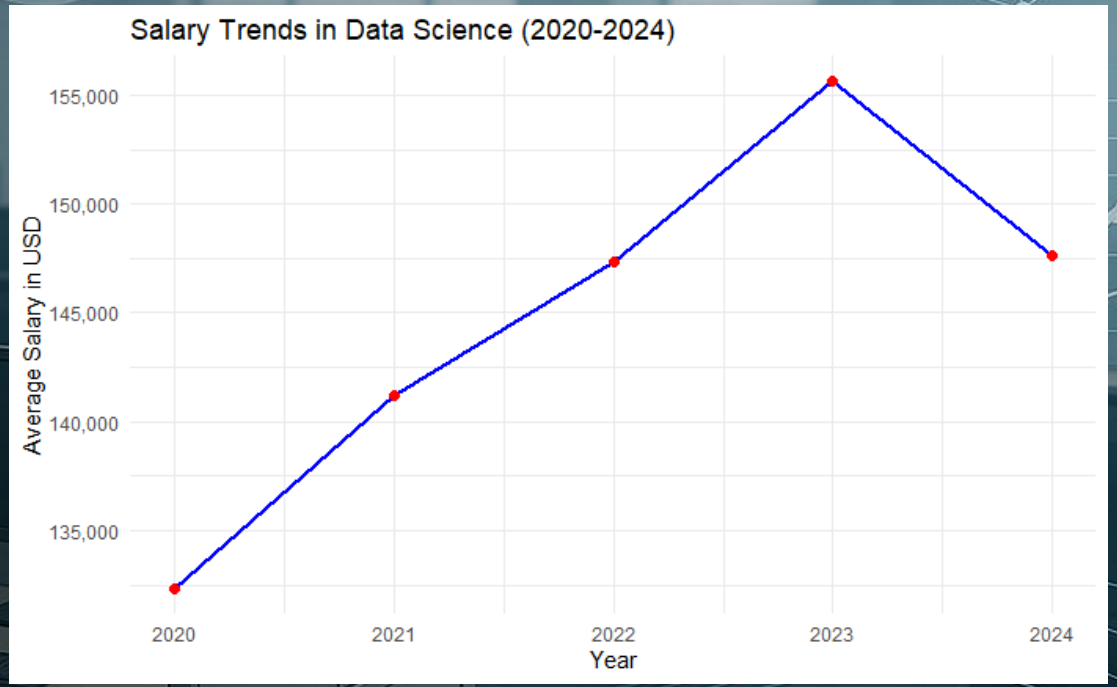
Data Science Salary Predictions

Previous Research Question:

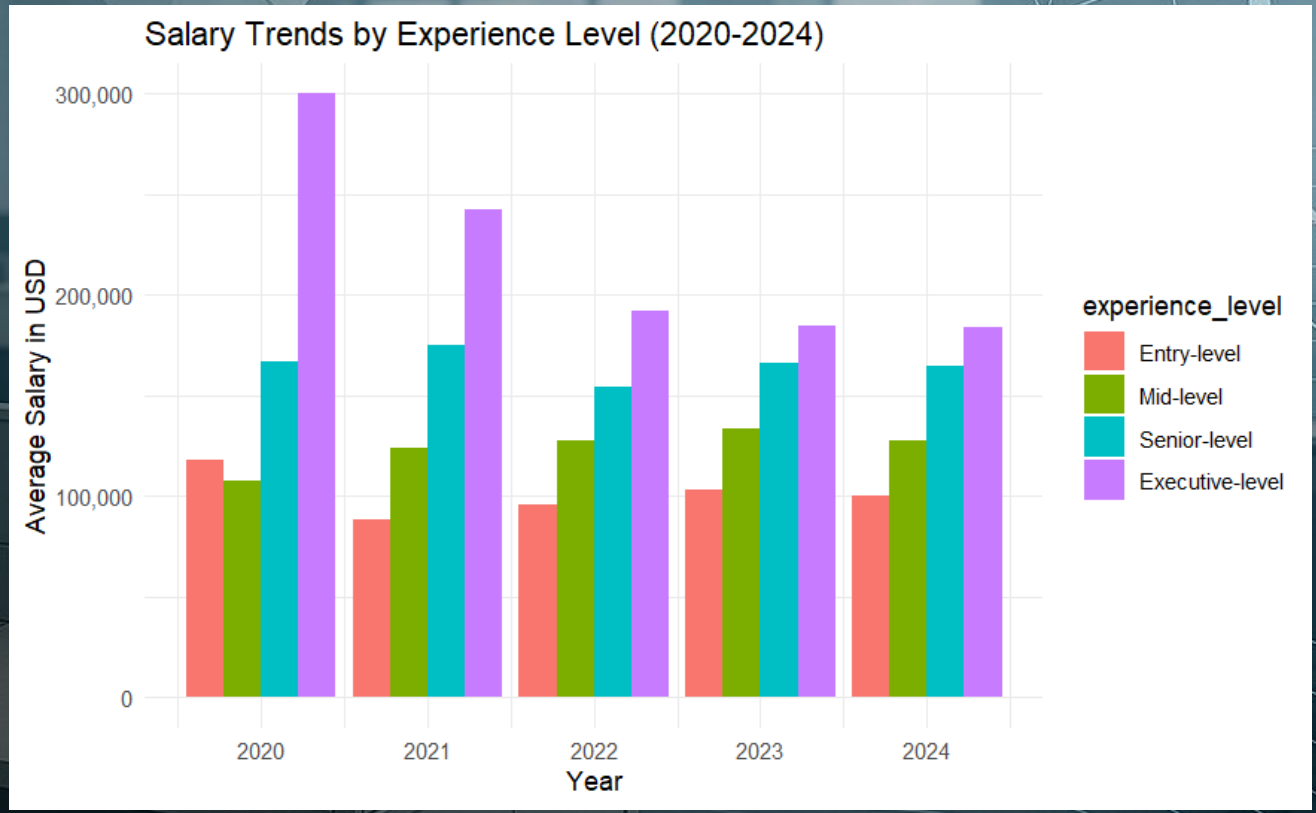
We analyzed salary trends from 2020 to 2024, exploring how salaries have changed over time and whether there are any identifiable trends in salary growth?



