Fall 2021 DS GA 1001 Capstone Project:
Analyzing Engineering Salary Data

The Money Team

By: Giulio Duregon (gjd9961), Joby George (jg6615), Howell Lu (hl4631), Jonah Poczobutt (jp6422), Alexandre Vives (av2926)

Professors Wallisch and Bradonjic
Due 12/22/2021

1.0 Introduction:

As a team of graduate students at New York University's Center for Data Science we were naturally curious about the potential salaries of data scientists and engineers in different companies and industries. When looking for different datasets that would be suitable for the Capstone Project, our team found two datasets that seemed well suited to model this problem:

1. A Kaggle Dataset titled *"Data Science and STEM Salaries"* [1]
2. A Dataset providing general information on Fortune 1000 companies [2]

1.1 Dataset Descriptions:

The primary dataset that we used to model salaries is *Data Science and STEM Salaries*. It is a scraped compilation of 62,000 records from the anonymous forum, *levels.fyi*, where employees post which company they work for, their specific job role (i.e. software engineer, data scientist, etc.), their job level (L7, Principal data scientist) and more. For a more comprehensive description of the features and values, see *Figure* 1, **Data Description Table** in the Appendix.

The auxiliary dataset used in our analysis was used to identify a company's sector (i.e. technology, finance, healthcare, etc.), so that the team could learn more about the effect of industry on salary.

1.2 Data Preparation and Cleaning:

As our first step to understand the data, the team wanted to know which features were sparsely filled versus being present for the majority of our rows. We visualized the feature presence using missingno plots[3], which can be seen in *Figure 2,* **Feature Misingno Plots**.

Variables that had large amounts of missing data were Race, Education, Gender, and OtherDetails. Aside from OtherDetails, the team decided to keep these features given the potential explanatory power in predicting salary.

From here, we wanted to determine which features were noisy and would not be valuable in modeling. Based on the large number of potential values, the variables, *level*, *cityid* were removed. Furthermore, the variables *base salary, stock grant value,* and *bonus* were also excluded, as this would give our model compensation data as an input, when the goal of the model is to predict total yearly compensation. Lastly, to keep the scope relevant to the United States, the team discarded any respondents that were not located in the United States.

---

[1] https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries
[2] https://www.kaggle.com/winston56/fortune-500-data-2021
[3] https://github.com/ResidentMario/missingno

This left us with an analytical dataset containing the following features:

1. Company
2. Title
3. Region
4. Sector
5. Race
6. Education
7. Gender
8. Years of experience
9. Years at company
10. Tag
11. Year of survey response
12. Total Yearly Compensation (the variable we will be predicting)

For brevity's sake, we will refer to the Total Yearly Compensation variable as **TYC.**

2.0 Exploratory Data Analysis and Hypothesis Testing

We set out to both learn more about the dataset, and investigate some pieces of career advice that many of us had heard from alumni, mentors and peers through hypothesis testing.

Many of the below hypothesis tests assess whether we can say there is a significant difference in pay between groups based on some other feature. One feature in our data set that is, unsurprisingly, highly correlated with total pay is years of experience.

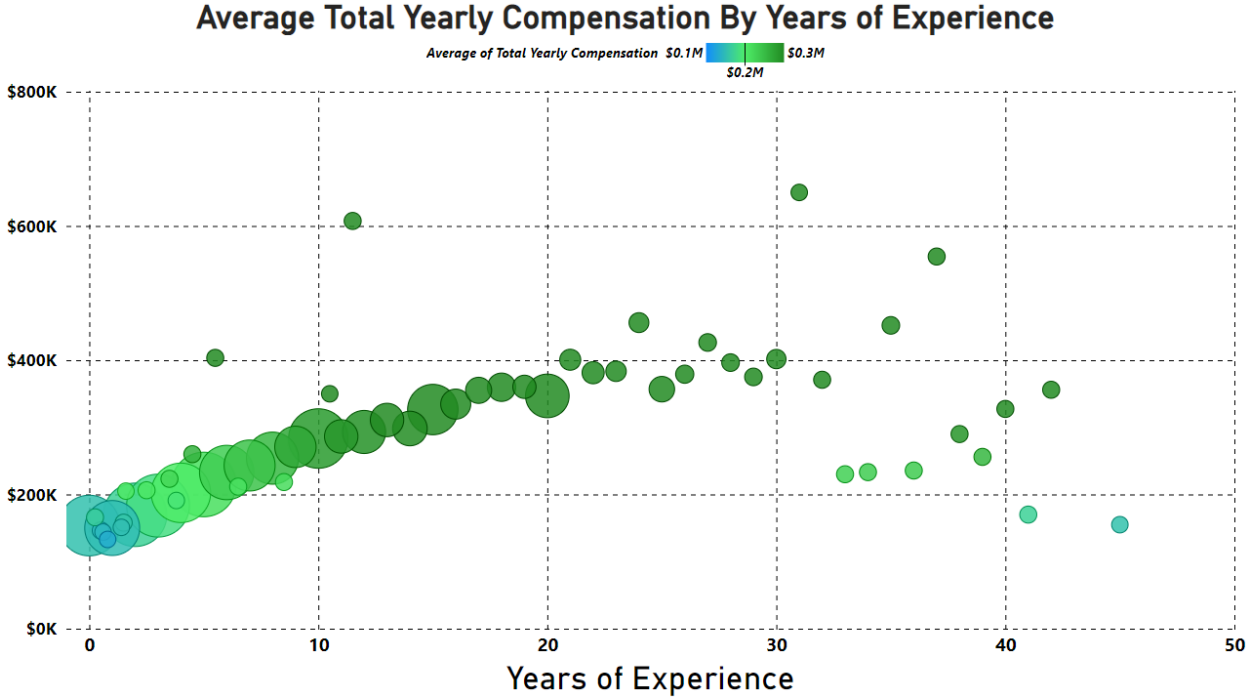


Figure 3: Average TYC vs Y.O.E.

