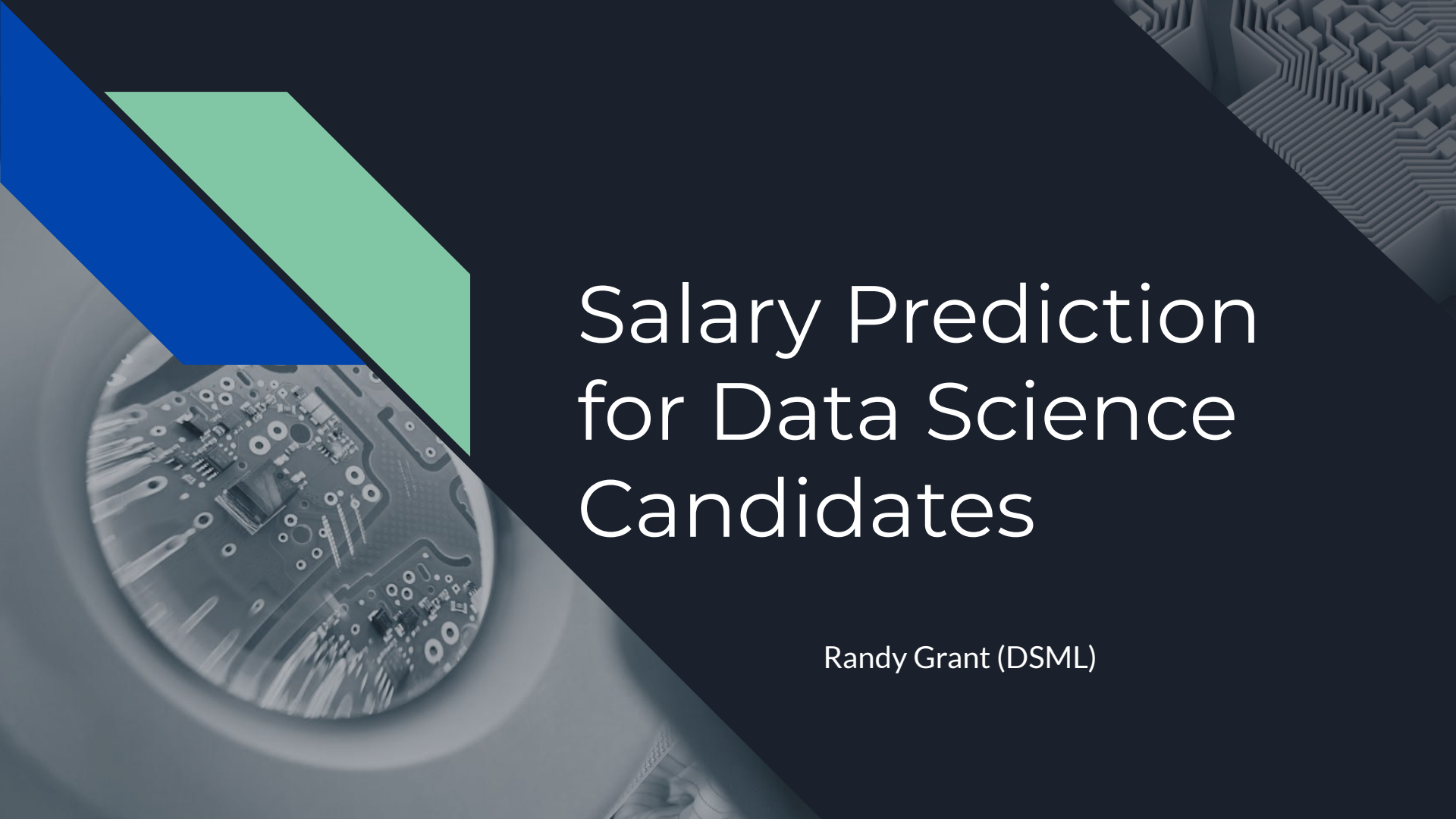


Salary Prediction for Data Science Candidates

Randy Grant (DSML)

Introduction

- Issue: The client, a recruiting agency, needs a way to predict how much a data scientist will get paid so they can recruit more effectively

- Solution: With data created from web scraping indeed.com, predicting salary should be able to be to be done based on a scoring of “work happiness”

