Ironhack final project

Job market analysis

Content

Introduction

Dataset and EDA

Data visualisation

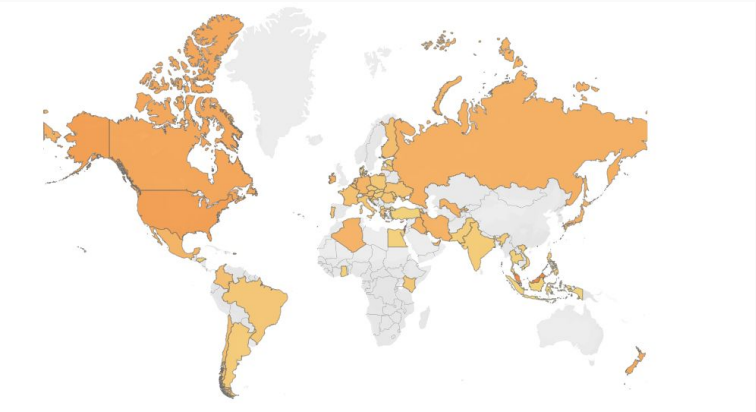
SQL queries

Machine learning / Streamlit UI

Conclussions / further research

Introduction

The intention of this project is providing a thorough analysis of a dataset regarding it-related job offerings, where we analyse relevant aspects such as the job title, location and gross average salary.



Dataset

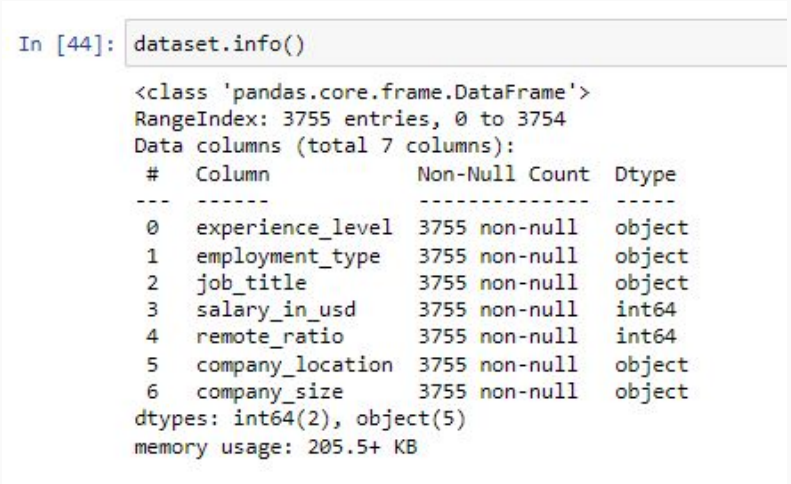
3755 observations

Total of 93 unique role names

72 locations worldwide

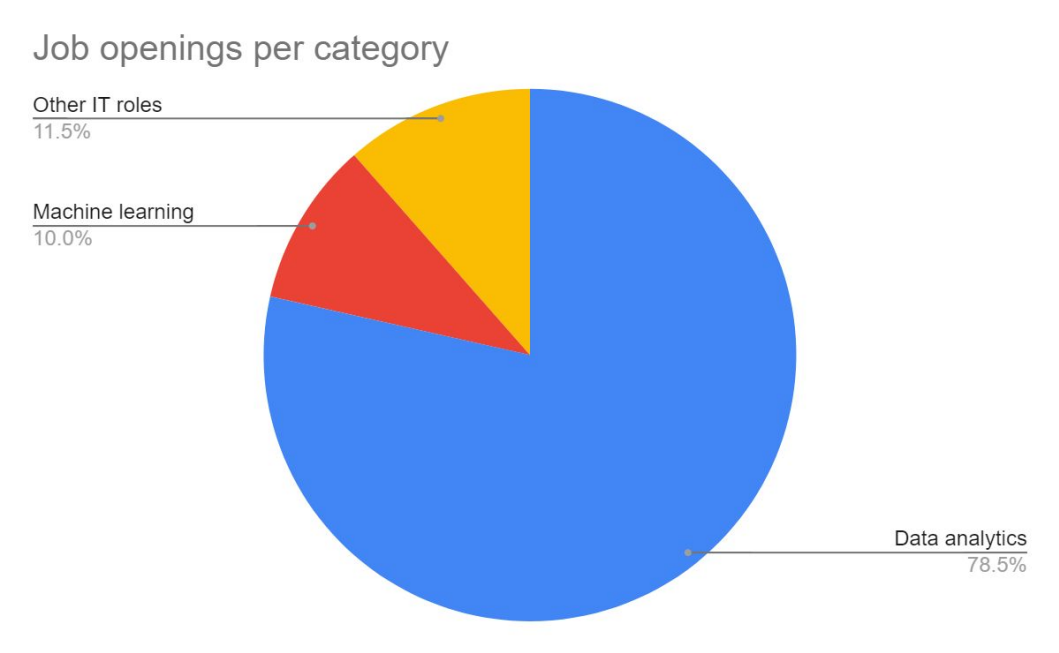
Source:

kaggle

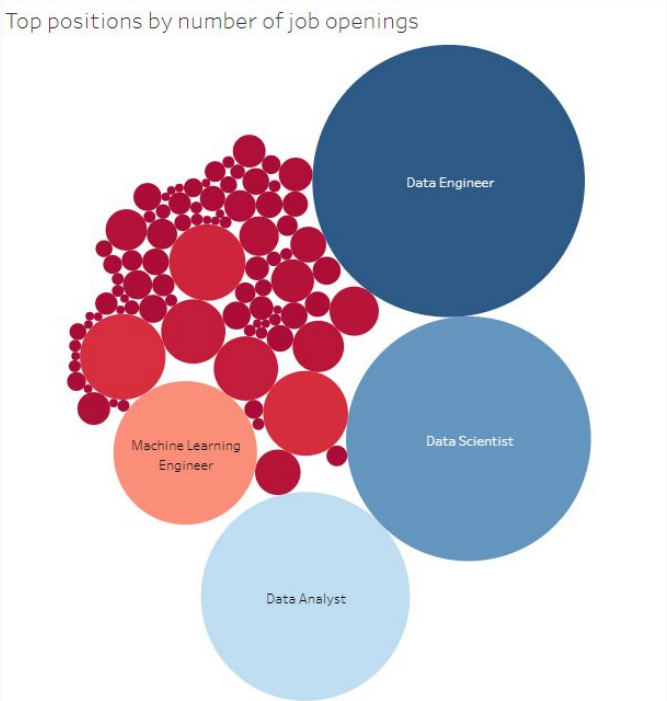


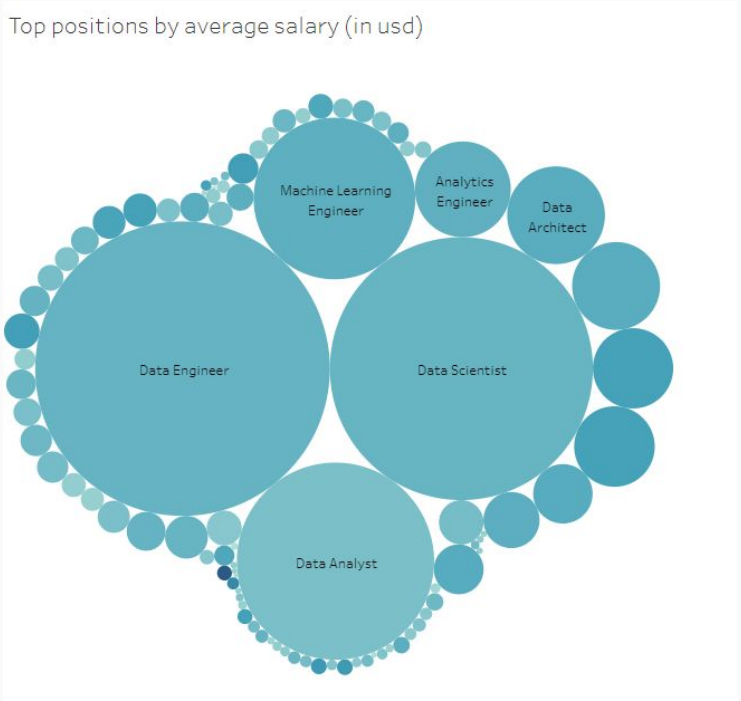
```
In [44]: dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3755 entries, 0 to 3754
Data columns (total 7 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   experience_level  3755 non-null   object
 1   employment_type   3755 non-null   object
 2   job_title         3755 non-null   object
 3   salary_in_usd     3755 non-null   int64 
 4   remote_ratio      3755 non-null   int64 
 5   company_location  3755 non-null   object
 6   company_size      3755 non-null   object
dtypes: int64(2), object(5)
memory usage: 205.5+ KB
```