- Recap
- Current Research Question
- Variable Manipulation
- Regression Model 1

Previous Research Question:

We analyzed salary trends from 2020 to 2024, exploring how salaries have changed over time and whether there are any identifiable trends in salary growth?

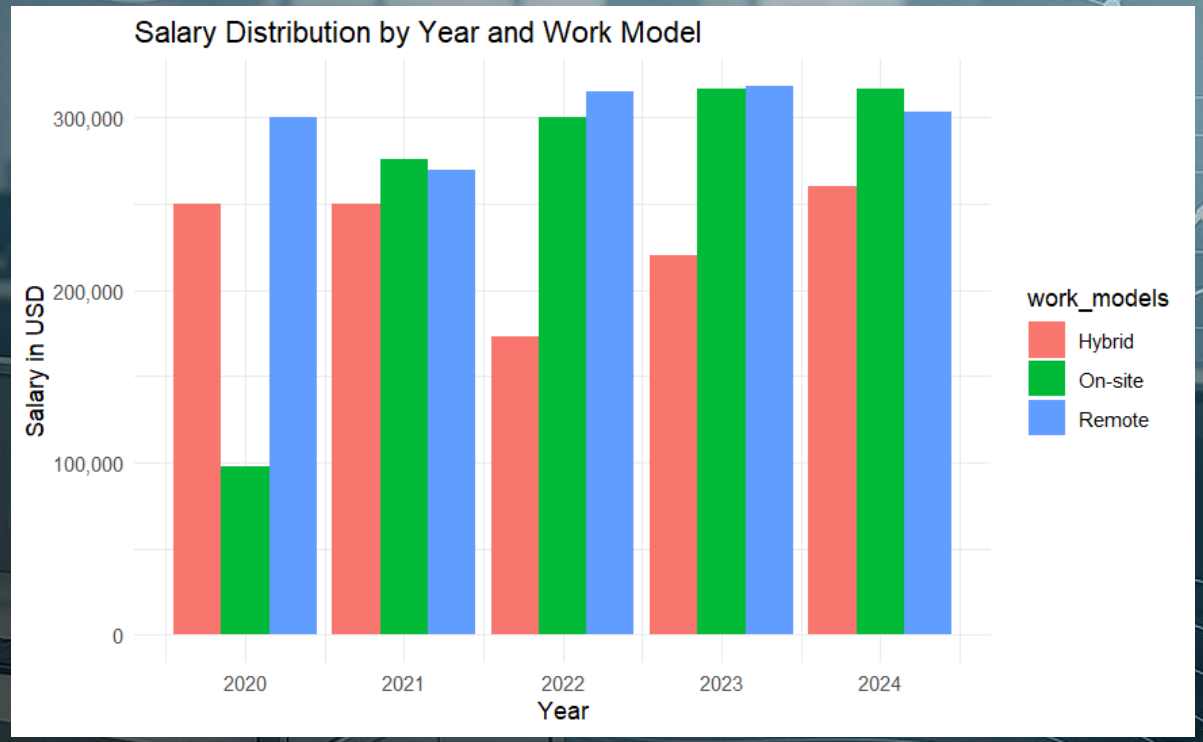


- Recap
- Current Research Question
- Variable Manipulation
- Regression Model 1
- Regression Model 2

Previous Research Question:

We analyzed salary trends from 2020 to 2024, exploring how salaries have changed over time and whether there are any identifiable trends in salary growth?

