Capstone Project

Jackson Staffing Agency

MELVIN LEWIS | DATA ENGINEER

Agenda

Bio

Project context

Define

Design

Deliver

Summary, Conclusions, & next steps

Bio

Education : Virginia State University - Information Logistics Technology

Data science experience:

Institute of Data (*Virginia Commonwealth University* Accredited)
Feb 2024

Project Context:

In the context of our employment placement agency, we aim to optimize our approach to data science job placements by strategically harnessing information related to salary structures, experience levels, roles, and job industries within the data science domain. This project seeks to explore how data-driven insights can be utilized to refine our recruitment and placement strategies, ultimately enhancing our ability to connect qualified candidates with relevant opportunities in the ever-evolving data science sector.

Industry: Employee Placement

Problem Area:
The current challenge lies in efficiently utilizing data pertaining to salary structures, experience levels, roles, and job industries to determine what compensation package we can give a new candidate based on salaries given to us in this dataset.

Define

- Business aspects
  - Business Question: How can our staffing agency use available data to improve recruitment, matching, and outcomes for clients and candidates?
- Data science aspects
  - Data Description :

work_year: The year in which the data was recorded. This field indicates the temporal context of the data, important for understanding salary trends over time.
job_title: The specific title of the job role, like 'Data Scientist', 'Data Engineer', or 'Data Analyst'. This column is crucial for understanding the salary distribution across various specialized roles within the data field.

job_category: A classification of the job role into broader categories for easier analysis. This might include areas like 'Data Analysis', 'Machine Learning', 'Data Engineering', etc.

salary_currency: The currency in which the salary is paid, such as USD, EUR, etc. This is important for currency conversion and understanding the actual value of the salary in a global context.

salary: The annual gross salary of the role in the local currency. This raw salary figure is key for direct regional salary comparisons.

salary_in_usd: The annual gross salary converted to United States Dollars (USD). This uniform currency conversion aids in global salary comparisons and analyses.

employee_residence: The country of residence of the employee. This data point can be used to explore geographical salary differences and cost-of-living variations.

experience_level: Classifies the professional experience level of the employee. Common categories might include 'Entry-level', 'Mid-level', 'Senior', and 'Executive', providing insight into how experience influences salary in data-related roles.

employment_type: Specifies the type of employment, such as 'Full-time', 'Part-time', 'Contract', etc. This helps in analyzing how different employment arrangements affect salary structures.

