AI-JOBS.NET

DATA SCIENCE JOB SALARIES

ELMIR DASHTIYEV I ELSHAN DASHTIYEV

01 DATA UNDERSTANDING (EDA)

02 DATA PREPARATION

03 MODEL SELECTIONS

04 EVALUATIONS AND COMPARISON

05 CONCLUSION

DATA UNDERSTANDING

- Formulating Questions
- Visualizing the Data
- Getting a better understanding and finding key data points

DATA UNDERSTANDING

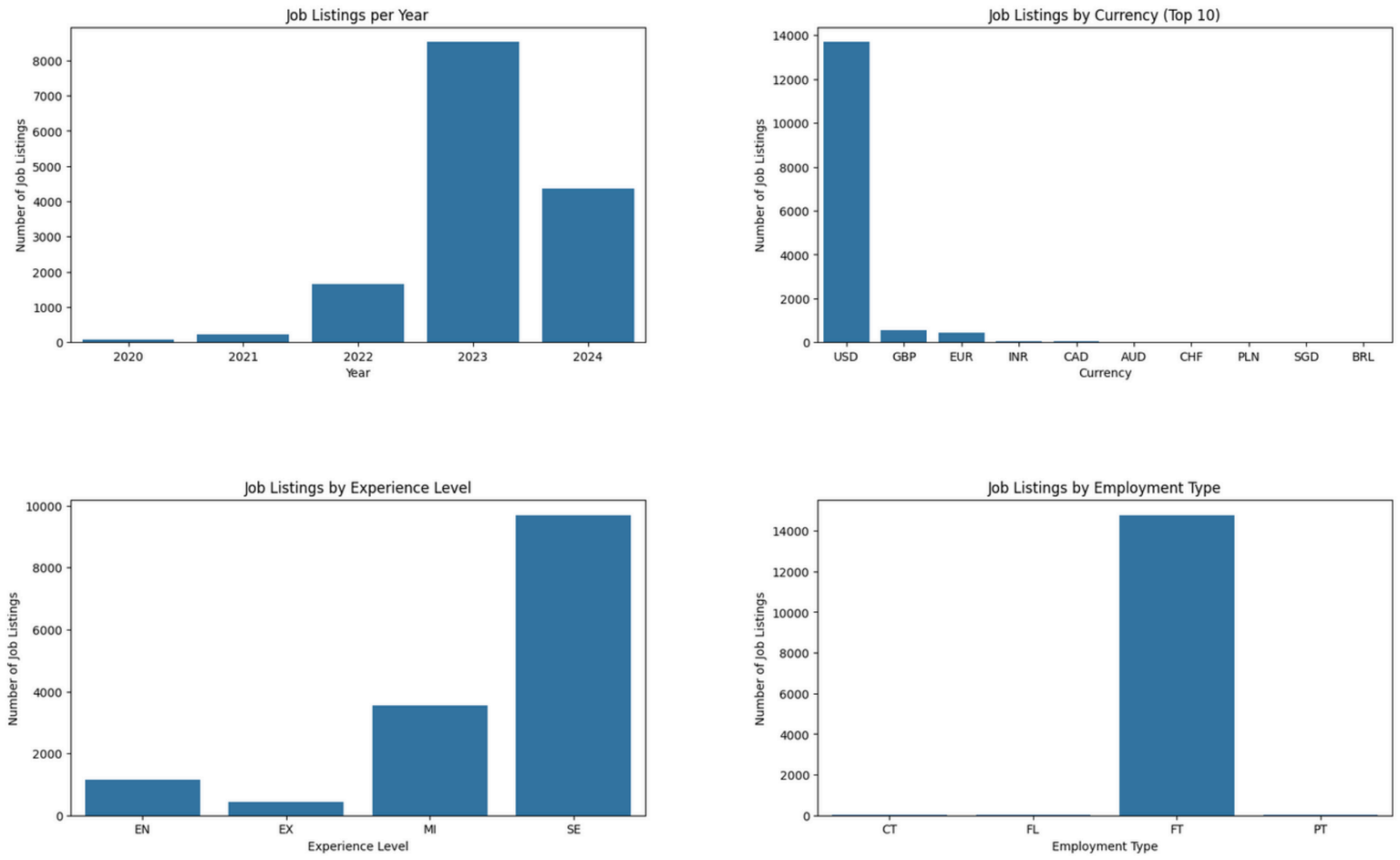
work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
2021	MI	FT	Data Scientist	30400000	CLP	40038	CL	100	CL	L
2021	MI	FT	BI Data Analyst	11000000	HUF	36259	HU	50	US	L
2020	MI	FT	Data Scientist	11000000	HUF	35735	HU	50	HU	L
2021	MI	FT	ML Engineer	8500000	JPY	77364	JP	50	JP	S
2022	SE	FT	Lead Machine Learning Engineer	7500000	INR	95386	IN	50	IN	L

