stack overflow

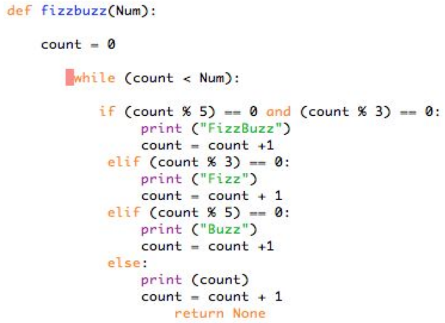
2019 Developer Survey

A Linear (Regression) Story

Jude Buenaseda

Do different backgrounds of developers make an impact on their salary?

+

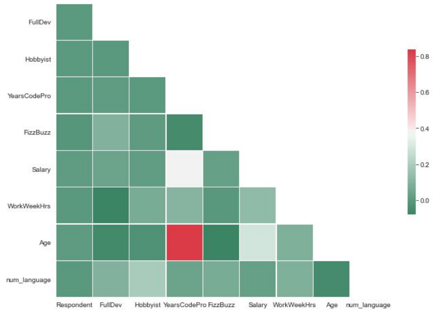


```
def fizzbuzz(Num):

    count = 0

       while (count < Num):

           if (count % 5) == 0 and (count % 3) == 0:
                print ("FizzBuzz")
                count = count +1
            elif (count % 3) == 0:
                print ("Fizz")
                count = count + 1
            elif (count % 5) == 0:
                print ("Buzz")
                count = count +1
            else:
                print (count)
                count = count + 1
                    return None
```

=

Using linear regression, we can finally find out if that fizzbuzz code was worth spending all night memorizing before an interview.

Process



AMERICA'S NEXT top model

1

Clean & Filter

Filter data by Country (US) and Employment (full-time)

2

Choose Features

Include features that would likely impact income like years coding and education level

3

Feature Engineer

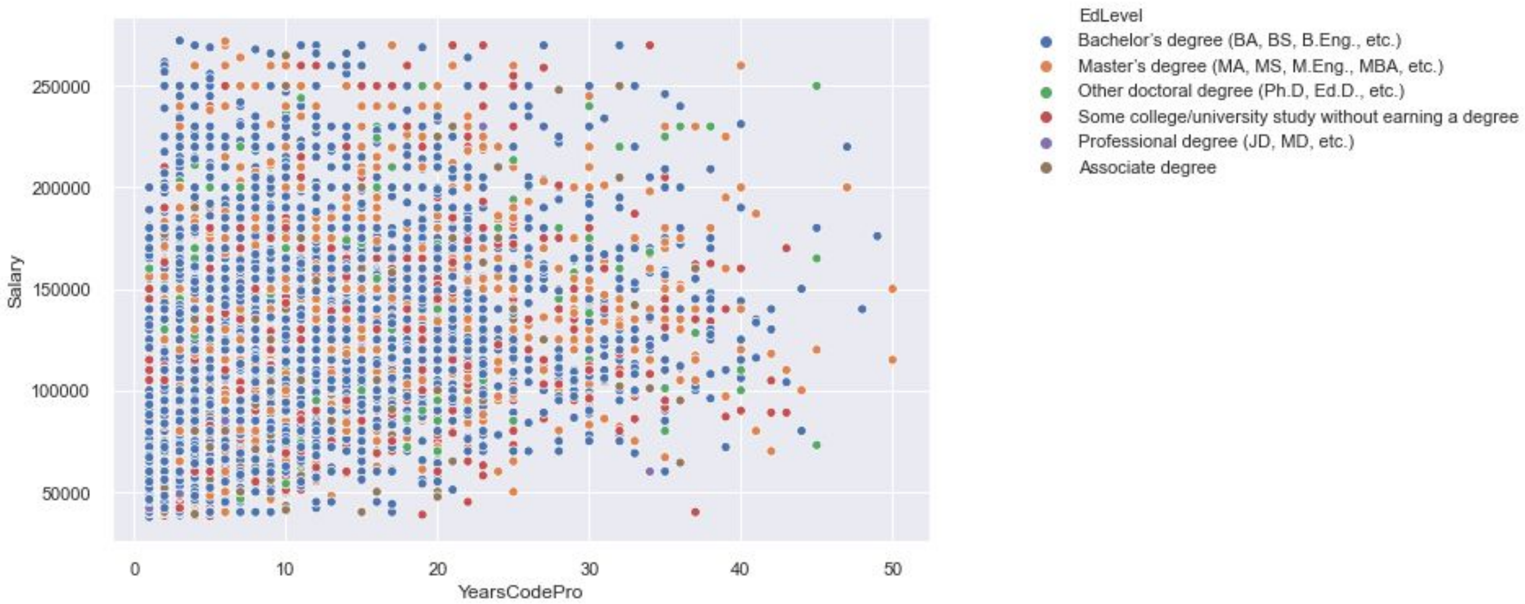
Applying log transformations and creating dummy variables when necessary