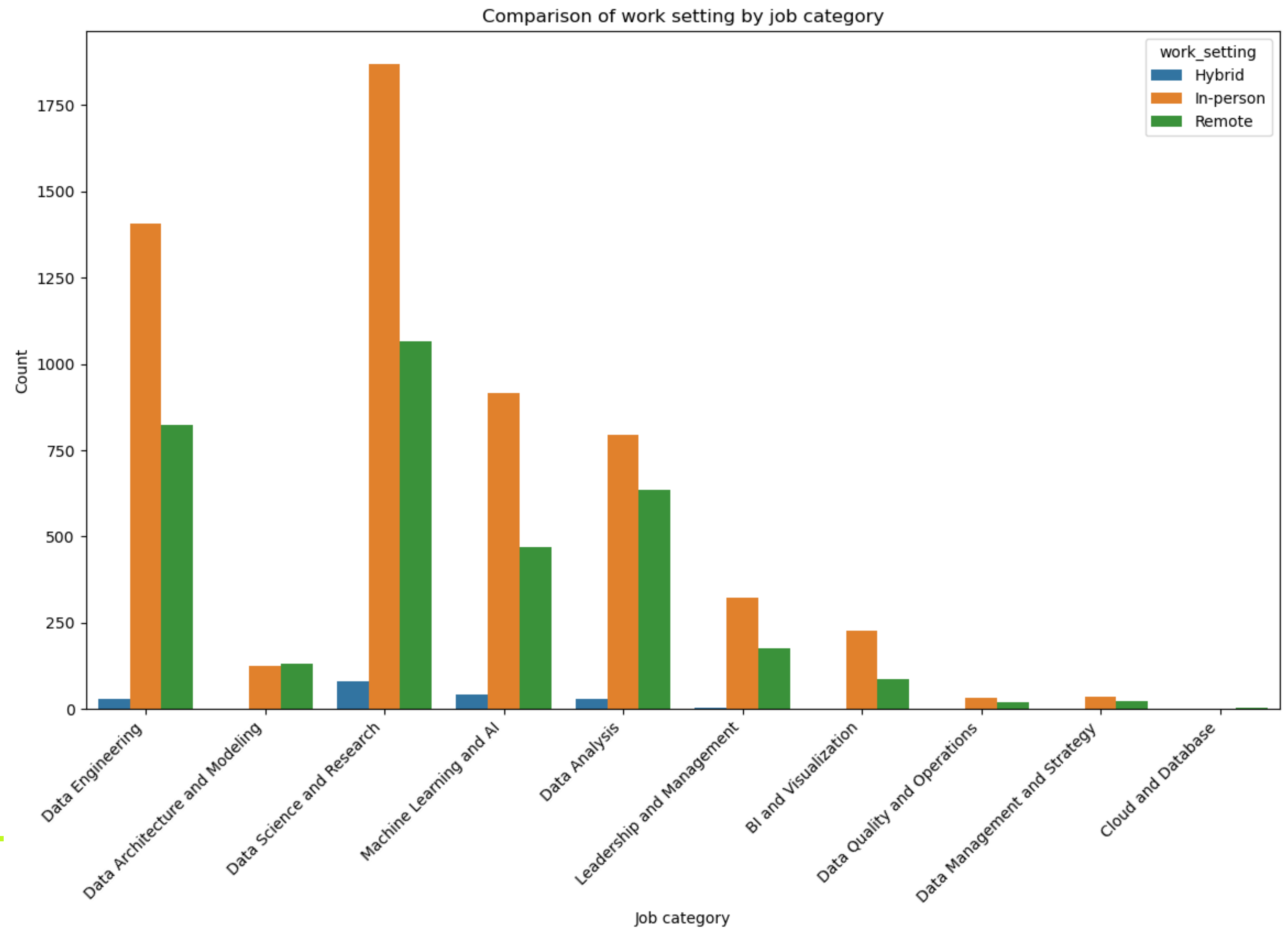
work_setting: The work setting or environment, like 'Remote', 'In-person', or 'Hybrid'. This column reflects the impact of work settings on salary levels in the data industry.

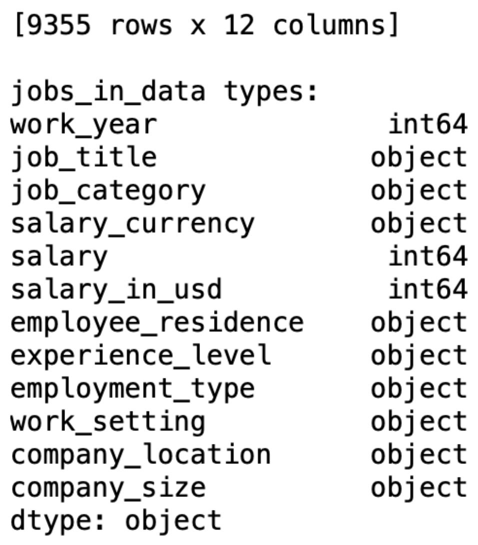
company_location: The country where the company is located. It helps in analyzing how the location of the company affects salary structures.

company_size: The size of the employer company, often categorized into small (S), medium (M), and large (L) sizes. This allows for analysis of how company size influences salary.