This presented a potential pitfall for our analysis. For example, The share of respondents with 15+ years of experience that are male is higher than the shares of respondents with less than 5 years of experience that are male. In order to better assess the relationship between gender and pay, therefore, we should control for years of experience. For all of our tests, we attempted to control for years of experience by bucketing our data based on the experience feature (0-5years, 10-15 years, 15+ years, for example.) While not perfect, bucketing in this manner allows us to evaluate the relationship between our feature of interest and total pay for each bucket, thereby reducing the potential effect of the correlation between the feature of interest and years of experience, while also maintaining enough sample data for each test to make a meaningful assessment of our hypothesis. This is what is meant in the following test descriptions when we say we have controlled for experience.

2.1 Years at Company & TYC

Members of our group had previously heard the advice that leaving a company after a couple of years and accepting a new role at a new company was a way to quickly drive up yearly compensation. The intuition behind this colloquial advice is that if your market value increases faster than annual raises and promotion increases, you won't be able to realize it unless you look for a different employer that will be more incentivized to increase salary for your skills.

We controlled for total experience, and then separated data points into two groups, one with years at the current company greater than the median for data points in that bucket, and one with fewer years at the current company. We then ran a Mann-Whitney U test on these groups within each total experience bucket to compare median compensation.

H_0: $\boldsymbol{Mdn_f = Mdn_m}$ The salaries of those who have above median years at their current company do not differ from those who have below median years at their current company, controlling for experience.

$\boldsymbol{H_a}$**:** $\boldsymbol{Mdn} \; ! = \boldsymbol{Mdn_m}$ The salaries of those who have above median years at their current company differ from those who have below median years at their current company, controlling for experience.

Median salaries for low years vs high years at firm and P-Values

	Low Years of Exp	High Years of Exp	P-Value
0-5 Years	167000	172000	$2.4e^{-13}$
5-10 Years	225000	210000	$2.6e^{-24}$
10-15 Years	270000	250000	$2.93e^{-19}$
15+	306000	280000	$4.09e^{-21}$

Takeaways: Our results seem to indicate that those who stay with their companies have better pay early in their careers, but this relationship flips at higher levels of experience, consistent with the advice we had heard. However, we believe this result is due more to the nature of our bucketing function than an actual meaningful relationship for early career respondents. In the low experience bucket, respondents who have been with their current companies for a longer period of time are more likely to be above the bucket average for total years of experience in general. This effect dissipates over time, as people move between employers. Because of this, we are not really making a fair comparison between these two groups, as our attempt to control for experience in the 0-5 year bucket is being undone due to this relationship. As we decrease the size of the buckets, this effect does seem to diminish, and the group with more years at their current company are only more highly paid in the 0-2 year experience bucket. For most levels of

experience, we see that median total compensation for employees who have been with their current employer for a shorter amount of time is higher than median compensation for employees who have worked at their current company longer, and that this difference is significant at alpha = .005.

2.2 Gender vs TYC

Numerous studies have shown that women, when controlling for industry, experience, education, and other qualitative factors earn less than their male counterparts. In 2021, The New York Times reported, on average American women will earn 82 cents for every dollar a man earns when controlling for various factors.[4] We observe that our dataset is deeply imbalanced, with male respondents making up 82% of those who listed a gender in their response.

H_0: $\boldsymbol{Mdn_f = Mdn_m}$ The salaries of men and women are not different, controlling for experience.
$\boldsymbol{H_a: Mdn_f != Mdn_m}$ The salaries of men and women are different, controlling for experience.

We separated respondents into Male and Female groups for each experience bucket based on the gender category. Many entries within this category do not have a response for this feature, which led to us dropping ~⅓ of the rows of our dataset for this test. This could invalidate our test if a large enough portion of non-respondents are high or low earners within their gender category, as we do not have access to this data. We view this as unlikely, however. This test also focuses solely on those who self-report gender as Male or Female. We once again ran a Mann-Whitney U test by bucket to compare the two groups.

P-values and Median Compensation by Gender

	Male	Female	P-val
0-5 Years	$169000	$165000	$2.1e^{-5}$
5-10 Years	$219000	$200000	$2.5e^{-22}$
10-15 Years	$256000	$228000	$4.23e^{-41}$

Takeaways: We reject the null hypothesis for all experience buckets. At each level of experience, median compensation for men and women differs, and this difference is significant at alpha = 0.005, with men always being paid more than women.