- Goal: Demonstrate this concept to prove to the client that they can use this data as an effective tool to recruit

Methods

Data Used

- Indeed.com
- Tools Used:
  - Pandas for data exploration
  - Scikit-learn for models
  - Seaborn for plots
- Metrics Created:
  - Spread of Salary Data Across 1,000+ jobs
  - Error Metrics of Trained Models
  - Scatter Plots of Data Fit

Sample Data

Raw:

	job_title	location	salary	company	company_rating	job_locations_available
7678	newRemote - Staff /Lead Data Scientist	Remote in Frisco, TX 75033 75033+6 locations	150000	Shopify	4.100000	+2 locations
8280	Director, Data Science	Remote	data_not_available	Slack	3.400000	+1 location
9199	Principal Data Scientist	Remote in Colorado+43 locations	data_not_available	Verizon	3.700000	+13 locations
9895	Junior Data Scientist	Remote in New York, NY	data_not_available	SmartAsset	4.200000	+3 locations
9401	newData Scientist	United States	data_not_available	Fund For Public Health In New York Inc	4.200000	+1 location

```
Int64Index: 11465 entries, 0 to 11662
Data columns (total 6 columns):
```

Final Snippet:

	salary	company_rating	job_locations_available	happiness	flexibility	learning	achievement	appreciation	inclusion	support	purpose	energy
0	175000.0	3.2	1.0	65.0	77.0	71.0	70.0	70.0	69.0	69.0	67.0	64.0
1	90000.0	3.8	9.0	55.0	55.0	60.0	62.0	61.0	55.0	59.0	59.0	56.0
2	124800.0	3.9	4.0	65.0	72.0	68.0	70.0	75.0	73.0	73.0	70.0	67.0
3	90000.0	4.1	3.0	76.0	84.0	80.0	78.0	82.0	79.0	80.0	78.0	74.0
4	155000.0	3.7	17.0	67.0	69.0	68.0	72.0	71.0	68.0	70.0	70.0	68.0

```
RangeIndex: 1001 entries, 0 to 1000
Data columns (total 25 columns):
```

Tool Details

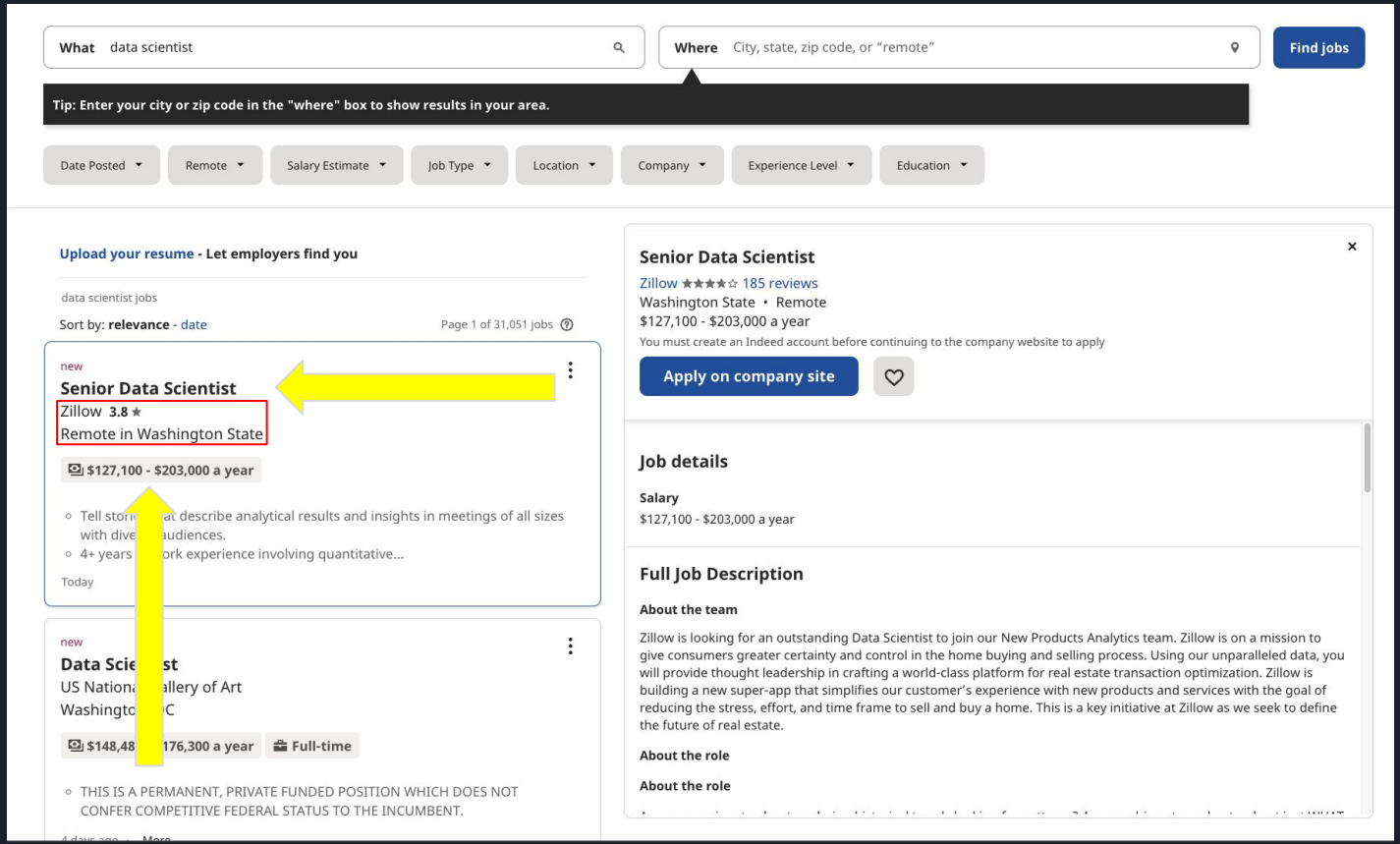
My IDE that contains all of the rest of the awesomeness