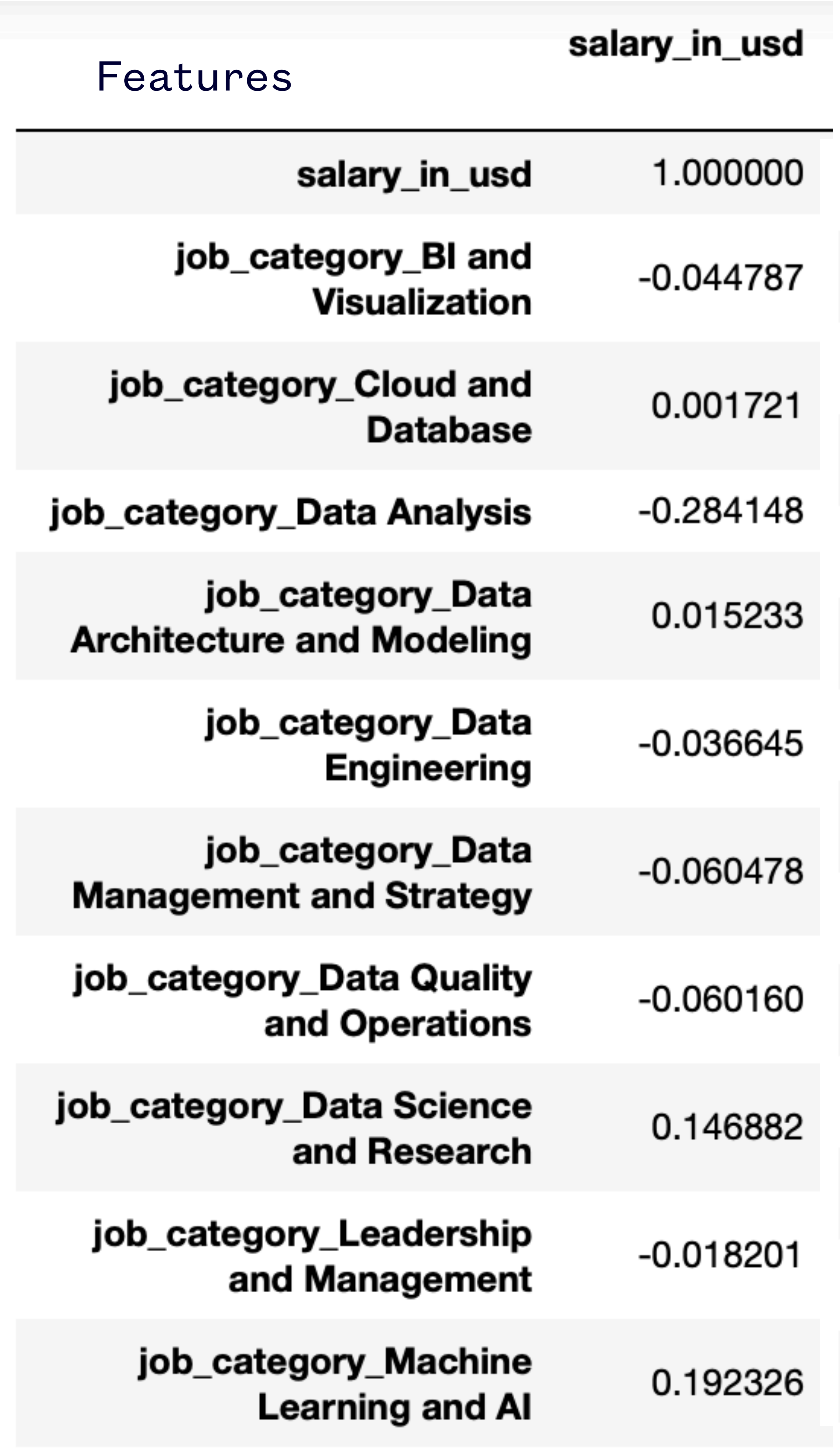
Design

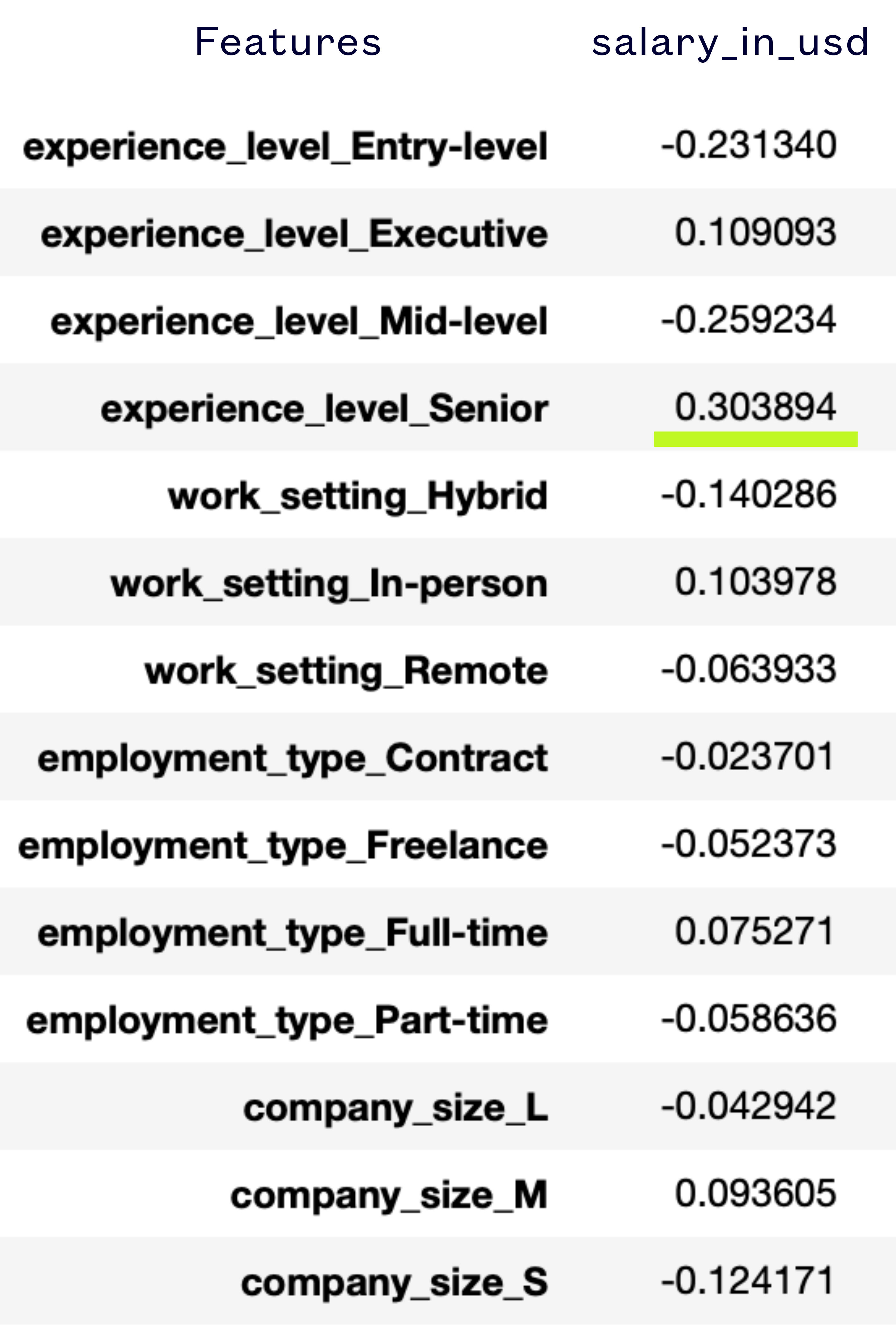
- Data exploration, analysis and visualization
  - We will explore more of EDA while walking through the coding.