[4] https://www.nytimes.com/2021/03/24/us/equal-pay-day-explainer.html

2.3 Education vs TYC

Given the student loans incurred to pay for New York University's Master's Program, we were deeply interested in learning whether median compensation for those with Master's degrees differs from other groups.

H_0: $\boldsymbol{Mdn_{ed1} = Mdn_{ed2}}$ Salaries do not differ between groups with different levels of educational attainment, when controlling for experience.
$\boldsymbol{H_a: Mdn_{ed1} != Mdn_{ed2}}$ Salaries differ between groups with different levels of educational attainment, when controlling for experience.

Testing this hypothesis required us to run 6 pairwise tests for each experience bucket in order to compare data points with the following highest levels of educational attainment [high school, bachelors, masters, PhD]. For each experience bucket.

Takeaways: The differences in median total compensation for all categories were significantly different when compared with each other, with the exception of the pairwise test for those with bachelor's degrees vs those with only a high school degree. So we conclude that these groups have differing median levels of compensation, and differences between groups are statistically significant. (With the exception of those that have only high school degrees and those that have only bachelor's degrees) We did not include tables here, due to the large number of comparison tests, though these can be viewed in the notebook.

2.4 Industry vs TYC

When looking at New York University's MS in Data Science program industry placements for 2020-2021 alumni, approximately 75% of alumni were working in the technology or finance industries. [5] Anecdotally, we've heard that companies in those industries are usually the highest paying. We set out to see which of these two industries pays the most

H_0: $\boldsymbol{Mdn_t = Mdn_f}$ Tech and Finance salaries do not differ when controlling for experience.
$\boldsymbol{H_a: Mdn_t != Mdn_f}$ Tech and Finance salaries differ when controlling for experience.

In order to test compensation for tech workers v finance workers, we utilized the industry feature from our auxiliary dataset to categorize different employers into their respective industries. We then ran Mann-Whitney U tests to compare the salaries of tech and finance workers, after controlling for experience.

[5] https://cds.nyu.edu/placement-stats/

P-values and Median Compensation Tech vs Finance

	Tech	Finance	P-val
0-5 Years	$178000	$119000	0.0
5-10 Years	$233000	$155000	$1.35e^{-122}$
10+ Years	$302000	$200000	$3.08e^{-78}$

Takeaways: For all levels of experience, tech workers earn more than their finance counterparts, with this difference in medians being statistically significant at alpha = 0.005.

2.5 FAANG vs Non-FAANG TYC

Diving deeper into the question of which industries are the highest paying, specific companies, often referred to as FAANG (Facebook, now known as Meta, Apple, Amazon, Netflix and Google) are usually the most coveted internships and full time positions.

We were curious, even within the technology industry, are these companies paying a noticeable premium compared to competitors.

H_0: $\boldsymbol{Mdn_{faang} = Mdn_{nf}}$ The salaries of FAANG and non-FAANG tech employees do not differ when controlling for experience.
$\boldsymbol{H_a}$: $\boldsymbol{Mdn_{faang} != Mdn_{nf}}$ The salaries of FAANG and non-FAANG tech employees differ when controlling for experience.

Here we want to examine whether FAANG employees make more than their non-FAANG tech counterparts when controlling for total work experience. We sorted data points into two groups based on the “Employer” feature (FAANG employers v non-FAANG tech) and ran a Mann-Whitney U test by bucket to compare the groups.

Median Salaries for FAANG vs Non-FAANG and P-Values

	FAANG	Non-FAANG	P-Value
0-5 Years	200000	173000	$2.6e^{-220}$
5-10 Years	261000	220000	$7.5e^{-151}$
10-15 Years	320000	260000	$1.36e^{-116}$
15+	375000	290000	$1.05e^{-107}$

Takeaways: At all levels of experience, FAANG employees make more in total compensation than their non-FAANG tech counterparts. The difference in median compensation is statistically significant for all buckets at alpha = 0.005.

2.6 Region vs TYC

We were all similarly interested in how TYC differs between regions. Anecdotally, it seemed that the highest salaries were on the West Coast.

H_0: $\boldsymbol{Mdn_{R1} = Mdn_{R2}}$ Salaries do not differ by region when controlling for experience.
$\boldsymbol{H_a}$: $\boldsymbol{Mdn_{R1}} != \boldsymbol{Mdn_{R2}}$ Salaries differ by region when controlling for experience

In order to compute this test, we transformed our location column (city, state) into a region column in order to reduce the number of possible values to 4 [Northeast, West, South, Midwest]. This gave us more data to work with by experience bucket. Once again, we controlled for experience and ran 6 pairwise Mann-Whitney U tests per bucket in order to compare median salaries by region.