I used this to bring in the data, filter it down, group it, count it, all to make it pretty

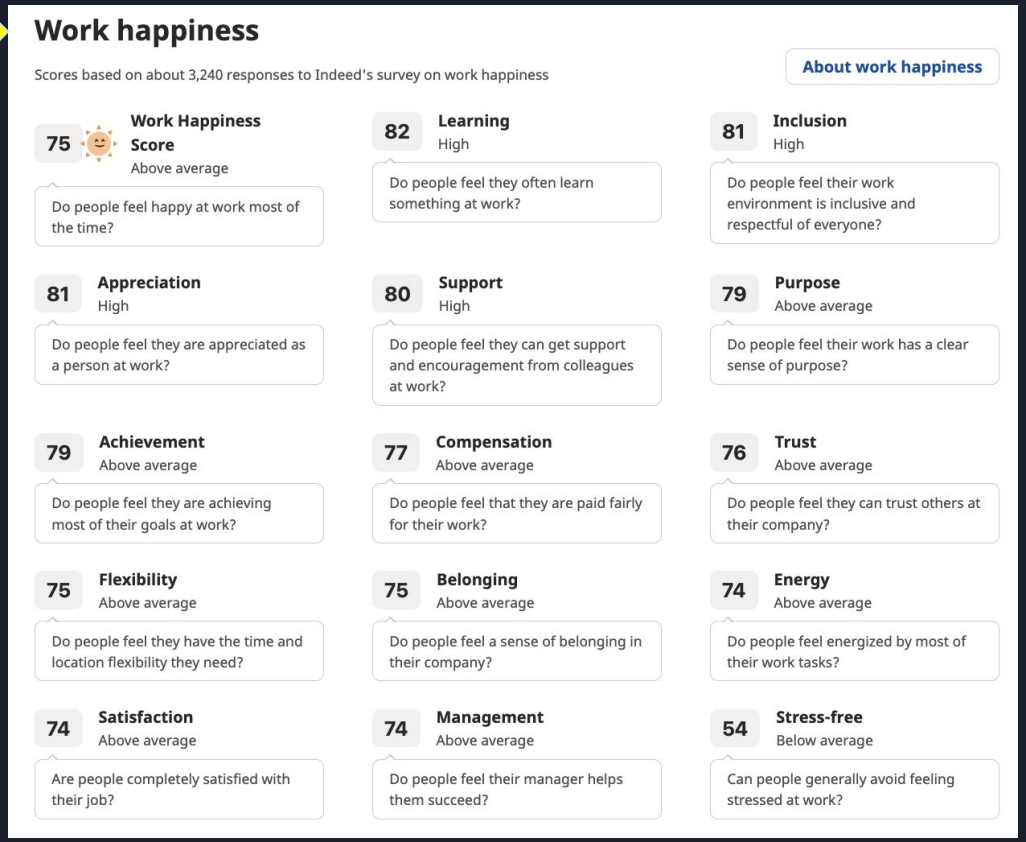
This was used to graph out all of the daily entries for a 2 year period