- Recap
- Current Research Question
- Variable Manipulation
- Regression Model 1
- Regression Model 2



Current Research Question:

- Recap
- Current Research Question
- Variable Manipulation
- Regression Model 1
- Regression Model 2
- Regression Model 3

How accurately can we predict the 2025 average salary of data science professionals using 2020–2024 data, based on:

Job Title	Experience Level	Employment Type	Work Model	Company Size

- Recap
- Current Research Question
- Variable Manipulation
- Regression Model 1
- Regression Model 2
- Regression Model 3
- Analyzing Regression Results

From Char type to Factor:

Job Title
Experience Level
Employment type
Work Model
Company Size

Linear Regression

- Recap
- Current Research Question
- Variable Manipulation
- **Regression Model 1**
- Regression Model 2
- Regression Model 3
- Analyzing Regression Results
- Question Solved?

Metric <chr>	Value <dbl>
RMSE	52697.3175
MAE	42920.4273
R2	0.2576
MAPE	35.6495

4 rows

Random Forest Regressor

- Current Research Question
- Variable Manipulation
- Regression Model 1
- **Regression Model 2**
- Regression Model 3
- Analyzing Regression Results
- Question Solved?
- Regression to Classification

Metric <chr>	Value <dbl>
RMSE	52982.3613
MAE	43377.5408
R2	0.2098
MAPE	35.7811

4 rows

- Current Research Question
- Variable Manipulation
- Regression Model 1
- Regression Model 2
- **Regression Model 3**
- Analyzing Regression Results
- Question Solved?
- Regression to Classification
- Classification Results

eXtreme Gradient Boosting (XGB)

Metric <chr>	Value <dbl>
RMSE	0.1910
MAE	0.1564
R2	0.1783
MAPE	34.1654

4 rows

- Variable Manipulation
- Regression Model 1
- Regression Model 2
- Regression Model 3
- **Analyzing Regression Results**
- Question Solved?
- Regression to Classification
- Classification Results
- Confusion Matrix

Definition

“Random Forest is a supervised machine learning algorithm that builds multiple decision trees using random subsets of data and features. It relies on ***variability*** in the input features to learn meaningful patterns”

job_title	experience_level	employment_type	work_models	work_year	salary_in_usd	company_size
Data Scientist	Senior-level	Full-time	Remote	2024	195500	Medium
Data Scientist	Senior-level	Full-time	Remote	2024	141300	Medium

Research Question:

How accurately can we predict the 2025 average salary of data science professionals using 2020–2024 data, based on “Job Title”, “Experience Level”, “Employment Type”, “Work Models”, and “Company Size”?

We cannot

No variability in Data.

Missing Important Information

Can this be solved by classification?

- Regression Model 2
- Regression Model 3
- Analyzing Regression Results
- Question Solved?
- **Regression to Classification**
- Classification Results
- Confusion Matrix