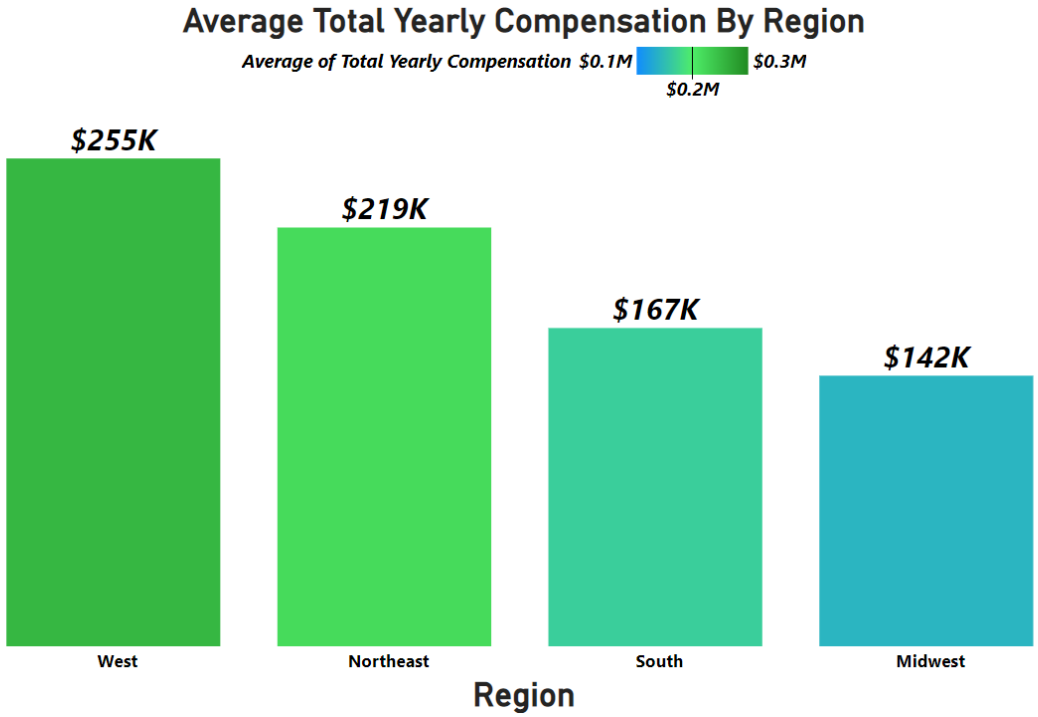


Figure 4: Average TYC by Region

Takeaways: For each experience bucket, we rejected the null hypothesis for every pairwise test. Median salaries by region were highest in the West, followed by the Northeast, South and Midwest, in that order. This pattern was the same for all experience buckets with all differences being significant at alpha = 0.005. Again, due to the number of comparisons, we have not included tables here, though these are visible in the notebook.

3.0: Predictive Modeling

Here, we turned to the primary challenge of predicting salary given our explanatory variables. Before building a model, we had to massage our data into the right format. Transforming categorical variables into one hot encoded variables was an essential step towards building our model. Afterwards, to normalize for different magnitudes of standard deviations in our explanatory numerical variables, we turned our quantitative variables into Z-scores by subtracting the mean of the variable from each value, and dividing by the standard deviation.

This allowed us to better understand a per unit normalized impact of our explanatory variable on a normalized version of salary.

3.1 Model choices

As a starting point, our team decided to use simple linear regression of all of the explanatory variables to understand how a base level linear model would perform. To understand how well our model performed on any specific subset of the data, our team used 10-fold cross validation and computed the mean squared error and R^2 of each fold.

The output of the simple linear regression can be found in **Figure 5: Ordinary Linear Regression Outputs**.

The model had an R^2 of .461 with 62 independent variables trying to predict TYC. Unsurprisingly, the most impactful variables, determined by looking at the significance of the regression weights were:

1. Normalized Years of Experience with a T statistic of 57.14 and weight of .462
2. Non FAANG Indicator with a T statistic of -26.5 and a weight of -.3984
3. Title of Software Engineering Manager with a T Statistic of 17.92 and weight of .899
4. PhD holders with a T Statistic of 15.3 and a weight of .4882

To prevent overfitting, and remove potentially correlated independent variables from creating noise in our predictions, we ran a LASSO regression on the dataset testing λ values of .0001 to 1 in increments of .0001. In tuning the LASSO parameter, we found the optimal L1 penalizer was .001.

The LASSO regression created a more balanced model, where mean training and testing error was .544 and .548, respectively for the optimized value of λ = .001. The R^2 of the LASSO model was .451, seeing a slight decrease in performance compared to OLS

However, a benefit of LASSO regression is that it can set regression weights of our independent variables to 0, indicating these variables were not important. After analyzing the regression coefficients we found the following variables to have a 0 regression weight:

1. Industry: Food & Drug Stores
2. Industry: Energy
3. Industry: Food, Beverages & Tobacco
4. Industry: Chemicals
5. Industry: Household Products
6. Industry: Industrials
7. Tag: MediaTek
8. Industry: Motor Vehicles & Parts
9. Tag: Product
10. Tag: User Experience (UX)
11. Industry: Hotels, Restaurants & LEISURE
12. Industry: Retailing
13. Tag: other
14. Tag: Hardware Engineer
15. Tag: Mechanical Engineer
16. Industry: Apparel
17. Tag: Solution Architect
18. Tag: Technical Program Manager
19. Education: Highschool

and were thus excluded from our final model. The results of the optimal LASSO regression coefficient weights can be found in **Figure 7: Optimized LASSO regression coefficients**.

Believing that the relationship between TYC and our independent variables was not fully linear, we sought to use a model that could better capture non-linear relationships. The team decided to use the XGBoost model, given it's fast computation, ability to capture non-linear relationships, and an ensemble approach that would minimize variance in model performance when comparing different datasets.