Linear models were created and graphed out using this amazing library

What is the Data



What is the Data



Work happiness

Scores based on about 3,240 responses to Indeed's survey on work happiness

About work happiness

75 **Work Happiness Score**
Above average
Do people feel happy at work most of the time?

82 **Learning**
High
Do people feel they often learn something at work?

81 **Inclusion**
High
Do people feel their work environment is inclusive and respectful of everyone?

81 **Appreciation**
High
Do people feel they are appreciated as a person at work?

80 **Support**
High
Do people feel they can get support and encouragement from colleagues at work?

79 **Purpose**
Above average
Do people feel their work has a clear sense of purpose?

79 **Achievement**
Above average
Do people feel they are achieving most of their goals at work?

77 **Compensation**
Above average
Do people feel that they are paid fairly for their work?

76 **Trust**
Above average
Do people feel they can trust others at their company?

75 **Flexibility**
Above average
Do people feel they have the time and location flexibility they need?

75 **Belonging**
Above average
Do people feel a sense of belonging in their company?

74 **Energy**
Above average
Do people feel energized by most of their work tasks?

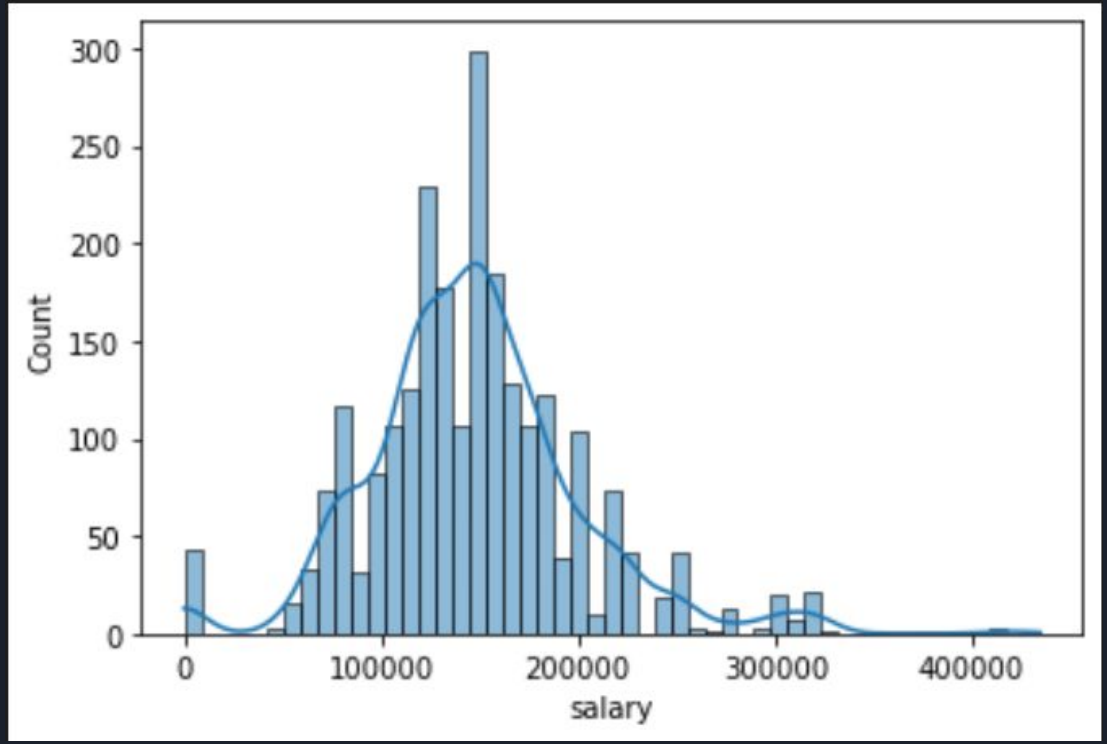
74 **Satisfaction**
Above average
Are people completely satisfied with their job?