Total number of job listings: 14838

Average salary: $149874.72

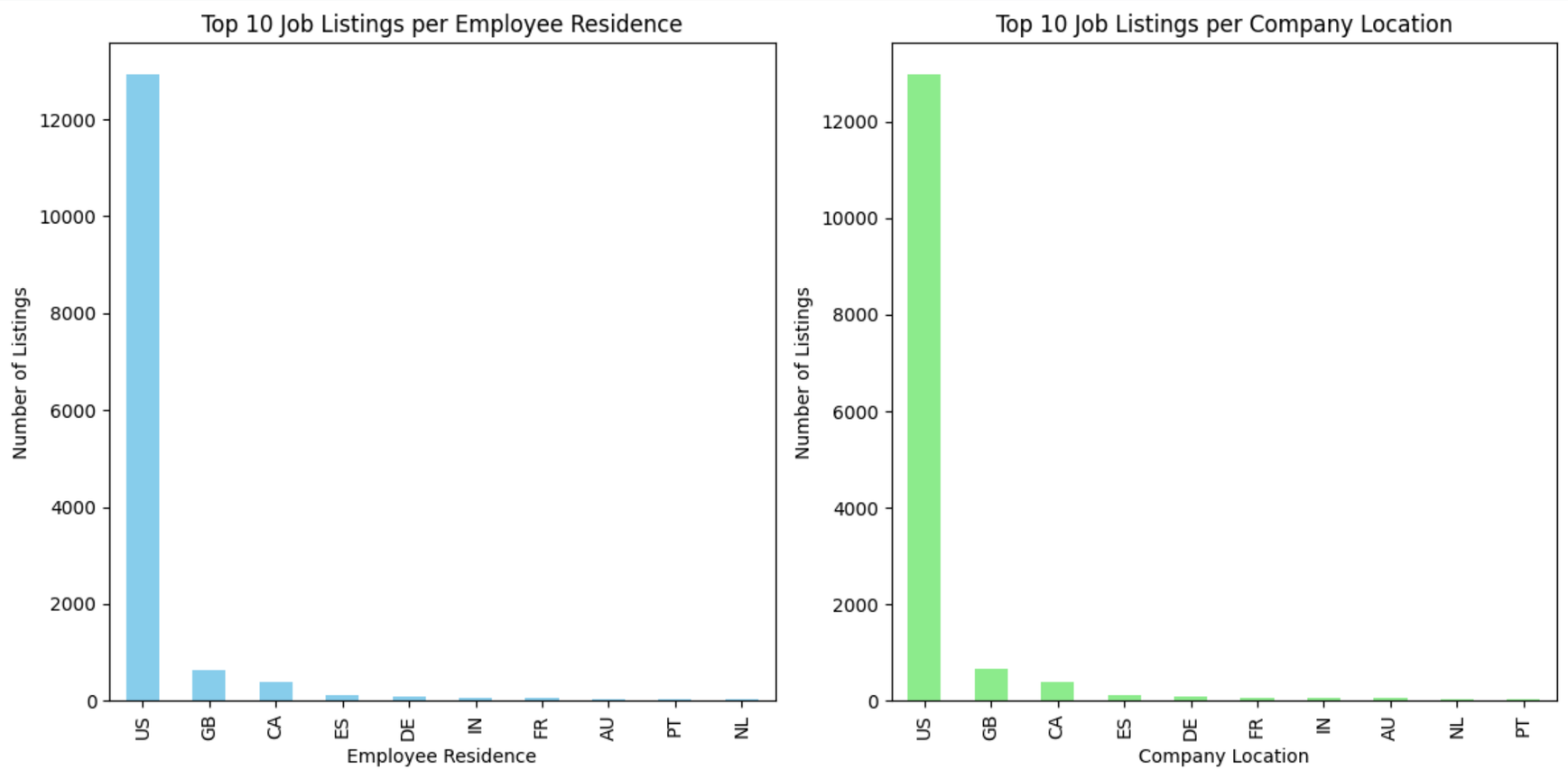
DATA UNDERSTANDING