Comparison of work setting by job category

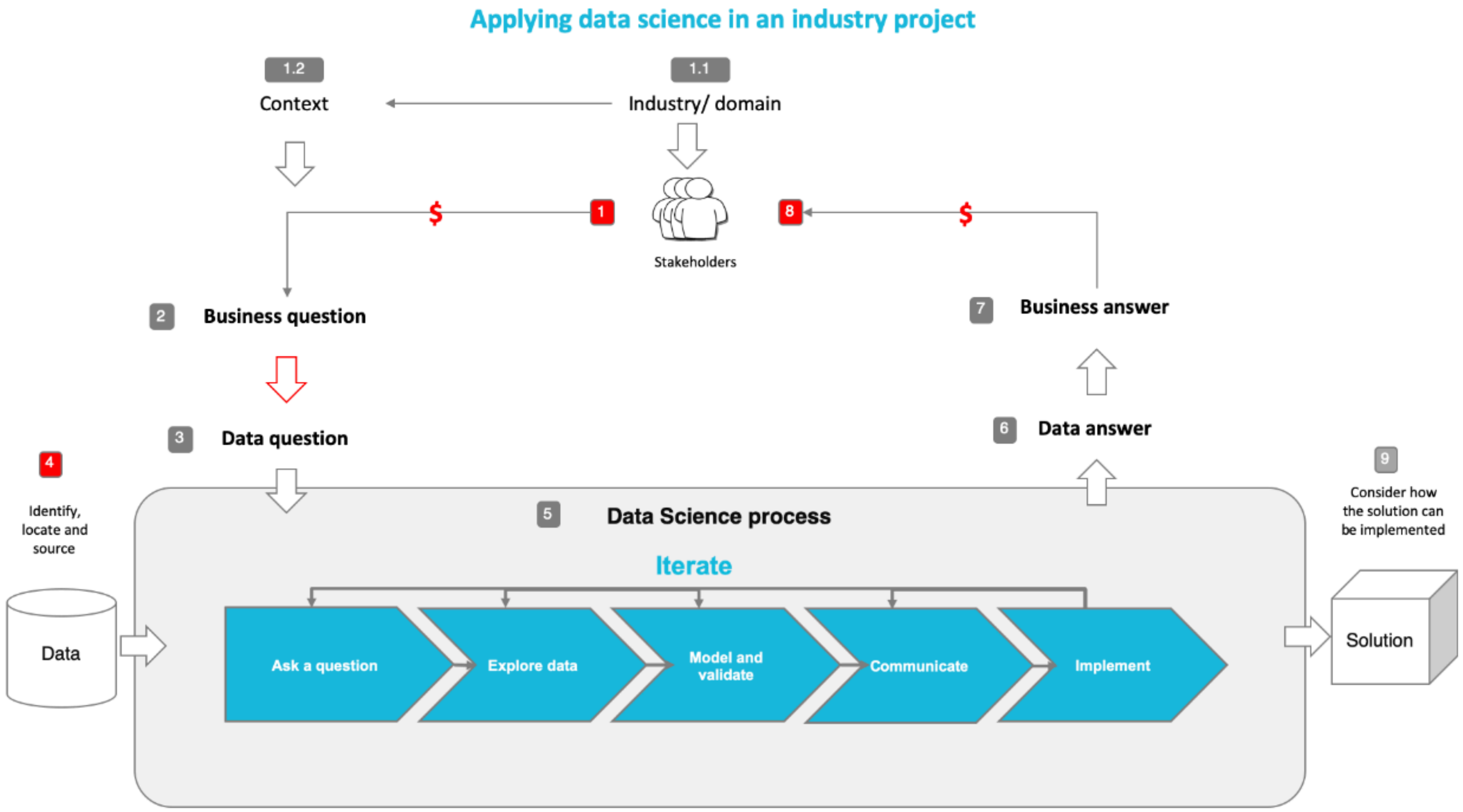

```
[9355 rows x 12 columns]

jobs_in_data types:
work_year              int64
job_title             object
job_category          object
salary_currency       object
salary                 int64
salary_in_usd          int64
employee_residence    object
experience_level      object
employment_type       object
work_setting          object
company_location      object
company_size          object
dtype: object
```

Correlation(features/target variable)


Features	salary_in_usd
salary_in_usd	1.000000
job_category_BI and Visualization	-0.044787
job_category_Cloud and Database	0.001721
job_category_Data Analysis	-0.284148
job_category_Data Architecture and Modeling	0.015233
job_category_Data Engineering	-0.036645
job_category_Data Management and Strategy	-0.060478
job_category_Data Quality and Operations	-0.060160
job_category_Data Science and Research	0.146882
job_category_Leadership and Management	-0.018201
job_category_Machine Learning and AI	0.192326