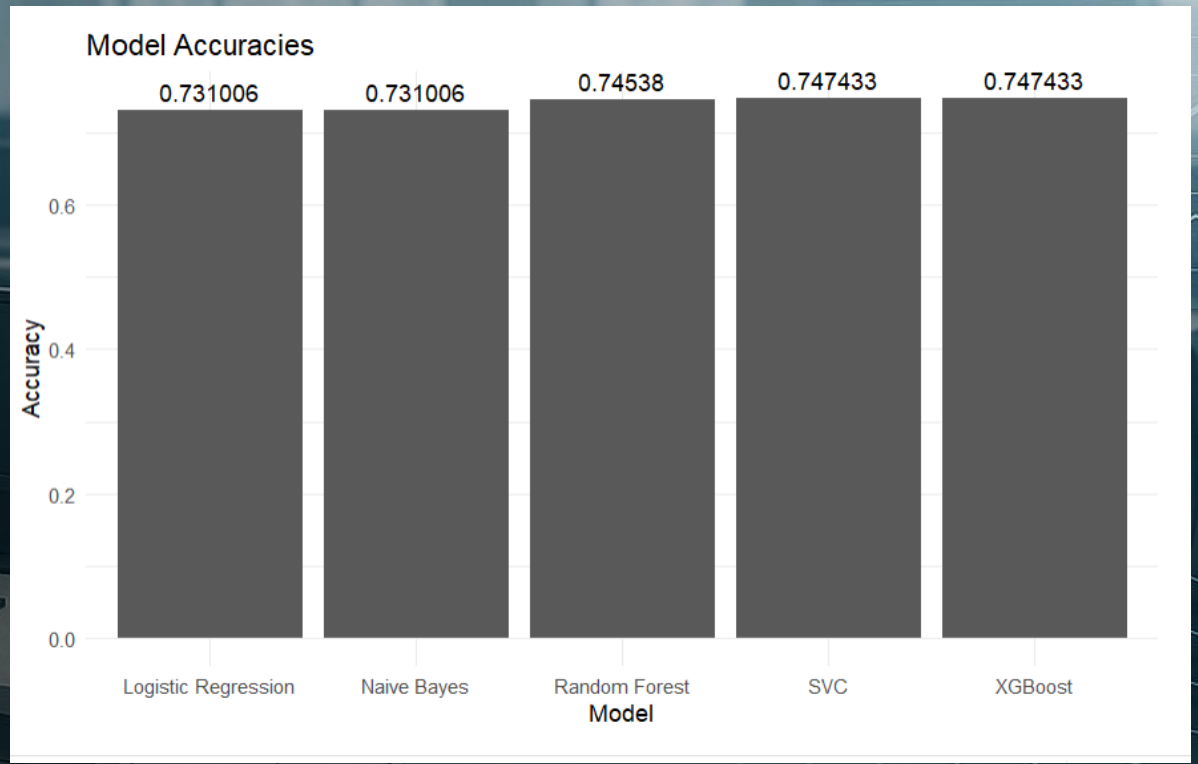
We used U.S. Bureau of Labor Statistics (BLS) ranges:

Amount	Category
< $30,000	Low Income
$30,000 to $49,999	Lower Middle Income
$50, 000 to $ 74,999	Middle Income
$75,000 to $99, 999	Upper Middle Income
$100,000+	High Income

- Regression Model 3
- Analyzing Regression Results
- Question Solved?
- Regression to Classification
- **Classification Results**
- Confusion Matrix

	Model <chr>	Accuracy <dbl>
1	SVC	0.7474333
4	XGBoost	0.7474333
2	Random Forest	0.7453799
3	Logistic Regression	0.7310062
5	Naive Bayes	0.7310062