Job listings by different data types



DATA UNDERSTANDING

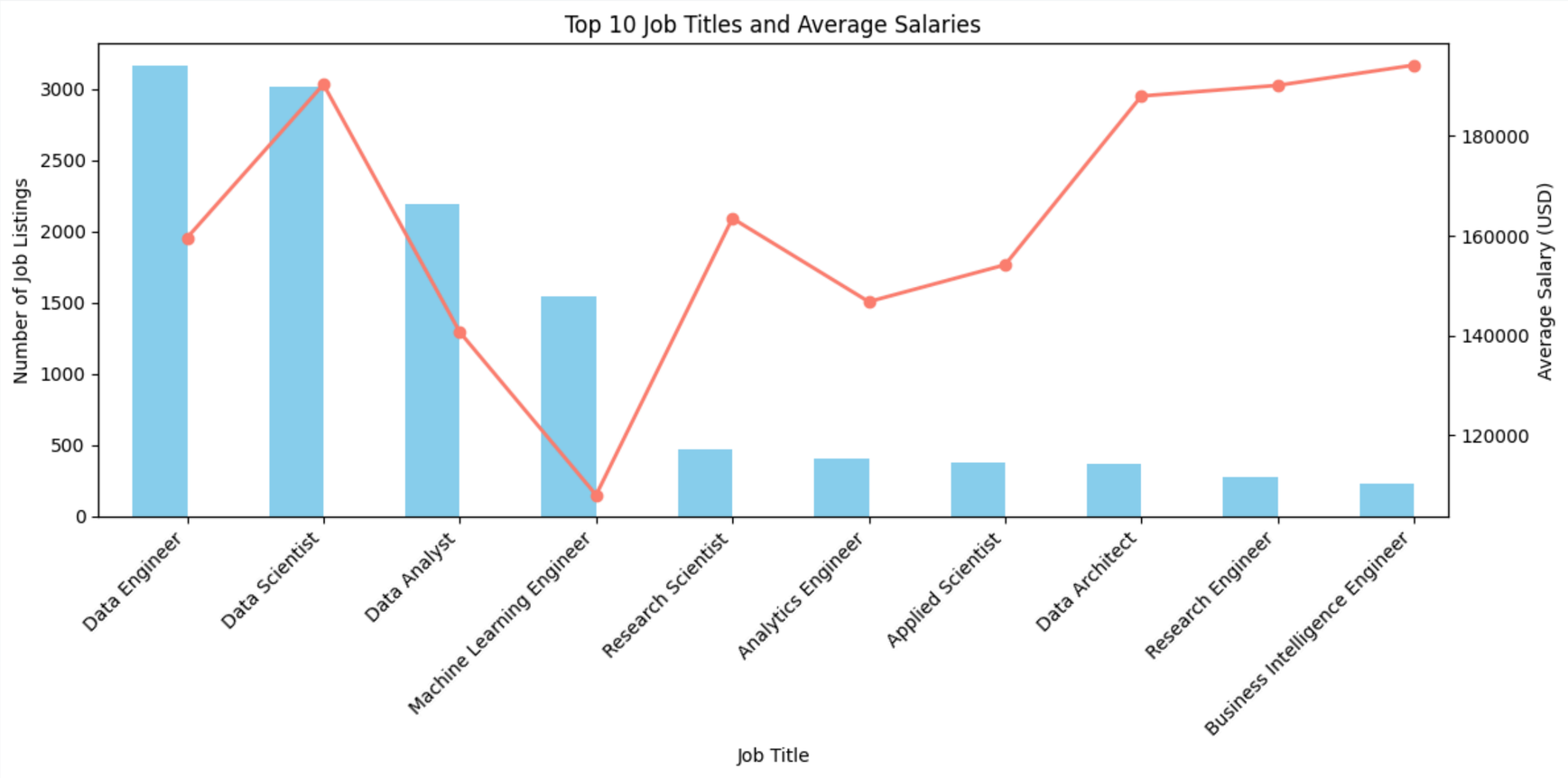
Comparing Employee Residence and Company Location



