each square in the plot represents a different combination of salary category and gender, and the size of the square is proportional to the number of employees in that category.

The plot shows that the highest salary category (>250k) is dominated by males, with other genders representing roughly 20% of employees in this category. The gender gap is also significant in the 100k-250k salary category, with other genders representing only 40% of employees. However, the gender gap narrows in the lower salary categories, with other genders representing close to 50% of employees in the 5k-20k and 20k-50k salary categories.

This plot provides a clear visualization of the gender pay gap prevalent in industries. It also highlights the fact that the gender gap is particularly pronounced in the highest salary categories.



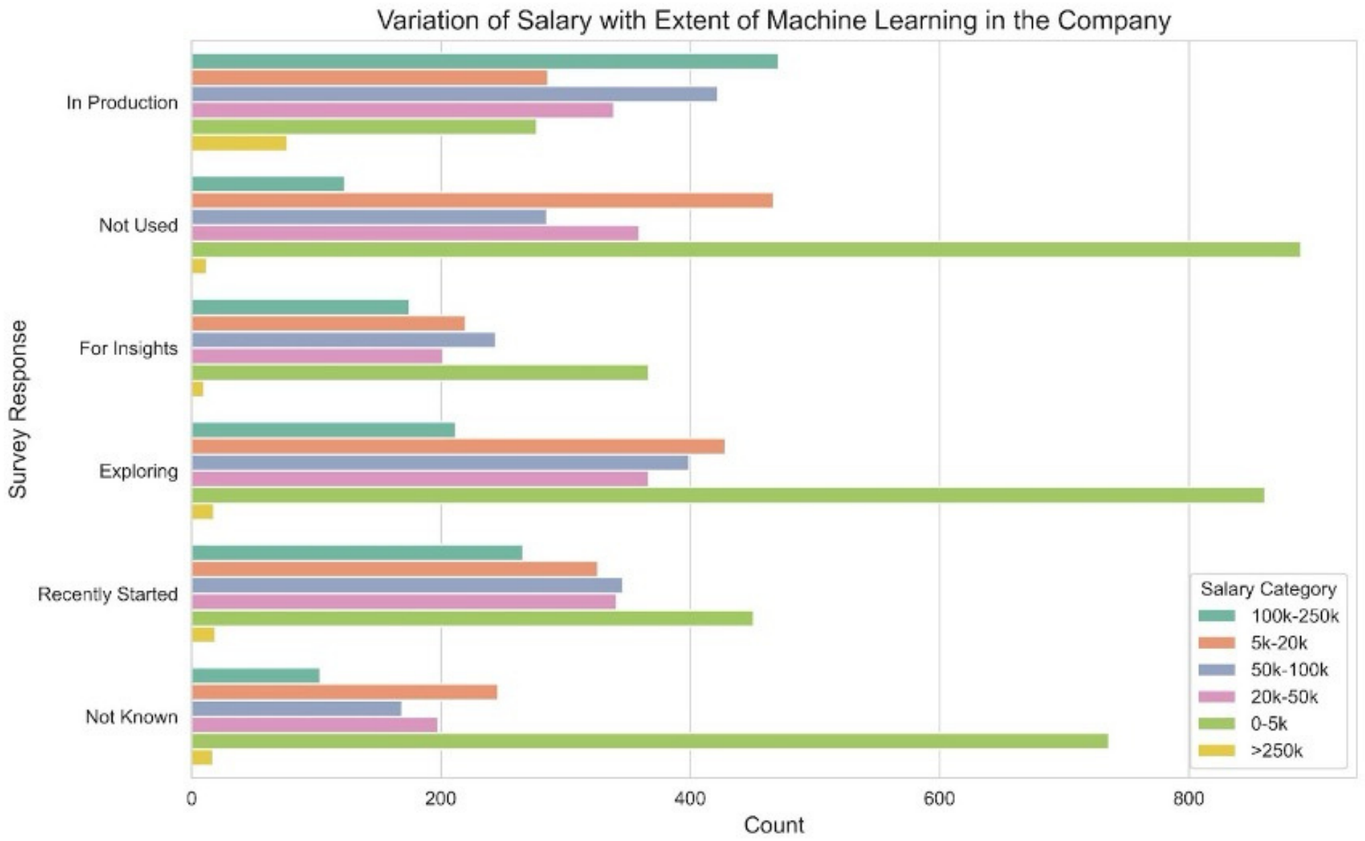
Created in CoLab

DATA EXPLORATION:

Exploring Company's ML Implementation vs Salary

This bar chart shows the distribution of salary broken down by the extent of machine learning. The data is based on a survey of employees, and shows the percentage of employees in each salary category who reported using machine learning in their work.

The chart suggests that in companies where Machine Learning is being used in some way or the other, the total number of employees being paid more than 100k is much higher than companies where there is no usage of ML.



Created in CoLab

ML MODELS

We'll start with Logistic Regression with L1 regularization for an efficient baseline that also helps in feature reduction and mitigating overfitting.

We'll then progress to tree-based ensembles like Random Forests and Gradient Boosting Machines, which are effective for high-dimensional data and assist in feature importance evaluation.

Finally, we will use deep neural networks for their proficiency in complex pattern recognition. Thorough cross-validation will be essential at each step to prevent overfitting and ensure our model's generalizability to new data.

READ MORE

Logistic Regression (L1)

Serves as a fundamental baseline, can reduce overfitting and dimensionality

Random Forest, Gradient Boosting

Work well with high dimensionality

Deep Learning

Excel at learning complex patterns in large-scale data