4

Fit Model

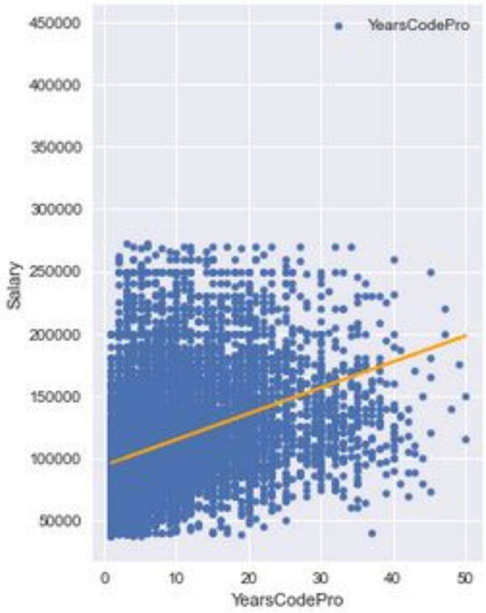
Test multiple linear regression models to best explain the relationship between the dependent and independent variables

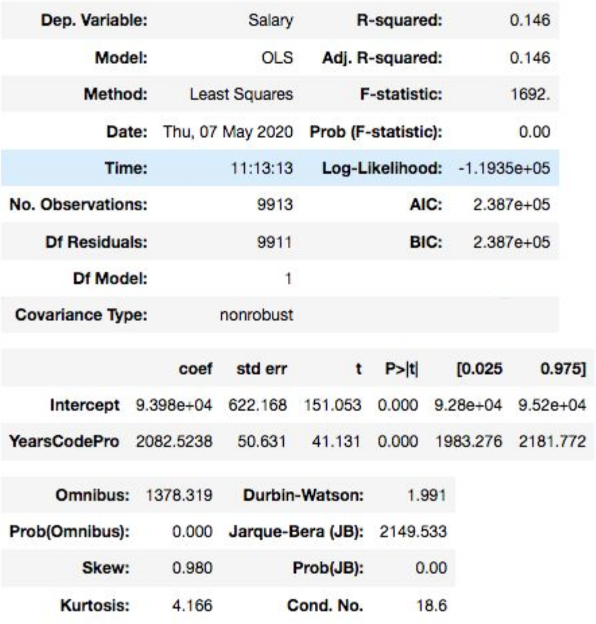
Education Level - Does it matter?



The salary means for each level is different but with a p-value of 1.46 using ANOVA, it isn't statistically significant.

Baseline Model - Simple Linear Regression





Dep. Variable:	Salary	**R-squared:**	0.146
Model:	OLS	**Adj. R-squared:**	0.146
Method:	Least Squares	**F-statistic:**	1692.
Date:	Thu, 07 May 2020	**Prob (F-statistic):**	0.00
Time:	11:13:13	**Log-Likelihood:**	-1.1935e+05
No. Observations:	9913	**AIC:**	2.387e+05
Df Residuals:	9911	**BIC:**	2.387e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.398e+04	622.168	151.053	0.000	9.28e+04	9.52e+04
YearsCodePro	2082.5238	50.631	41.131	0.000	1983.276	2181.772

Omnibus:	1378.319	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2149.533
Skew:	0.980	**Prob(JB):**	0.00
Kurtosis:	4.166	**Cond. No.**	18.6

This baseline model consists of only YearsCodePro as the feature.

Adjusted R^2: ***.146***
Standard Error: ***40k***

Multivariable Linear Regression Models

19.1%

***Adj R^2* - All features (26)**

Using all features picked from the dataset & dummy variables

- An adjusted r-square of 19.1 isn't much of an improvement from the 14.6 of the single linear regression model.
- Standard Error: 39k

23.7%

***Adj R^2* - All features + interactions (29)**

Performing log transformations and adding the top 3 interactions based on improvement of adjusted r-squared

- An adjusted r-square of 23.7 is a major improvement from previous model.
- Standard Error: 39K