3.2 Final Predictive Model

Our final predictive model was an XGBoost regressor on the remaining 43 variables. The model had a mean squared train error of .0327 versus a mean squared test error of .461, and an R^2 of **.538**

To determine the optimal value of the XGBoost hyperparameters, we manually tried values of:

1. n_estimators in the set of {100, 250, 500, 100}
2. max_depth in the set of {2,4,8,16}

When looking at model performance on a cross validated training and testing set using different hyperparameters, we found that the optimal set of hyperparameters were n_estimators = 500, max_depth = 3.

The increase in performance from the XGBoost can be attributed to non-linear relationships between our independent variables and TYC. For example, with some sub-groups an increase in years of experience is not as strongly correlated with an increase in salary, most notably non-engineer and technical roles. By taking into account these non-linear relationships, XGBoost is able to create a more accurate prediction, and since it is an ensemble model composed of relatively shallow learners the variance between train and test error is reasonable.

4.0 Conclusion and Take-aways

4.1: Findings on Data Science and Tech Salaries

In our modeling approach, we noticed a noticeable improvement in R^2 when using the XGBoost model as compared to OLS. This highlights a fundamental concept in modeling, the Bias-Variance tradeoff.[6] Illustrated well in page 25 of Introduction To Statistical Learning (see **Figure 6: Bias variance Tradeoff Graphic from ISLR)**, the LASSO regression is one of the most biased models, but also one of the most easily interpretable. Whereas XGBoost, a boosted model, is one of the most flexible, yet uninterpretable models.

Our team's approach of first understanding what linear relationships exist between our independent and dependent variable through OLS, feature pruning using LASSO and then deriving a high performing boosted model.

4.2: Takeaways on Predictive Modeling

As a learning assessment, it seems that our model did not improve much throughout the process of model building and tuning. However, it is worth noting that there were many substantial iterative improvements that helped us to capture more signal in predicting salary.

Standardizing our quantitative variables, further data cleaning to include tag, region, FAANG vs non FAANG all drove lift in model performance as we sought to better improve the predictive power.

When comparing our initial simple linear regression, that did not use the above information to predict salary, we observed an R^2 of .31 compared to a much improved performance of an R^2 of .538 after incorporating these new features and transformations.

[6] https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf page 25

Our group’s main takeaway from this process is that to build a better performing model, the quality of your predictor variables is more important than the model used. A linear regression can outperform complex, non-linear models such as XGBoost if the linear model includes more variables that have a strong relationship with the dependent variable.

Appendix:

Figure 1: Kaggle Dataset Description

Column Name	Data Description	Data Type	Example Values
Timestamp	Timestamp when the respondent submitted their response	Timestamp	11:13:42 2018/05/03
Company	Company of the respondent	String	Google, IBM, etc.
Level	Specific job level of the respondent	String	L7, Senior SDE, etc.
Title	Broader job function for a respondent	String	Data Scientist, Software Engineer
Total Yearly Compensation	Respondent's yearly compensation including base pay, and bonuses	Float	$325,000
Location	City, State, Country of a respondent	String	Cupertino, CA, US
Years of Experience	Respondent's total years of experience	Float	7.75
Years at Company	Respondent's total years at specified company	Float	2.5
Tag	Specific expertise of a respondent	String	Full Stack, ML / AI, Web, etc.
Base Salary	Base Salary of a respondent	Float	$121,100.50
Stock Grant Value	Value of stock grant compensation of a respondent	Float	$25,550.75
Bonus	Value of yearly bonus of a respondent	Float	$11,250
Gender	Gender of respondent	String	Male, Female, Other, etc.
Other Details	Free form answers from the respondent	String	$10k Relocation bonus, etc.
City Id	Numerical Identifier for the respondent's city	Integer	7472 (Sunnyvale CA), etc.
Education	Categorical variable that describes highest level of education	String	Masters, PHD, etc.
Masters Degree	Binary variable indicating an MS is the highest education obtained	Integer	1,0
Bachelors Degree	Binary variable indicating a BS is the highest education obtained	Integer	1,0
Doctorate Degree	Binary variable indicating a PHD is the highest education obtained	Integer	1,0
High School	Binary variable indicating high school is the highest education obtained	Integer	1,0
Some College	Binary variable indicating some college is the highest education obtained	Integer	1,0
Race	Categorical variable that describes highest level of education	String	White, Hispanic, Black
Race Asian	Binary variable indicating respondent is Asian	Integer	1,0
Race White	Binary variable indicating respondent is White	Integer	1,0
Race Two or More	Binary variable indicating respondent is multi-racial	Integer	1,0
Race Black	Binary variable indicating respondent is Black	Integer	1,0
Race Hispanic	Binary variable indicating respondent is Hispanic	Integer	1,0
Sector	Categorical variable describing the industry of the company	String	Finance, Healthcare, Technology, etc.
Region	Categorical variable describing the region of the company	String	West, Northeast, South, Midwest

