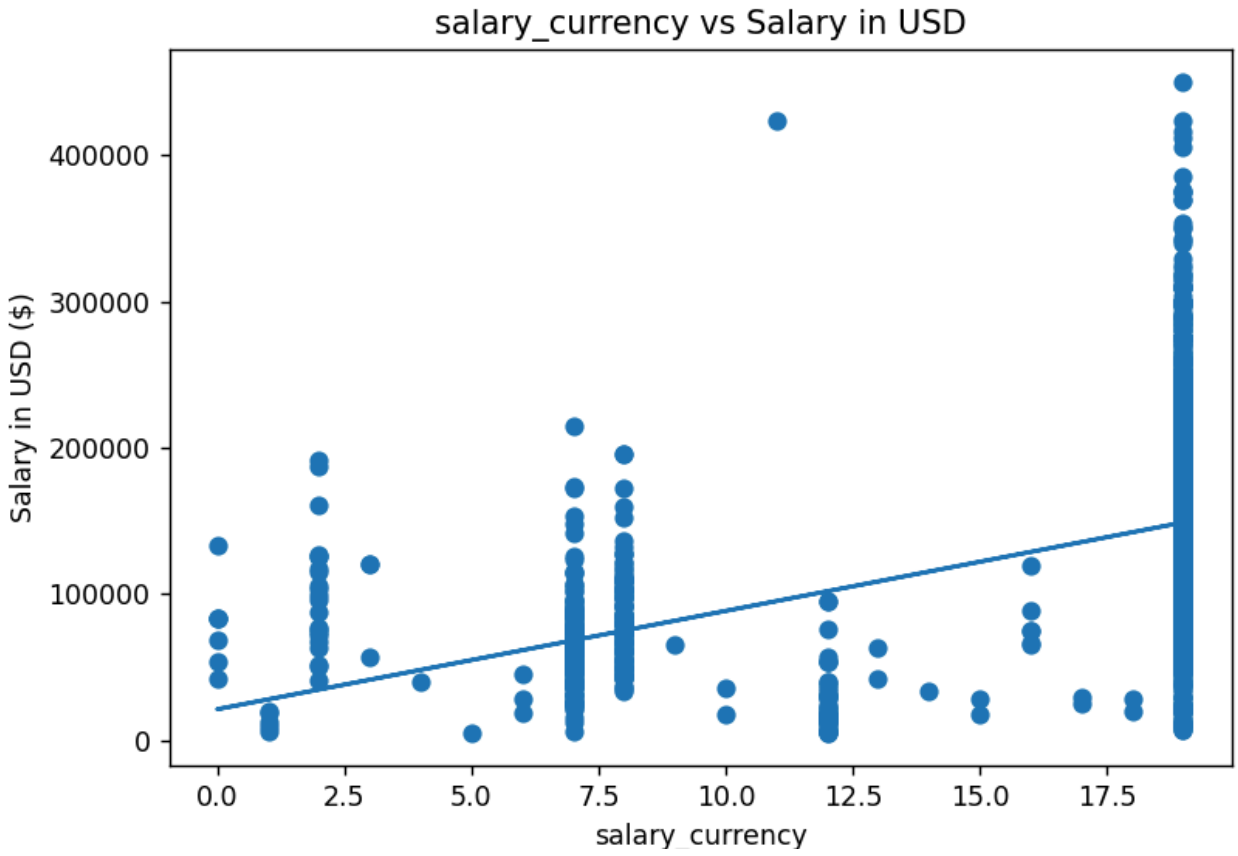
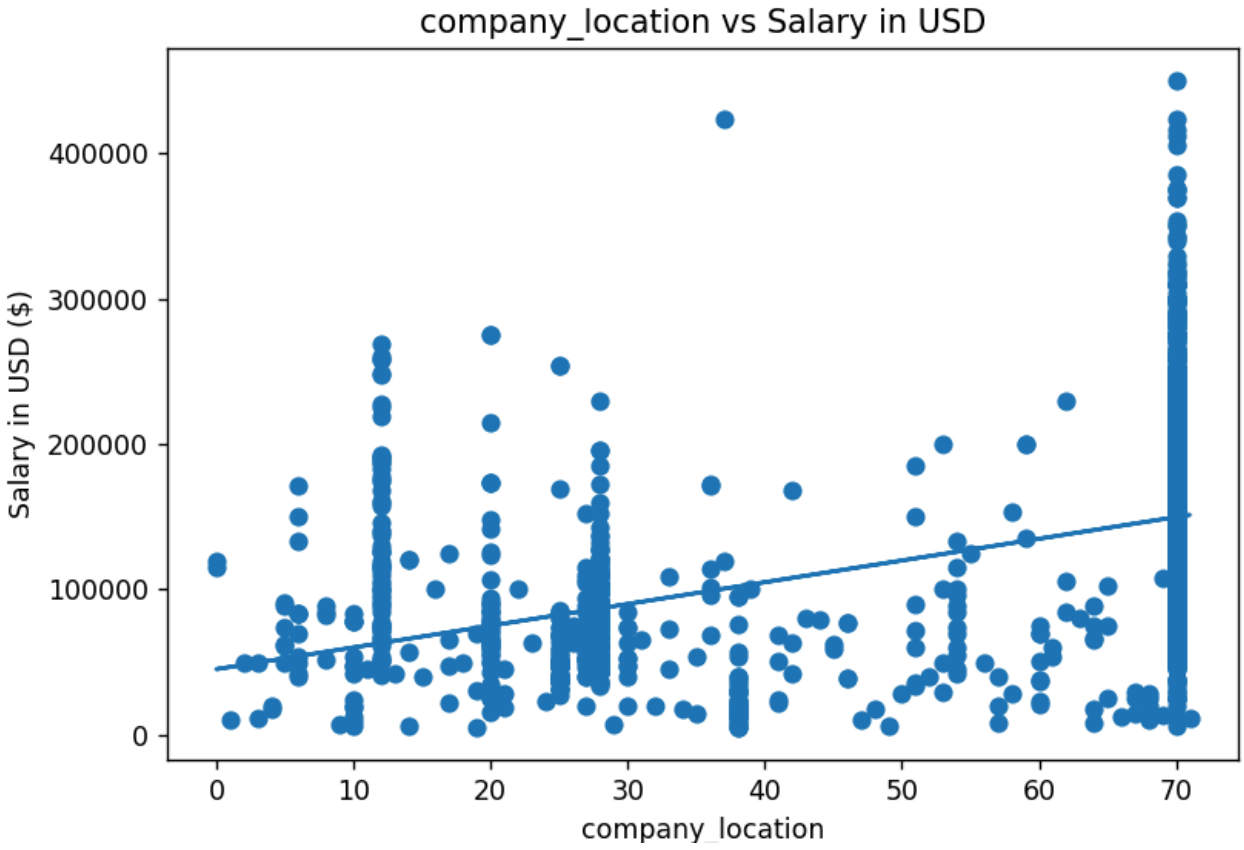
Team Name: Algo Warriors
Members: Rahul Basak, Aden Zhao, Arnav Kaul
PM: Samara Silverman
PM Meeting insights: She helped us resolve some Jupyter Notebook dependency errors.