74 **Management**
Above average
Do people feel their manager helps them succeed?

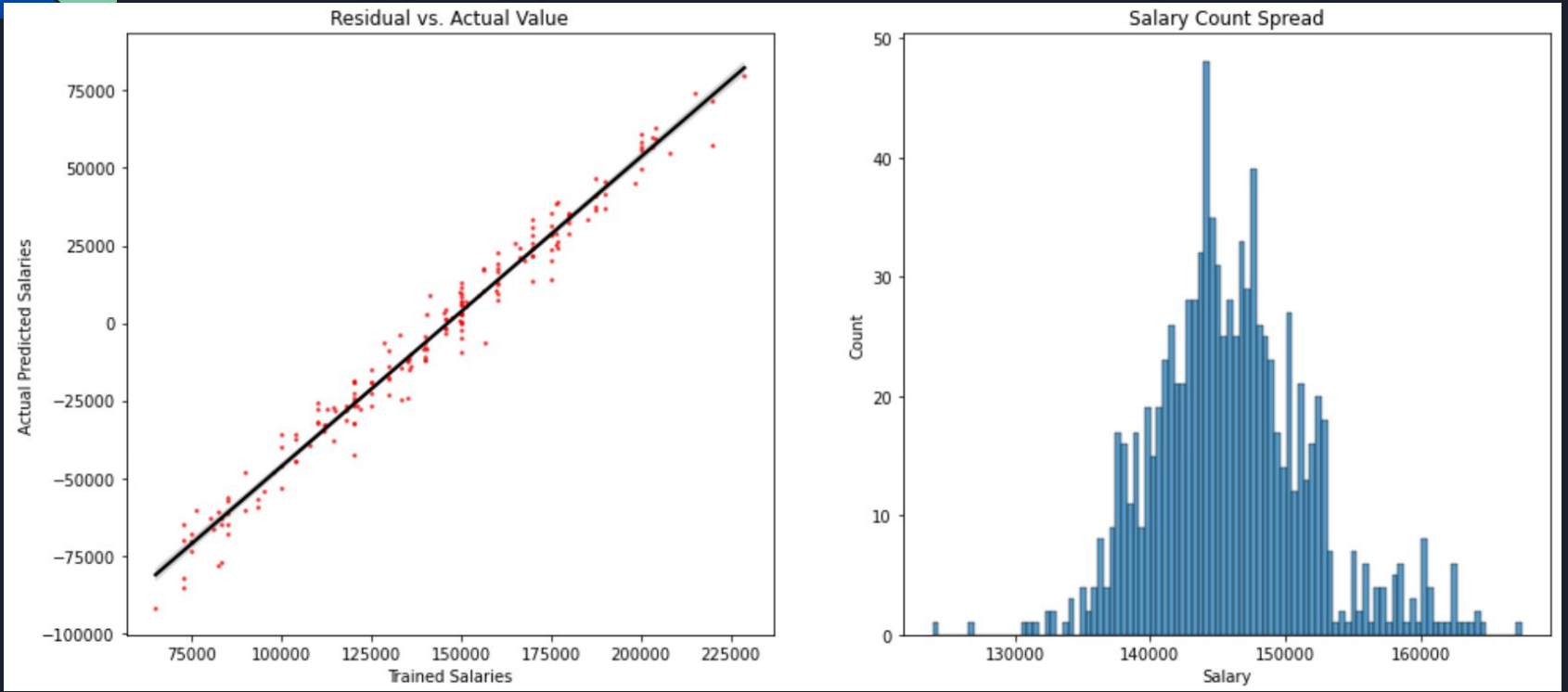
54 **Stress-free**
Below average
Can people generally avoid feeling stressed at work?

Metrics!!!