AI-JOBS.NET

DATA UNDERSTANDING

Job listings and salaries for different job titles



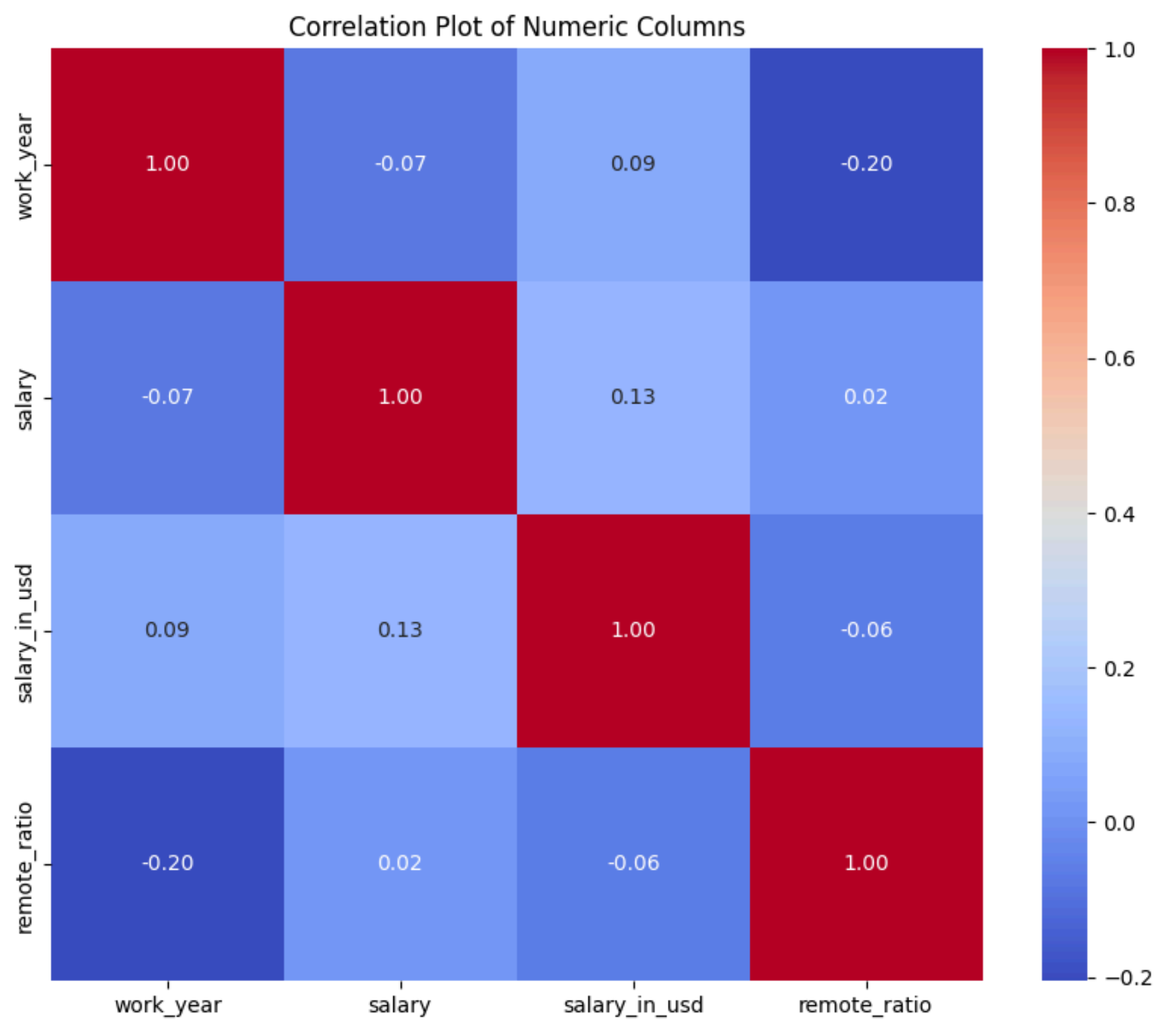
AI-JOBS.NET

DATA UNDERSTANDING

Avarage salary by different data types

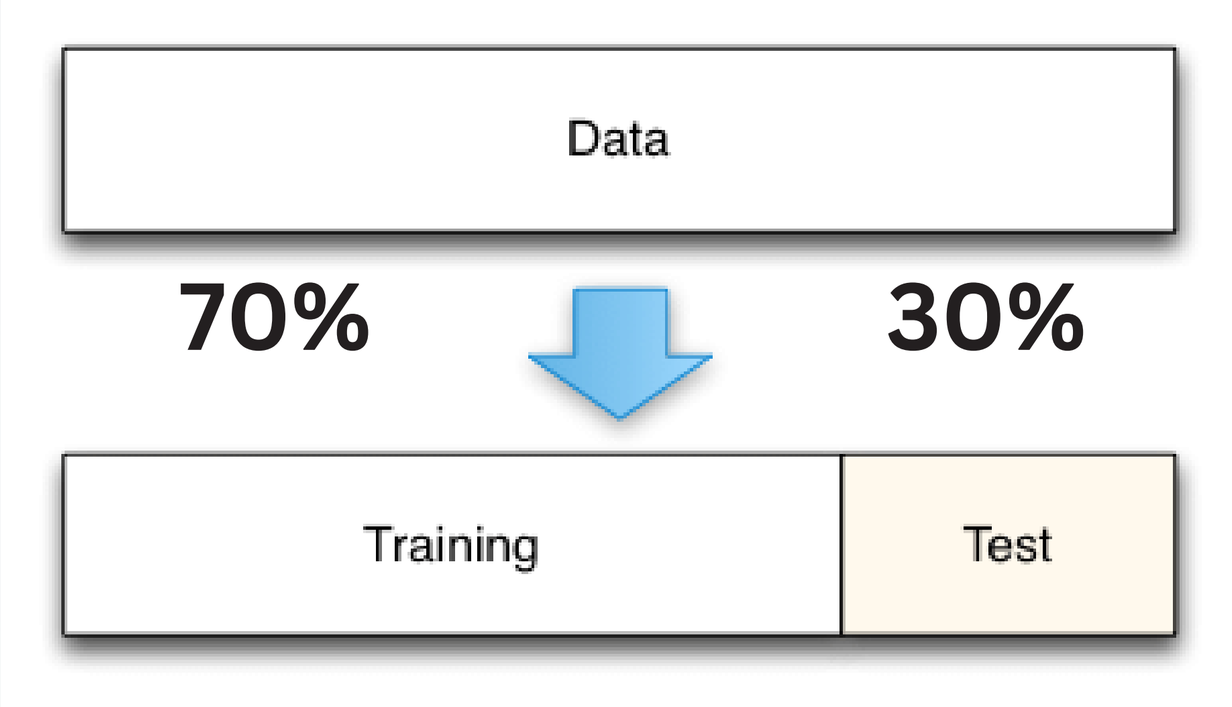
DATA UNDERSTANDING

Correlation Plot



DATA PREPARATION

- 11 Features - 4 numerical, 7 catogorical
- get_dummies & one-hot encoder used for catogorical features
- Dropped “salary”
- Target: “salary_in_usd”

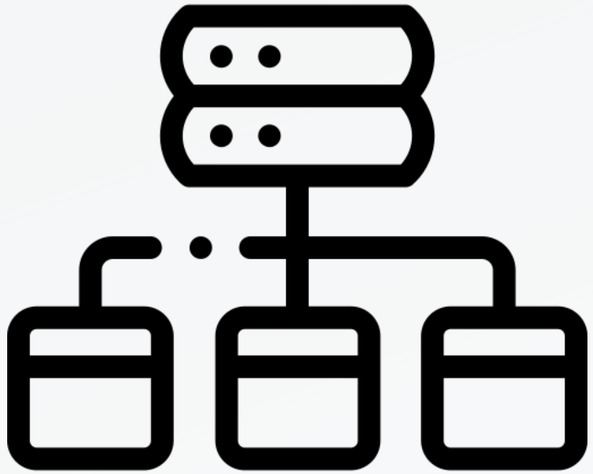


MODEL SELECTION

- Linear Regression
  Ridge & Lasso

- RandomForest
  With & without Gridsearch

- XGBoost



$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

MODEL SELECTION

Linear Regression

```
features = ['work_year', 'experience_level', 'employment_type', 'salary_currency']
target = 'salary_in_usd'
```

Result:

```
Train RMSE: 61772.17
Train R2: 0.20
Test RMSE: 61307.43
Test R2: 0.20
```

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \ldots + \beta_n \cdot x_n$$