23.1%

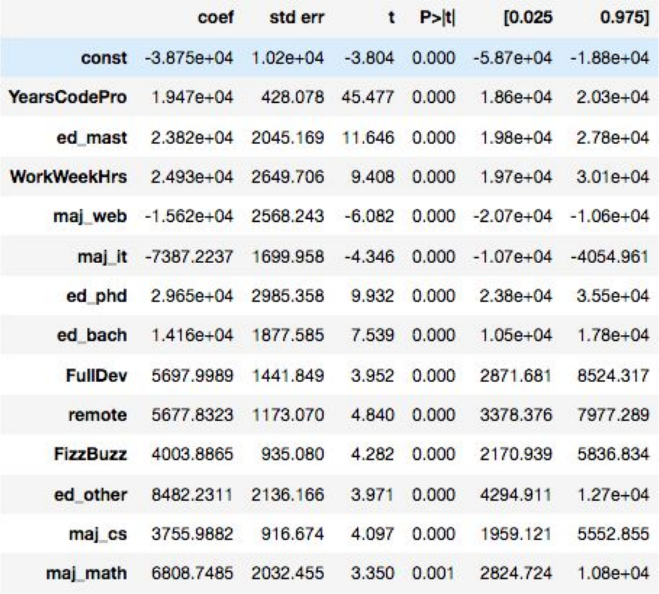
***Adj R^2* - Features using stepwise selection (13)**

Including only features that have the lowest p-value and contributing to the adjusted r-squared using a stepwise selection function

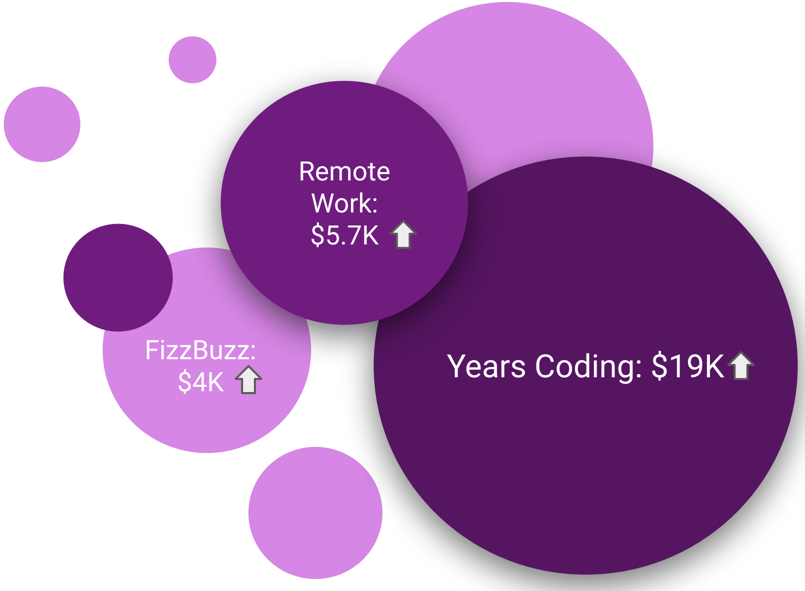
- Similar percentage to second model but with less features
- Standard Error: 38k

Coefficients

Features used in final model:



	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.875e+04	1.02e+04	-3.804	0.000	-5.87e+04	-1.88e+04
YearsCodePro	1.947e+04	428.078	45.477	0.000	1.86e+04	2.03e+04
ed_mast	2.382e+04	2045.169	11.646	0.000	1.98e+04	2.78e+04
WorkWeekHrs	2.493e+04	2649.706	9.408	0.000	1.97e+04	3.01e+04
maj_web	-1.562e+04	2568.243	-6.082	0.000	-2.07e+04	-1.06e+04
maj_it	-7387.2237	1699.958	-4.346	0.000	-1.07e+04	-4054.961
ed_phd	2.965e+04	2985.358	9.932	0.000	2.38e+04	3.55e+04
ed_bach	1.416e+04	1877.585	7.539	0.000	1.05e+04	1.78e+04
FullDev	5697.9989	1441.849	3.952	0.000	2871.681	8524.317
remote	5677.8323	1173.070	4.840	0.000	3378.376	7977.289
FizzBuzz	4003.8865	935.080	4.282	0.000	2170.939	5836.834
ed_other	8482.2311	2136.166	3.971	0.000	4294.911	1.27e+04
maj_cs	3755.9882	916.674	4.097	0.000	1959.121	5552.855
maj_math	6808.7485	2032.455	3.350	0.001	2824.724	1.08e+04



Seniority, working remotely and whether an interviewer asks you about fizzbuzz ***MIGHT*** earn you more money on average.

Conclusion

We're still in the dark...