Challenges that arose: Web scraping from wikipedia took a lot of trial and error to make sure we were filtering the results from beautiful soup correctly.
What went well: Our web scraper works very well; it's a pretty simple script, but it's really cool to see it working without errors.

Plots and tables
Scatterplots and linear regression predictions (combined) of top 3 features with highest correlation to the target of salary in USD:




employee_residence vs Salary in USD
Salary in USD ($)
0
100000
200000
300000
400000
0
10
20
30
40
50
60
70
80
employee_residence


company_location vs Salary in USD
Salary in USD ($)
0
100000
200000
300000
400000
0
10
20
30
40
50
60
70
company_location

Feature	Train Error	Validation Error	Min Feature Test Error
salary_currency	3299058810.567853	2961873372.983696	3035563968.788183
employee_residence	3331902323.646222	2979869257.972082	
company_location	3375656800.578246	3013291181.9098163	

What did you learn from linear regression? We learned that linear regression models with categorical features are very dependent on how the features are encoded and dispersed.
What did you learn about your dataset? Since all of our highly-correlated features were initially categorical, the different encodings we used (one-hot, label, etc.) did not provide reasonable models, giving such large dataset errors.
What is the question you are interested in exploring? What do you need to do to start answering it? Since linear regression models did not work well on these features, we are interested in exploring how other models can be used to help predict salary. To do so, we will potentially look into classification models, or possibly use features that translate well to a linear basis.

Signatures: Arnav Kaul, Aden Zhao,