X

Figure 2: Feature Missingno Plots

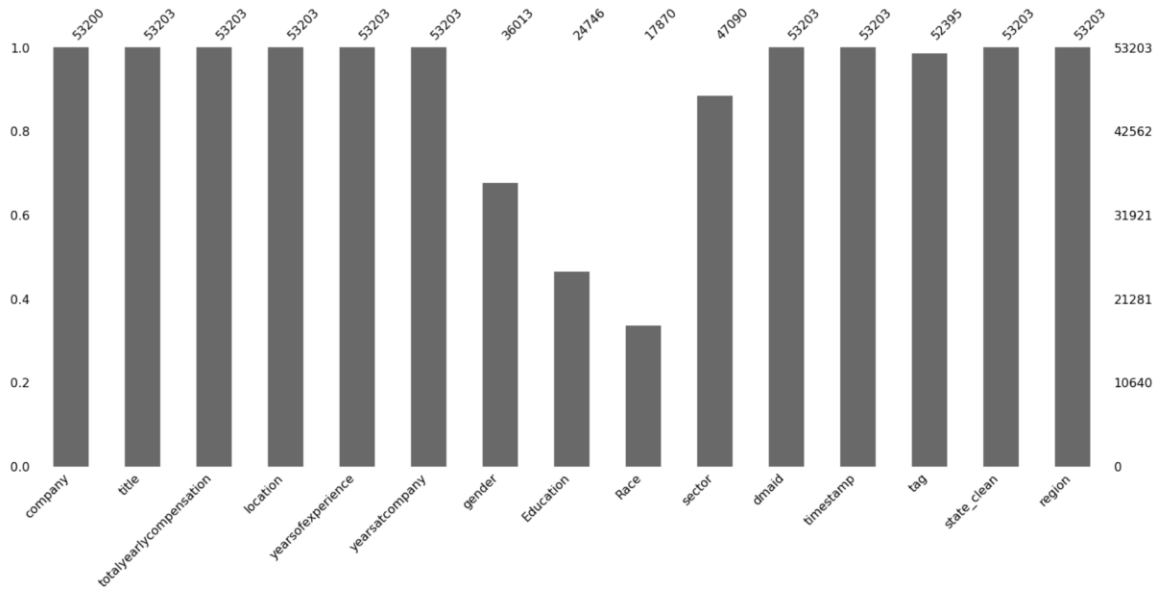
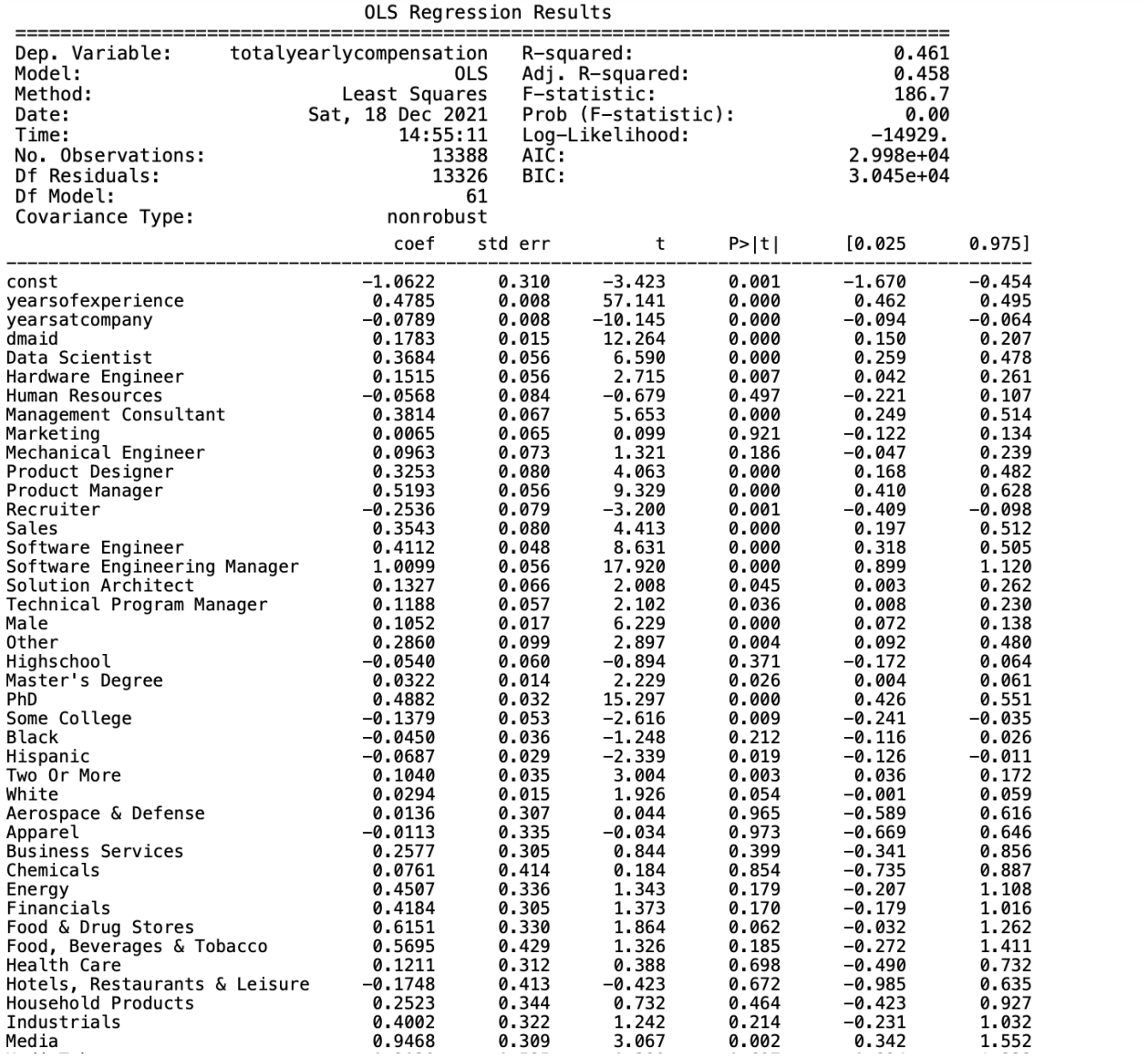