Distribution of job openings



Data visualisation: count and average





Avg. salary: company size and exp. Level

Experience level:

- Entry level (EN)
- Middle (MI)
- Senior (SE)
- Experienced / Manager (EX)

Dataset

Out[46]:

	salary_in_usd	remote_ratio
0	85847	100
1	30000	100
2	25500	100
3	175000	100
4	120000	100
...	...	...
3750	412000	100
3751	151000	100
3752	105000	100
3753	100000	100
3754	94665	50

3755 rows × 2 columns

Num. variables

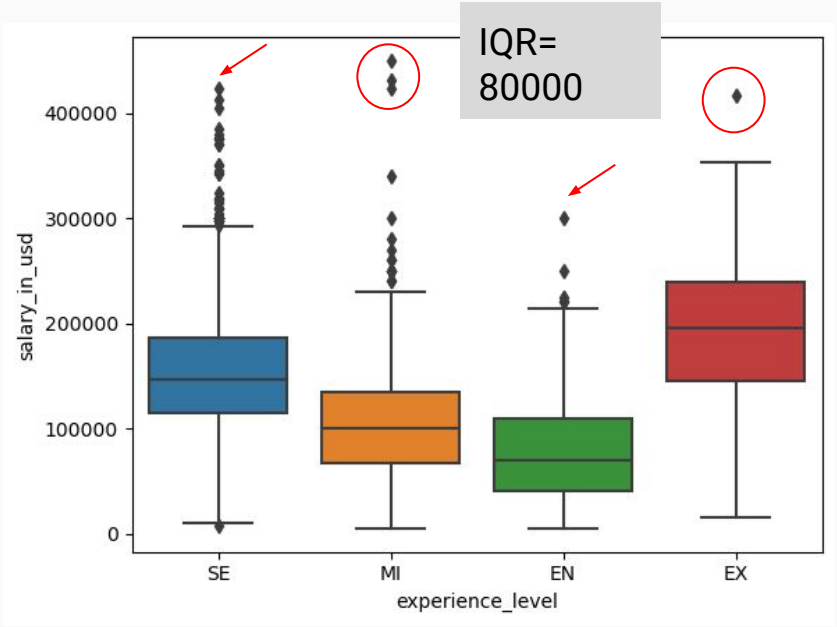
Out[47]:

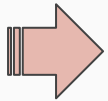
	experience_level	employment_type	job_title	company_location	company_size
0	SE	FT	Principal Data Scientist	ES	L
1	MI	CT	ML Engineer	US	S
2	MI	CT	ML Engineer	US	S
3	SE	FT	Data Scientist	CA	M
4	SE	FT	Data Scientist	CA	M
...	...	...	...	...	...
3750	SE	FT	Data Scientist	US	L
3751	MI	FT	Principal Data Scientist	US	L
3752	EN	FT	Data Scientist	US	S
3753	EN	CT	Business Data Analyst	US	L
3754	SE	FT	Data Science Manager	IN	L