- Salary ranged from 0 to more than 400k per year
- These were outliers and were removed
- This would throw off the predictions if left in



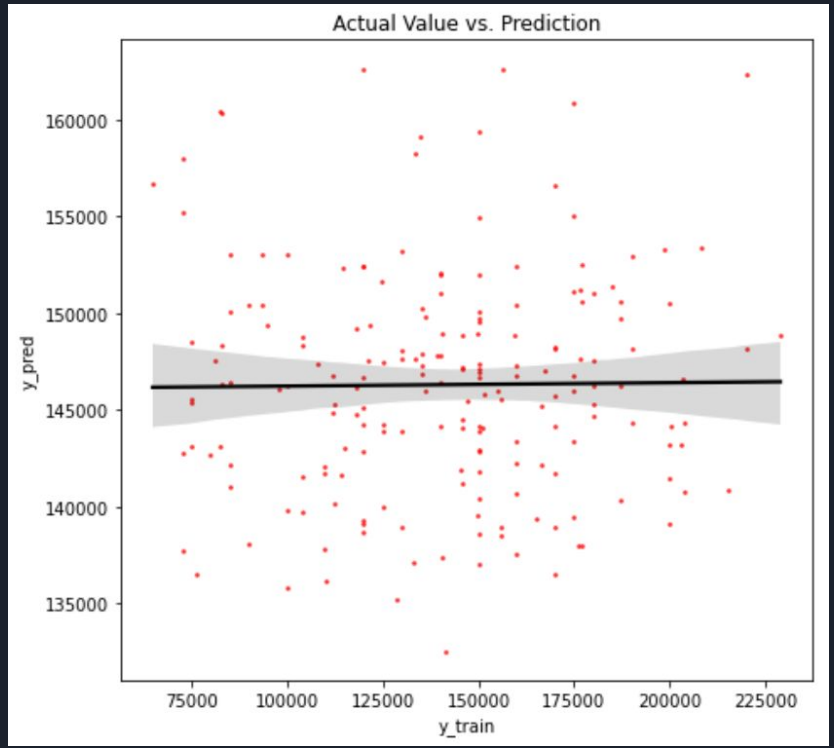
Metrics!!!


Residual vs. Actual Value
Actual Predicted Salaries
Trained Salaries
75000
50000
25000
0
−25000
−50000
−75000
−100000
75000
100000
125000
150000
175000
200000
225000
Salary Count Spread
Count
Salary
50
40
30
20
10
0
130000
140000
150000
160000

Metrics again!!!

```
R² training score: 0.0241370488219681
R² test score: -0.043735956434078016
RMSE on training data: 139102.2838013116
RMSE on test data: 147436.2139067333
----------------
Intercept: 147476.39950021522
----------------
```

It appears most data that I thought was going to be viable wasn't the best decision. The multicollinearity between all features created a very low R^2 value for all linear models. The RMSE scores were also high. However, it's good to note that even though the features were highly correlated for this project, one could still create models in a similar way for a prediction with features with less correlation.



Metrics again!!!

This is the Ridge coefficients.

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.

job_title_mathematician	16080.738437
region_northeast	7516.322298
job_title_statistician	7393.872452
region_west	7232.830250
region_remote	5991.498914
job_title_machine learning engineer	4417.825738
region_south	1630.261903
flexibility	916.961818
compensation	193.285864
job_locations_available	28.637847
purpose	-197.208840
stress-free	-256.032581
achievement	-296.403414
management	-511.095248
company_rating	-947.925923

Appendix

https://www.indeed.com/

https://towardsdatascience.com/statistics-in-python-collinearity-and-multicollinearity-4cc4dcd82b3f

https://www.youtube.com/watch?v=Q_nKKx8L_qE

https://gist.github.com/rogerallen/1583593

https://datascienceparichay.com/article/pandas-split-column-by-delimiter/

https://stackoverflow.com/questions/39903090/efficiently-replace-values-from-a-column-to-another-column-pandas-dataframe

https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv

Appendix

https://www.csestack.org/compare-two-lists-python-return-non-match-elements/

https://stackoverflow.com/questions/312443/how-do-you-split-a-list-into-evenly-sized-chunks

https://stackoverflow.com/questions/13384841/swap-values-in-a-tuple-list-inside-a-list-in-python

https://www.kaggle.com/code/jack89roberts/top-7-using-elasticnet-with-interactions/notebook

https://en.wikipedia.org/wiki/Ridge_regression#:~:text=Ridge%20regression%20is%20a%20method,econometrics%2C%20chemistry%2C%20and%20engineering

365 Data Science - Linear Regression Practical Example (Part 5)

Thank you!

Questions?