Figure 5: Simple Linear Regression Outputs



OLS Regression Results

Dep. Variable:	totalyearlycompensation	R-squared:	0.461
Model:	OLS	Adj. R-squared:	0.458
Method:	Least Squares	F-statistic:	186.7
Date:	Sat, 18 Dec 2021	Prob (F-statistic):	0.00
Time:	14:55:11	Log-Likelihood:	-14929.
No. Observations:	13388	AIC:	2.998e+04
Df Residuals:	13326	BIC:	3.045e+04
Df Model:	61		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.0622	0.310	-3.423	0.001	-1.670	-0.454
yearsofexperience	0.4785	0.008	57.141	0.000	0.462	0.495
yearsatcompany	-0.0789	0.008	-10.145	0.000	-0.094	-0.064
dmaid	0.1783	0.015	12.264	0.000	0.150	0.207
Data Scientist	0.3684	0.056	6.590	0.000	0.259	0.478
Hardware Engineer	0.1515	0.056	2.715	0.007	0.042	0.261
Human Resources	-0.0568	0.084	-0.679	0.497	-0.221	0.107
Management Consultant	0.3814	0.067	5.653	0.000	0.249	0.514
Marketing	0.0065	0.065	0.099	0.921	-0.122	0.134
Mechanical Engineer	0.0963	0.073	1.321	0.186	-0.047	0.239
Product Designer	0.3253	0.080	4.063	0.000	0.168	0.482
Product Manager	0.5193	0.056	9.329	0.000	0.410	0.628
Recruiter	-0.2536	0.079	-3.200	0.001	-0.409	-0.098
Sales	0.3543	0.080	4.413	0.000	0.197	0.512
Software Engineer	0.4112	0.048	8.631	0.000	0.318	0.505
Software Engineering Manager	1.0099	0.056	17.920	0.000	0.899	1.120
Solution Architect	0.1327	0.066	2.008	0.045	0.003	0.262
Technical Program Manager	0.1188	0.057	2.102	0.036	0.008	0.230
Male	0.1052	0.017	6.229	0.000	0.072	0.138
Other	0.2860	0.099	2.897	0.004	0.092	0.480
Highschool	-0.0540	0.060	-0.894	0.371	-0.172	0.064
Master's Degree	0.0322	0.014	2.229	0.026	0.004	0.061
PhD	0.4882	0.032	15.297	0.000	0.426	0.551
Some College	-0.1379	0.053	-2.616	0.009	-0.241	-0.035
Black	-0.0450	0.036	-1.248	0.212	-0.116	0.026
Hispanic	-0.0687	0.029	-2.339	0.019	-0.126	-0.011
Two Or More	0.1040	0.035	3.004	0.003	0.036	0.172
White	0.0294	0.015	1.926	0.054	-0.001	0.059
Aerospace & Defense	0.0136	0.307	0.044	0.965	-0.589	0.616
Apparel	-0.0113	0.335	-0.034	0.973	-0.669	0.646
Business Services	0.2577	0.305	0.844	0.399	-0.341	0.856
Chemicals	0.0761	0.414	0.184	0.854	-0.735	0.887
Energy	0.4507	0.336	1.343	0.179	-0.207	1.108
Financials	0.4184	0.305	1.373	0.170	-0.179	1.016
Food & Drug Stores	0.6151	0.330	1.864	0.062	-0.032	1.262
Food, Beverages & Tobacco	0.5695	0.429	1.326	0.185	-0.272	1.411
Health Care	0.1211	0.312	0.388	0.698	-0.490	0.732
Hotels, Restaurants & Leisure	-0.1748	0.413	-0.423	0.672	-0.985	0.635
Household Products	0.2523	0.344	0.732	0.464	-0.423	0.927
Industrials	0.4002	0.322	1.242	0.214	-0.231	1.032
Media	0.9468	0.309	3.067	0.002	0.342	1.552