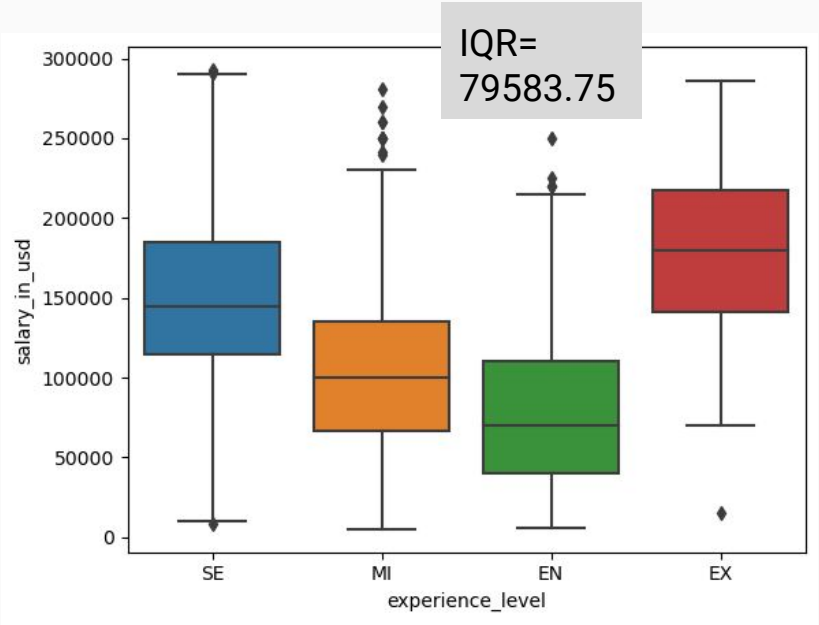
3755 rows × 5 columns

Cat. variables

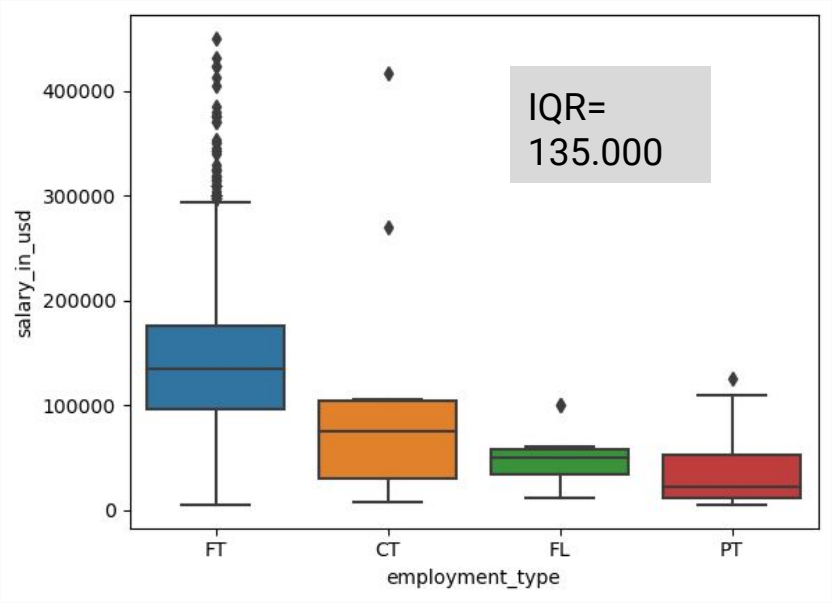
EDA comparison after removing outliers

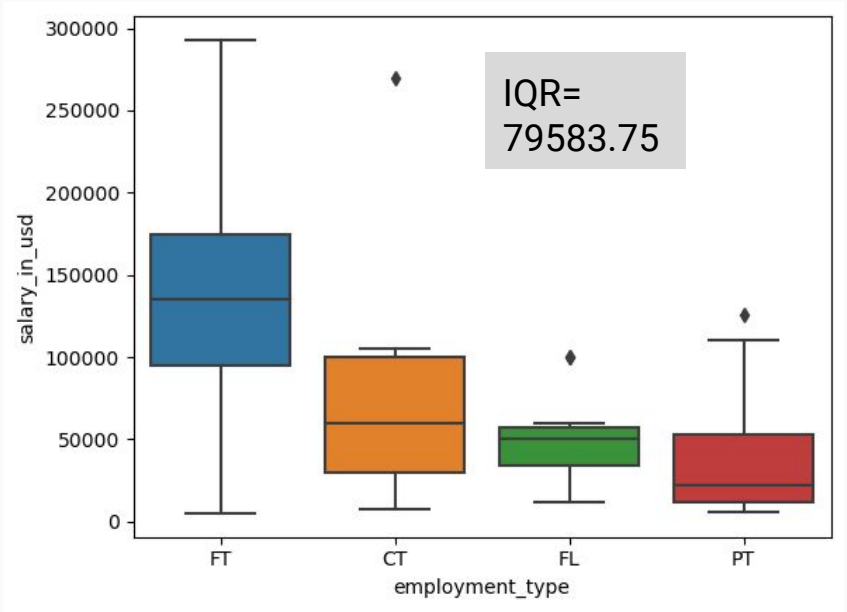




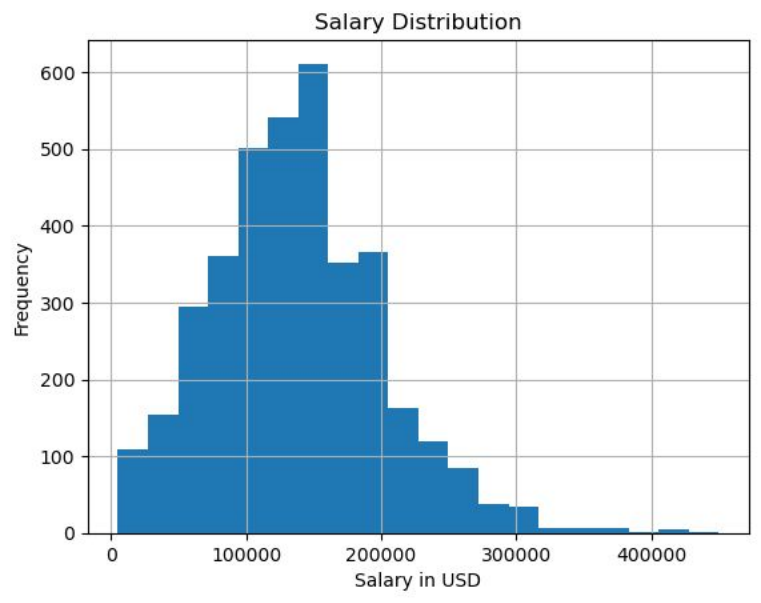


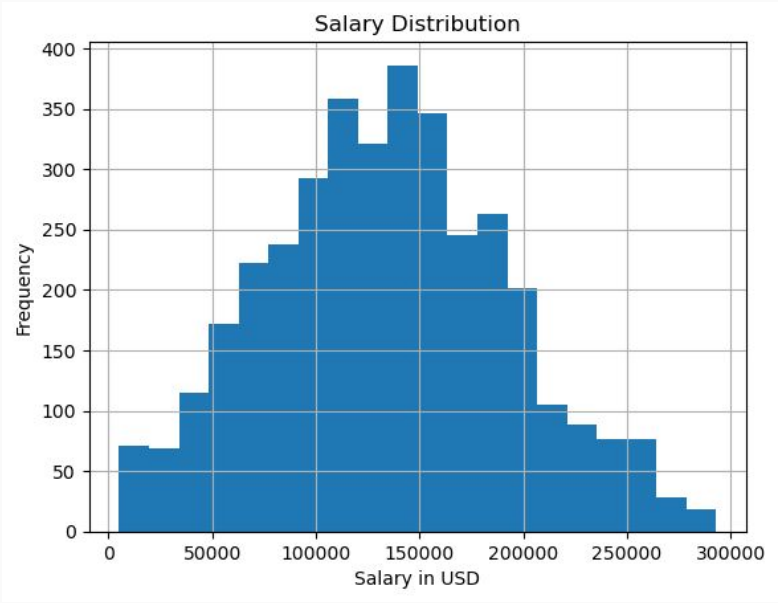
EDA comparison after removing outliers





EDA comparison after rmv. outliers





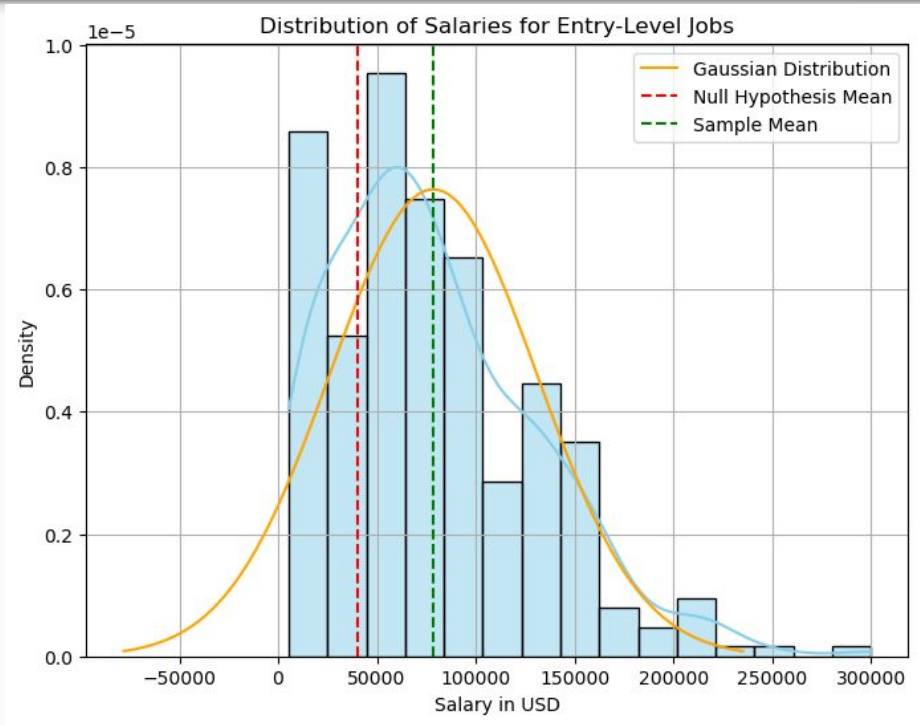
EDA comparison after removing outliers

	salary_in_usd	remote_ratio
count	3755.000000	3755.000000
mean	137570.389880	46.271638
std	63055.625278	48.589050
min	5132.000000	0.000000
25%	95000.000000	0.000000