MODEL SELECTION

Linear Regression: Ridge & Lasso

GridSearch done for both regularizations

```
param_grid = {
    'alpha': [0.1, 1.0, 10.0, 100.0, 1000.0]
}
```

Ridge

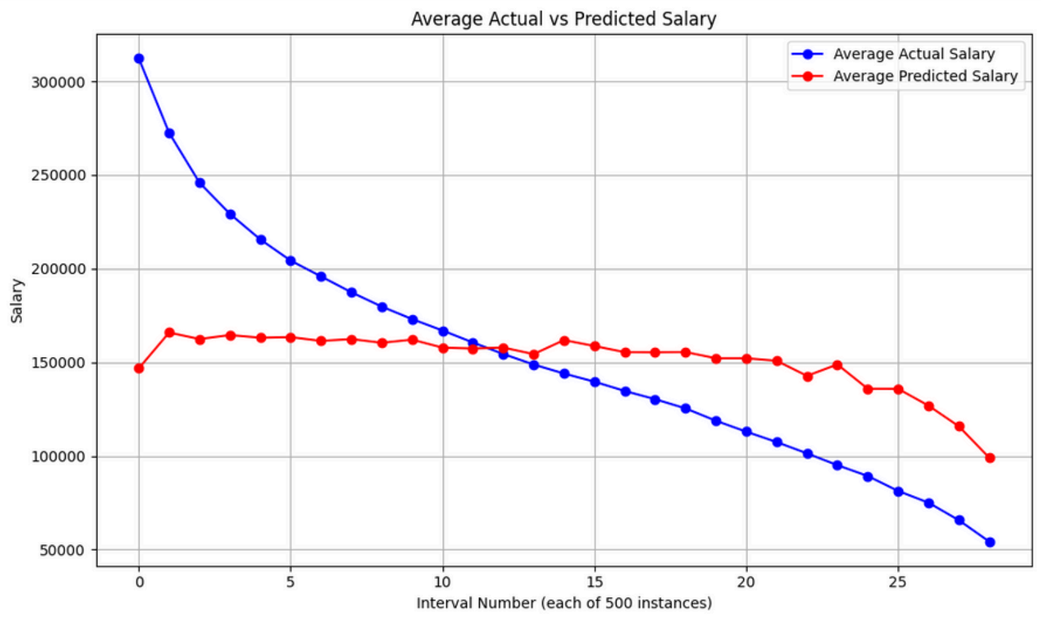
```
Train RMSE (Ridge): 61772.17
Train R2 (Ridge): 0.20
Test RMSE (Ridge): 61307.41
Test R2 (Ridge): 0.20
Best hyperparameter (alpha): 0.1
```

Lasso

```
Train RMSE (Lasso): 61777.18
Train R2 (Lasso): 0.20
Test RMSE (Lasso): 61292.56
Test R2 (Lasso): 0.20
Best hyperparameter (alpha): 100.0
```

MODEL SELECTION

Linear Regression: Visualization



MODEL SELECTION

RandomForest

```
features = ['work_year', 'experience_level', 'employment_type', 'job_title', 'salary_currency', 'employee_residence']
target = 'salary_in_usd'
```

```
param_grid = {
    'model__n_estimators': [100, 200, 300],
    'model__max_depth': [None, 10, 20, 30],
    'model__min_samples_split': [2, 5, 10],
    'model__min_samples_leaf': [1, 2, 4]
}
```

With GridSearch

```
Train RMSE: 54452.22
Train R2: 0.38
Test RMSE: 55971.40
Test R2: 0.34
```

Without Gridsearch

```
Train RMSE: 52772.18
Train R2: 0.42
Test RMSE: 56371.71
Test R2: 0.33
```

MODEL SELECTION

RandomForest: Removing Outliers (IQR)