MediaTek	0.2039	0.525	0.389	0.697	−0.824	1.232
Motor Vehicles & Parts	0.2903	0.309	0.938	0.348	−0.316	0.897
Retailing	0.3991	0.307	1.300	0.194	−0.203	1.001
Technology	0.5938	0.304	1.954	0.051	−0.002	1.189
Telecommunications	0.1814	0.308	0.589	0.556	−0.423	0.786
Transportation	1.0978	0.322	3.411	0.001	0.467	1.729
Northeast	0.4035	0.037	11.005	0.000	0.332	0.475
South	−0.0321	0.036	−0.896	0.370	−0.102	0.038
West	0.1852	0.041	4.511	0.000	0.105	0.266
2021	0.0702	0.013	5.282	0.000	0.044	0.096
non_faang	−0.3984	0.015	−26.484	0.000	−0.428	−0.369
Data	−0.0079	0.040	−0.198	0.843	−0.086	0.070
DevOps	−0.0899	0.050	−1.782	0.075	−0.189	0.009
Distributed Systems (Back-End)	0.1877	0.028	6.799	0.000	0.134	0.242
Full Stack	−0.0072	0.027	−0.266	0.790	−0.060	0.046
ML / AI	0.3471	0.037	9.347	0.000	0.274	0.420
Product	0.0638	0.050	1.267	0.205	−0.035	0.163
Technical	−0.0032	0.056	−0.057	0.954	−0.113	0.107
User Experience (UX)	0.0631	0.083	0.759	0.448	−0.100	0.226
Web Development (Front-End)	0.0462	0.038	1.209	0.227	−0.029	0.121
other	0.0404	0.027	1.495	0.135	−0.013	0.093

Omnibus:	5065.778	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30221.800
Skew:	1.706	Prob(JB):	0.00
Kurtosis:	9.521	Cond. No.	458.

Figure 6: Bias Variance Trade-off Graphic from ISLR

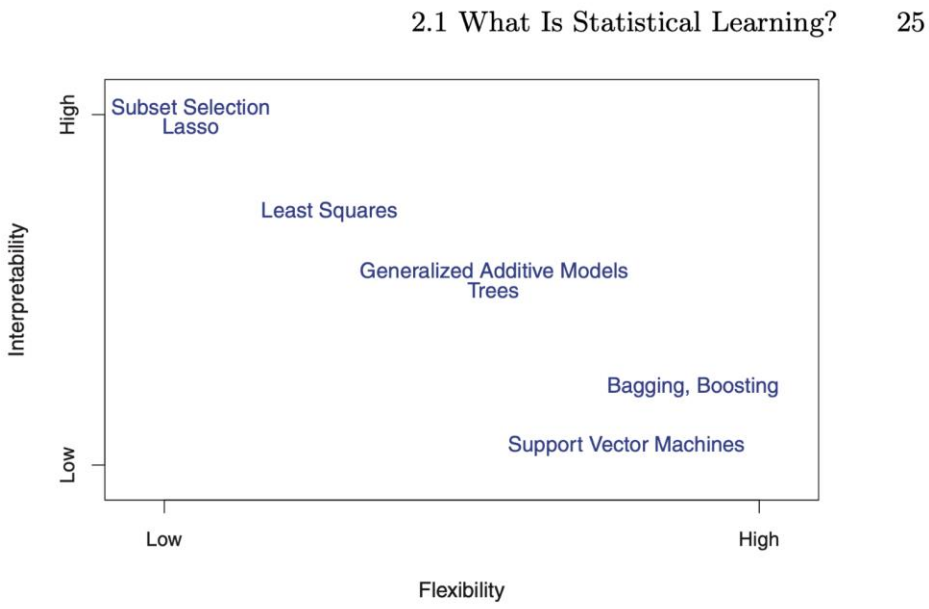


Figure 7: Optimized LASSO regression coefficients

	regression_coef
non_faang	-0.397089
Recruiter	-0.313211
Aerospace & Defense	-0.264672
Health Care	-0.119033
Telecommunications	-0.096513
yearsatcompany	-0.078184
DevOps	-0.067433
Marketing	-0.066647
Human Resources	-0.065266
Some College	-0.060029
Hispanic	-0.054831
Business Services	-0.049698
South	-0.040516
Black	-0.028123
Full Stack	-0.022712
Technical	-0.022641
Data	-0.004663
Food & Drug Stores	0.000000
Energy	0.000000
Food, Beverages & Tobacco	0.000000
Chemicals	-0.000000
Household Products	-0.000000
Industrials	0.000000
MediaTek	-0.000000
Motor Vehicles & Parts	-0.000000
Product	0.000000
User Experience (UX)	0.000000
Hotels, Restaurants & Leisure	-0.000000
Retailing	0.000000
other	0.000000

	regression_coef
Hardware Engineer	0.000000
Mechanical Engineer	-0.000000
Apparel	-0.000000
Solution Architect	-0.000000
Technical Program Manager	-0.000000
Highschool	-0.000000
Web Development (Front-End)	0.004543
White	0.022567
Master's Degree	0.032000
Financials	0.052764
Other	0.062745
2021	0.069184
Two Or More	0.081441
Male	0.096889
Sales	0.112233
Distributed Systems (Back-End)	0.163339
Management Consultant	0.168751
West	0.169685
dmaid	0.179009
Product Designer	0.179399
Data Scientist	0.201869
Technology	0.253646
Software Engineer	0.260478
ML / AI	0.313277
Product Manager	0.383785
Northeast	0.392114
yearsofexperience	0.477503
PhD	0.484665
Transportation	0.489744
Media	0.533163
Software Engineering Manager	0.840750