50%	135000.000000	0.000000
75%	175000.000000	100.000000
max	450000.000000	100.000000

	salary_in_usd	remote_ratio
count	3692.000000	3692.000000
mean	134262.993770	46.289274
std	57992.294349	48.589320
min	5132.000000	0.000000
25%	94916.250000	0.000000
50%	133916.000000	0.000000
75%	174500.000000	100.000000
max	293000.000000	100.000000

Hypothesis testing

Perform a hypothesis testing where we assume that **the average salary for an entry level job** (experience_level = 'EN') **is at least** (higher than or equal to) **40.000** (salary_in_usd). Note: this is a one-tail t-test where we take a **significance level of 0.05** for our hypothesis testing.



Machine learning modelling

- INITIAL MACHINE LEARNING MODEL
- MACHINE LEARNING **MIN-MAX** SCALER
- MACHINE LEARNING **NORMALISATION**
- HYPERPARAMETER **TUNING**
- **K-NEAREST**
- **RANDOM FOREST**

Initial machine learning modeling

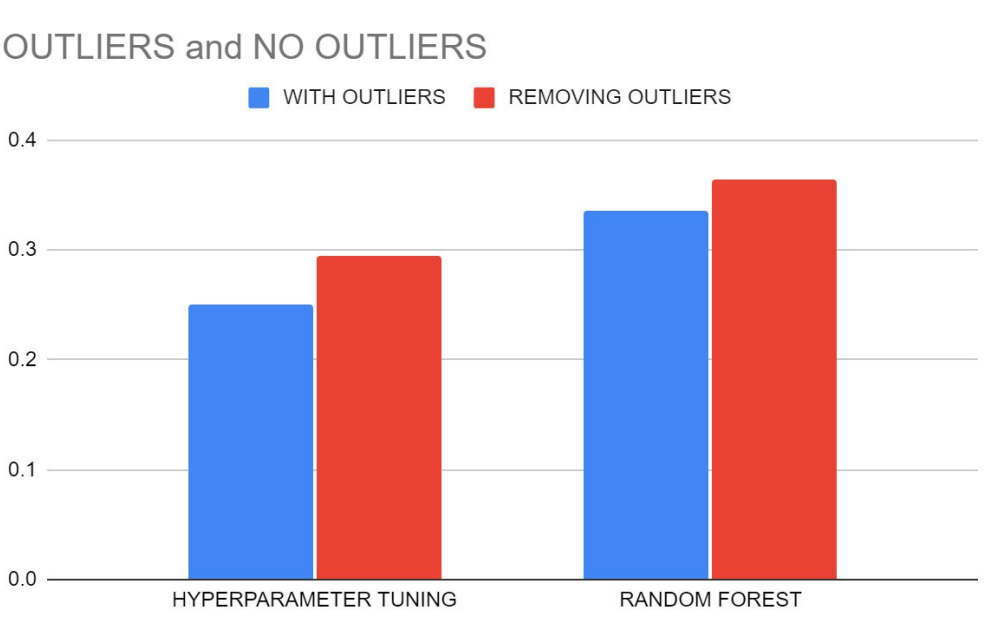
```
X = dataset[['experience_level', 'employment_type', 'remote_ratio', 'company_size', 'company_location']]
y = dataset['salary_in_usd']

X_encoded = pd.get_dummies(X, columns=['experience_level', 'employment_type', 'company_size', 'company_location'])

X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