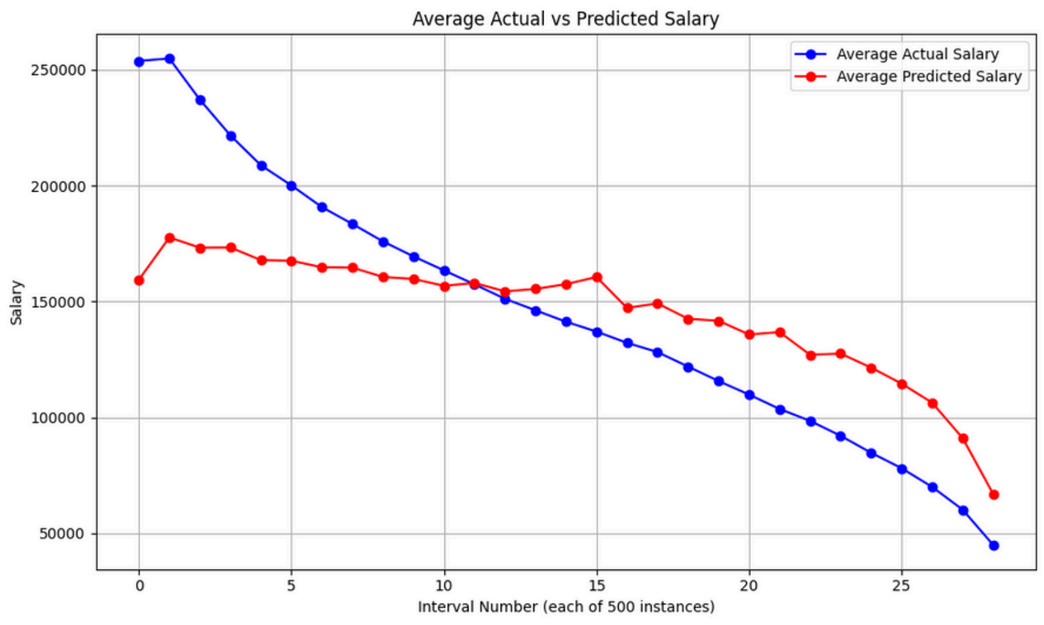
```
Q1 = df[target].quantile(0.25)
Q3 = df[target].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
df = df[(df[target] >= lower_bound) & (df[target] <= upper_bound)]
```

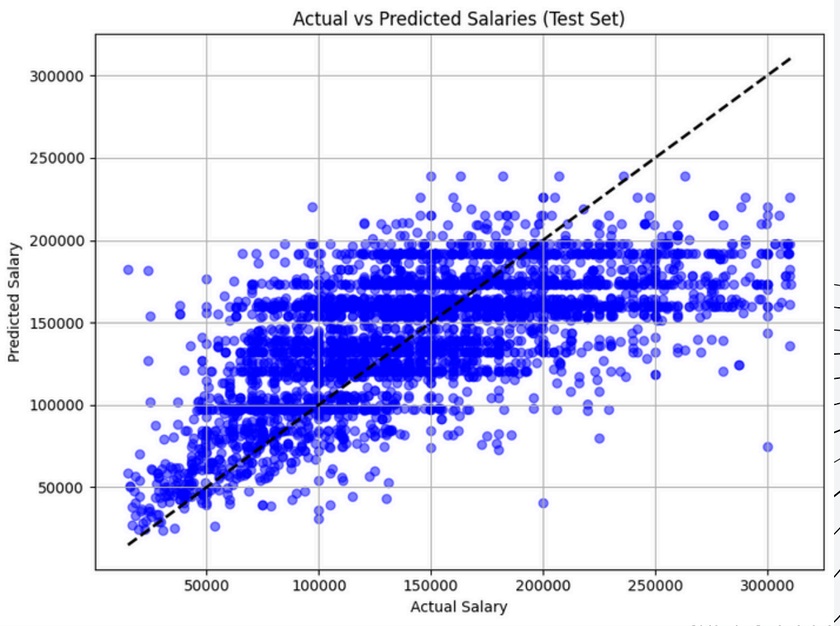
Result:

```
Train RMSE: 44926.35
Train R2: 0.43
Test RMSE: 46264.28
Test R2: 0.39
```

MODEL SELECTION

Random Forest: Visualization





MODEL SELECTION

XGBoost

```
features = ['work_year', 'experience_level', 'employment_type', 'job_title', 'salary_currency', 'employee_residence']
target = 'salary_in_usd'
```

