

What are your salary expectations,
after finishing the bootcamp?



Salary Prediction Tool for U.S.-Based Data Science Roles

Project Overview

Team

- Cristian Llanes (Square Role)
- Maria Sevillano (Triangle Role)
- Alejandra Villarreal (Circle Role)
- Sharof Abdoolayev (X Role)



Project Overview

Objective

The purpose of this project is to build a resource for job-seekers to predict the salary of a given career field, Data Science, based on set variables.

- Answer the "**What Are Your Salary Expectations?**" question that a hiring manager might pose during an interview process.
- Determine if they should accept or decline a job offer.



Project Overview

Data Source

Original data sets:

Levels_Fyi_Salary_Data.csv

Participants_Data.csv



Project Overview

Levels_Fyi_Salary_Data.csv

Levels_Fyi_Salary_Data																		
AutoSave OFF																		
Home Insert Draw Page Layout Formulas Data Review View Developer Tell me																		
Calibri (Body) 12 A^ A^ Paste B I U \$ % , < 0 .00 → 0 Conditional Formatting Insert Delete Sort & Filter Find & Select Analyze Data																		
General																		
Share Comments																		
fx responseid																		
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q		
responseid	timestamp	company	level	title	totalyearlycc	location	yearsofexper	yearsatscomp	tag	basalary	stockgrantva	bonus	otherdetails	cityid	dmaird	rowNumber		
A1	6/7/17 11:33	Oracle	L3	Product Manager	127000	Redwood City, CA	1.5	1.5	NA	107000	20000	10000	NA	7392	807	1		
A2	6/10/17 17:11	eBay	SE 2	Software Engineer	100000	San Francisco, CA	5	3	NA	0	0	0	NA	7419	807	2		
A3	6/11/17 14:53	Amazon	L7	Product Manager	310000	Seattle, WA	8	0	NA	155000	0	0	NA	11527	819	3		
A4	6/17/17 0:23	Apple	M1	Software Engineering	372000	Sunnyvale, CA	7	5	NA	157000	180000	35000	NA	7472	807	7		
A5	6/20/17 10:58	Microsoft	60	Software Engineer	157000	Mountain View, CA	5	3	NA	0	0	0	NA	7322	807	9		
A6	6/21/17 17:27	Microsoft	63	Software Engineer	208000	Seattle, WA	8.5	8.5	NA	0	0	0	NA	11527	819	11		
A7	6/22/17 12:37	Microsoft	65	Software Engineering	300000	Redmond, WA	15	11	NA	180000	65000	55000	NA	11521	819	12		
A8	6/22/17 13:55	Microsoft	62	Software Engineer	156000	Seattle, WA	4	4	NA	135000	8000	13000	NA	11527	819	13		
A9	6/22/17 23:08	Microsoft	59	Software Engineer	120000	Redmond, WA	3	1	NA	0	0	0	NA	11521	819	15		
A10	6/26/17 21:25	Microsoft	63	Software Engineer	201000	Seattle, WA	12	6	NA	157000	26000	28000	NA	11527	819	16		
A11	6/30/17 16:29	Salesforce	9	Software Engineering	450000	San Francisco, CA	16	3	NA	230000	100000	45000	NA	7419	807	18		
A12	7/2/17 14:16	Microsoft	Sde 2	Software Engineer	155000	Bellevue, WA	5	3	NA	126000	0	0	NA	11470	819	19		
A13	7/3/17 19:28	Microsoft	63	Product Manager	150000	Redmond, WA	10	10	NA	0	0	0	NA	11521	819	20		
A14	7/7/17 22:29	Microsoft	63	Software Engineer	191000	Seattle, WA	7	7	NA	152000	17000	22000	NA	11527	819	21		
A15	7/14/17 21:36	Amazon	L6	Software Engineering	287000	Seattle, WA	12	1	NA	160000	0	0	NA	11527	819	23		
A16	7/16/17 16:50	Amazon	L5	Software Engineer	218000	Seattle, WA	10	0	NA	150000	7000	61000	NA	11527	819	25		
A17	7/20/17 22:35	Facebook	E3	Software Engineer	168000	Menlo Park, CA	1	1	NA	0	0	0	NA	7300	807	27		
A18	7/22/17 22:20	Uber	5a	Software Engineer	160000	San Francisco, CA	9	1	NA	0	0	0	NA	7419	807	29		
A19	7/24/17 12:21	Apple	L4	Software Engineer	50000	London, EN, United Kingdom	2	2	NA	0	0	0	NA	12008	0	30		

Participants_Data.csv

20	
21	
23	
25	
27	
29	
30	

Project Overview

Technologies Used

- Pandas
- Postgres
- Amazon AWS
- SciKitLearn
- Tableau



Project Overview

Questions Data Set Will Answer


- Will salary for Data Science jobs continue to experience growth in the future?
- Based on the selected set of variables, what is the expected salary range?
- Determine salary trends based on specific factors.



Preliminary Machine Learning Model

Data Preprocessing

Preprocessing will involve the followings:

- Checking and handling imbalanced datasets.
 - Performing initial exploratory analysis, including scatter plotting and correlation.
 - Removing non-beneficiary columns.
 - Preparing the data by working with any missing values, scaling the data, and converting categorical variables by using the one-hot encoding scheme.
- 

Preliminary Machine Learning Model

Splitting the dataset

The dataset will be split into training and testing sets using the 80/20 Pareto principle resulting in a test size of 20%.



Preliminary Machine Learning Model

Supervised Machine Learning Model

We will use a supervised machine learning model since we are looking to predict a value. There are different models we can use:

- **Regression**
- **Classification / Ensemble Methods**



Preliminary Machine Learning Model

Regression

- Apply a Linear Regression to predict salary.
- We will also explore applying a Multilinear Regression Model to add other factors that might influence the salary prediction.



Preliminary Machine Learning Model

Classification / Ensemble Methods

We could use Random Forest Regression to discover the connection between the target and independent variables to determine a continuous value. This connection can then be used to predict salaries of data science jobs..



Preliminary Machine Learning Model

Model Evaluation

We will evaluate the models based on:

- **Explained Variance Score:** Similar to the R^2 score, with the notable difference that it does not account for systematic offsets in the prediction.
- **Model Score:** Returns the mean accuracy on the given test data.