```
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

print("Mean Absolute Error (MAE):", mae)
print("Mean Squared Error (MSE):", mse)
print("Root Mean Squared Error (RMSE):", rmse)
print("R-squared (R^2):", r2)
```

```
Mean Absolute Error (MAE): 42973770819.71571
Mean Squared Error (MSE): 3.4677846012106515e+23
Root Mean Squared Error (RMSE): 588878985973.4045
R-squared (R^2): -87841165456152.3
```

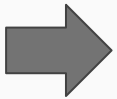
Treated “salary_in_usd” as the target variable to get an overview of the salary pay irrespective of the company’s location (there is a total of 73 locations where they use different currencies).

Machine learning model performance evaluation



SQL DB Queries

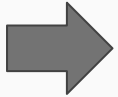
```
SELECT * FROM job_openings_db.`ds_salaries (1)`;

# Provide a top 20 of the entry-level positions with the highest pay in the year 2022

SELECT job_title, salary_in_usd FROM
job_openings_db.`ds_salaries (1)`

WHERE experience_level = 'EN' AND work_year = 2022

ORDER BY salary_in_usd DESC

LIMIT 20;
```



job_title	salary_in_usd
AI Developer	300000
AI Scientist	200000
Machine Learning Engineer	189750
Data Scientist	180000
Machine Learning Developer	180000
Data Scientist	180000
Data Scientist	168000
Data Engineer	160000
Data Engineer	160000
Data Engineer	160000
Data Engineer	160000
Data Engineer	160000
Data Engineer	160000
Data Analyst	150000
Computer Vision Software ...	150000
Machine Learning Engineer	140250
Data Engineer	135000

SQL DB Queries

```
# Calculate the average salary in EURO for
a machine learning engineer middle role

SELECT AVG(salary) AS average_salary

FROM job_openings_db.`ds_salaries (1)`

WHERE job_title = 'Machine Learning
Engineer' AND experience_level = 'MI' AND
salary_currency = 'EUR'
```

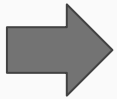


	average_salary
▶	58200

SQL DB Queries

Select the job openings in Spain that are not 'experienced level' (EX) and provide the first 20 positions, order by job title

```
SELECT salary, job_title, experience_level, company_size
FROM job_openings_db.`ds_salaries (1)`

WHERE company_location = 'ES' AND experience_level
!='EX'

ORDER BY job_title ASC

LIMIT 20;
```



salary	job_title	experience_level	company_size
55000	AI Scientist	SE	L
36000	AI Scientist	MI	L
30000	AI Scientist	EN	M
45000	Big Data Engineer	MI	M
20000	Business Data Analyst	EN	M
38000	Data Analyst	SE	M
38000	Data Analyst	SE	M
48000	Data Analyst	SE	M
35000	Data Analyst	SE	M
48000	Data Analyst	SE	M
35000	Data Analyst	SE	M
48000	Data Analyst	SE	M
35000	Data Analyst	SE	M
48000	Data Analyst	SE	M
38000	Data Analyst	SE	M
52000	Data Analyst	SE	M
48000	Data Analyst	SE	M

Streamlit user interface

URL to the application deployed:

https://job-market-analysis.streamlit.app/

Github repository:

https://github.com/carlosruiz-stack/streamlit_app