With GridSearch

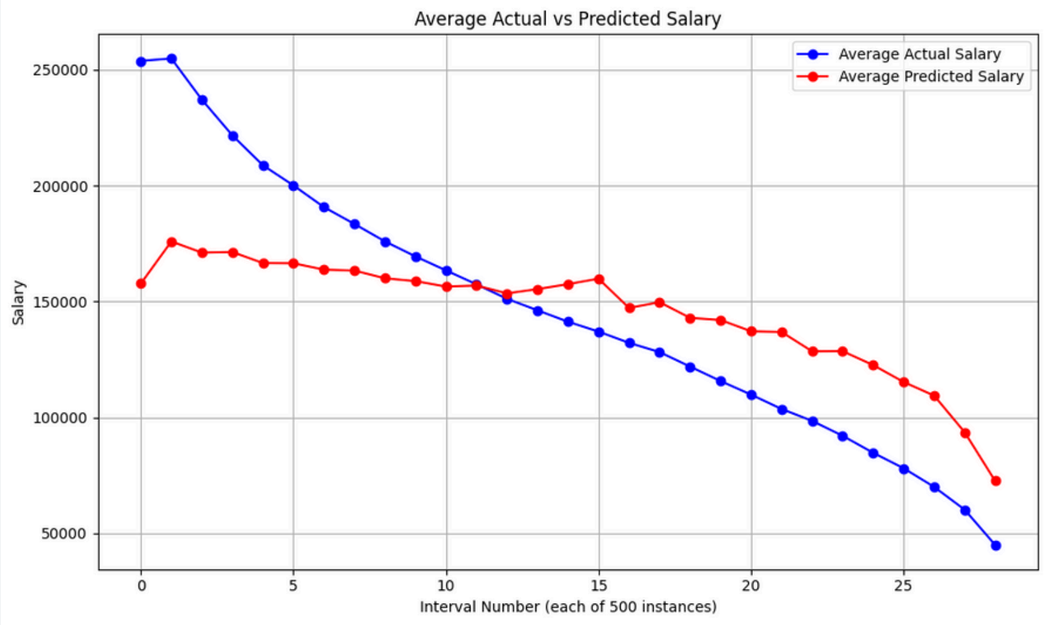
```
Train RMSE: 45292.78
Train R2: 0.42
Test RMSE: 46218.45
Test R2: 0.40
```

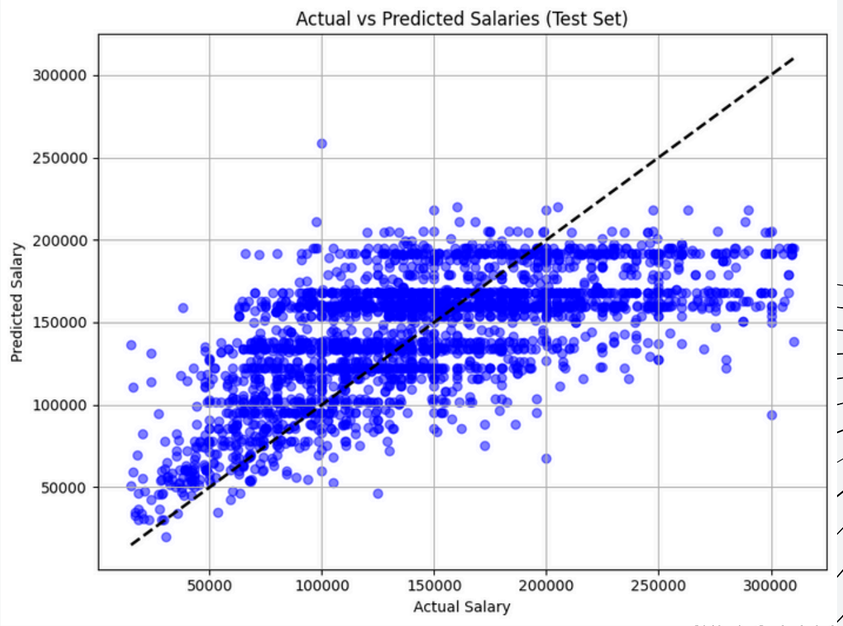
Without Gridsearch

```
Train RMSE: 44655.59
Train R2: 0.43
Test RMSE: 46084.66
Test R2: 0.41
```

MODEL SELECTION

XGBoost: Visualization





EVALUATION & COMPARISON

All Models

Linear Regression

```
Train RMSE: 61772.17
Train R2: 0.20
Test RMSE: 61307.43
Test R2: 0.20
```

Random Forest

```
Train RMSE: 44926.35
Train R2: 0.43
Test RMSE: 46264.28
Test R2: 0.39
```

XGBoost

```
Train RMSE: 44655.59
Train R2: 0.43
Test RMSE: 46084.66
Test R2: 0.41
```

THANKS FOR YOUR ATTENTION