Features	salary_in_usd
experience_level_Entry-level	-0.231340
experience_level_Executive	0.109093
experience_level_Mid-level	-0.259234
experience_level_Senior	0.303894
work_setting_Hybrid	-0.140286
work_setting_In-person	0.103978
work_setting_Remote	-0.063933
employment_type_Contract	-0.023701
employment_type_Freelance	-0.052373
employment_type_Full-time	0.075271
employment_type_Part-time	-0.058636
company_size_L	-0.042942
company_size_M	0.093605
company_size_S	-0.124171

Process Flow


Applying data science in an industry project

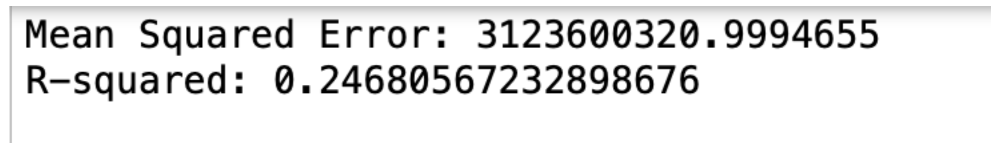
Deliver

- ## Feature engineering

  - The most important features that I will focus my attention on will be work setting, employment type, company size, experience level. These seem to have the highest correlation to Job salary in USD. This is significant to the business because it allows for targeted allocation of resources and strategic decision-making. By focusing on the features which have the highest correlation to job salary in USD, the business can better understand the factors driving compensation within the organization. This insight enables effective salary structuring, recruitment strategies, and talent management practices, ultimately leading to optimized workforce planning, enhanced employee satisfaction, and improved organizational performance.

- ## Machine Models :

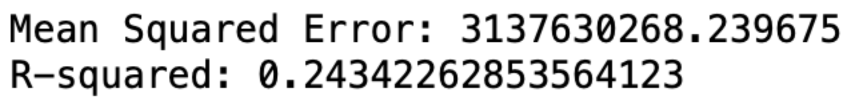
  - Linear Regression



```
Mean Squared Error: 3123600320.9994655
R-squared: 0.24680567232898676
```

Reasoning: Linear regression is best suited for situations where the relationship between the independent and dependent variables is assumed to be linear.

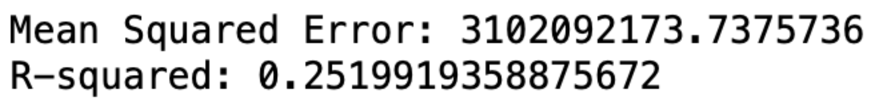
  - Random Forest Regressior



```
Mean Squared Error: 3137630268.239675
R-squared: 0.24342262853564123
```

Reasoning: Random forest regression is effective when dealing with complex nonlinear relationships between the features and the target variable.

  - Gradient Boosting Regression-



```
Mean Squared Error: 3102092173.7375736
R-squared: 0.2519919358875672
```

Reasoning: boosting regression is powerful for capturing complex relationships in data and often yields highly accurate predictions. It excels in situations where other models may struggle to capture subtle patterns or where there are interactions between features.

Overall: The scores for the machine models still need ample improvement. However, Gradient Boosting Regression produced the best results.

Summary, conclusion, and next steps

- In summary, the correlation between the features & target variables weren’t as strong as we hoped. However, we can still use the data found throughout the EDA process to help make decisions on employee compensation packages and with improvements to the gradient boosting model we could implement & distribute this model throughout our staffing agency.Additional variables in the data to be collected in the future may be helpful with coming up with a more predictive model.

Training
Experience
Skill
INTERVIEW
Talent
TOGETHER
RECRUITMENT
QUALIFICATION
MISSION
APPLICATION
30% Benefits
10% Bonus
Candidates

QUESTIONS?

Sources

- **https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data**