5 rows

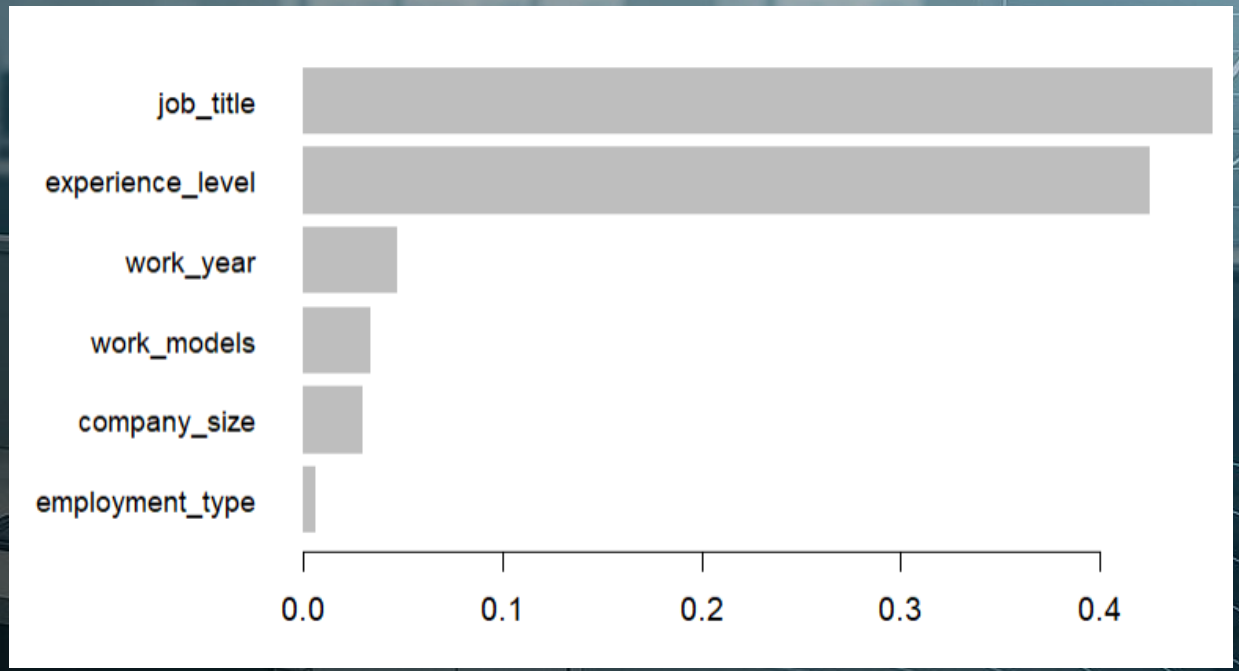


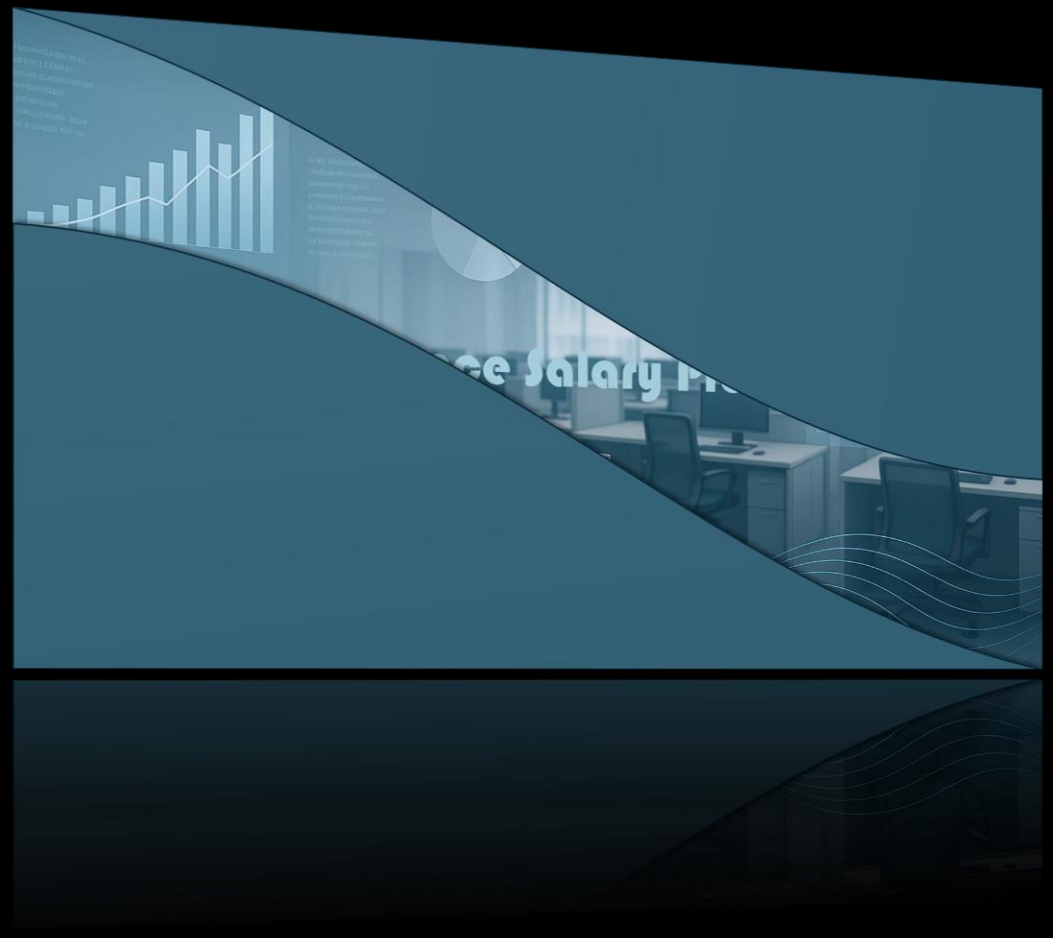
- Analyzing Regression Results
- Question Solved?
- Regression to Classification
- Classification Results
- **Confusion Matrix**

eXtreme Gradient Boosting (XGB)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2   3
         0   0   0   0   0
         1   0   0   0   2
         2   1   1   4   2
         3   3  37  77 360
```





Thank you

Presenters:

Sameer Batra

Jeongmin An

Yeobi Hobson

Aditi Shukla