Michael Loos

Show Me the Money

Searching for a job can be quite an arduous process. Applications, resumes, phone and in person interviews. Often, the applicant may go through this entire process without a good idea of what the position might pay until they have been selected for the position and salary negotiations begin. A company may provide a salary range during the application process, but this can be a wide range that does not provide meaningful information to the applicant. For example, I recently saw a posting that had a salary range of $80,000-$130,000. This range is hardly anything to go on at all. Someone making $100,000 could either completely waste their time if the salary ends up being on the lower end or significantly increase their pay if the salary is on the higher end of the range. Applicants (and employers) could save time and effort by providing an accurate salary based on the characteristics of the position and the applicant's attributes, such as education, experience, and skills. One way to make this prediction is to use linear regression to predict

Data Explanation

Data was collected through a 2019 Kaggle survey (See link in references, 1) to employees within the Data Science and Machine Learning industries. More than 19,000 responses were received from 171 countries and territories. The survey was sent through Kaggle channels, including email lists, forums, and social media. Suspected spam responses were excluded from the data. Responses were anonymized and left as raw data. Results of the survey include data about respondents' roles, respondent characteristics, and opinions on the best way for new data scientists to enter the industry.

The survey had 34 questions ranging from age to which relational database products do you use on a regular basis. Most of these questions allowed for multiple responses (select all that apply). Many questions were removed from the analysis as they did not seem likely to contribute to the predictability of salary. For example, "Who/what are your favorite media sources that report on data science topics?" While there could be correlation to salary, it is unlikely getting news from Twitter vs Reddit would have an impact on what an employer is likely to pay.

Questions used in analysis:

- What is your current yearly compensation (approximate $USD)?
- What is your age (# years)?
- What is your gender? - Selected Choice
- In which country do you currently reside?
- What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
- Select the title most similar to your current role (or most recent title if retired): - Selected Choice
- What is the size of the company where you are employed?

- What programming languages do you use on a regular basis? (Select all that apply)
- Which of the following ML algorithms do you use on a regular basis? (Select all that apply)
- Which categories of ML tools do you use on a regular basis? (Select all that apply)
- Which of the following machine learning frameworks do you use on a regular basis? (Select all that apply)

Additional alterations were made to the data set. All responses from outside of the United States were removed in order to make the data more relevant and to create a more predictive model. Any responses with no compensation listed were removed. Responses to "What is your current yearly compensation (approximate $USD)" were done in ranges. To prepare the data for linear regression, the mean of the range was assigned for the salary. Any salaries less than $15,000 were also removed as a salary below that would be less than minimum wage, assuming full-time position. Any salary greater than $500,000 was assigned a salary of $500,001 since there was no upper end to the range and as to not be overly influential. Finally, questions with multiple answers were converted to dummy variables in the model process.

Methods

A multiple linear regression model was determined to be the simplest and most transparent option for a model. Ideally, the model could incorporate all variables and be able to show the influence of each variable. The initial model used age, gender, education, role title, company size, programming language, and machine learning algorithms as input variables, with mean salary as the target variable. Non-predictive variables were removed until a final model was reached.

First, Python was used to initially view and clean the data. The data was filtered to only included responses from the United States and drop and responses without a response to the compensation question or salary under $15,000. Then, new variables were created to better format the salary ranges and to create a single salary value, the mean salary of the range, for each respondent.

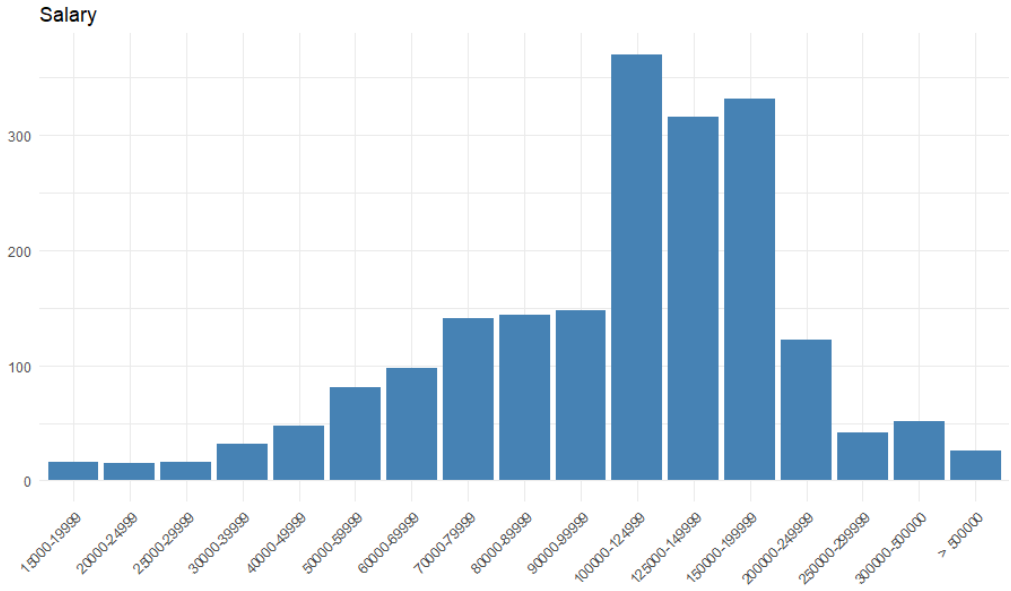
R was then used for some further formatting, such as factoring and re-ordering variables. Exploratory Data Analysis, including summary statistics and visualizations, were done to obtain a feel for the data and possible explanations to salary. Finally, a multiple linear regression model was created. The model was trained and tested using an 80/20 split.

Exploratory Data Analysis (EDA)

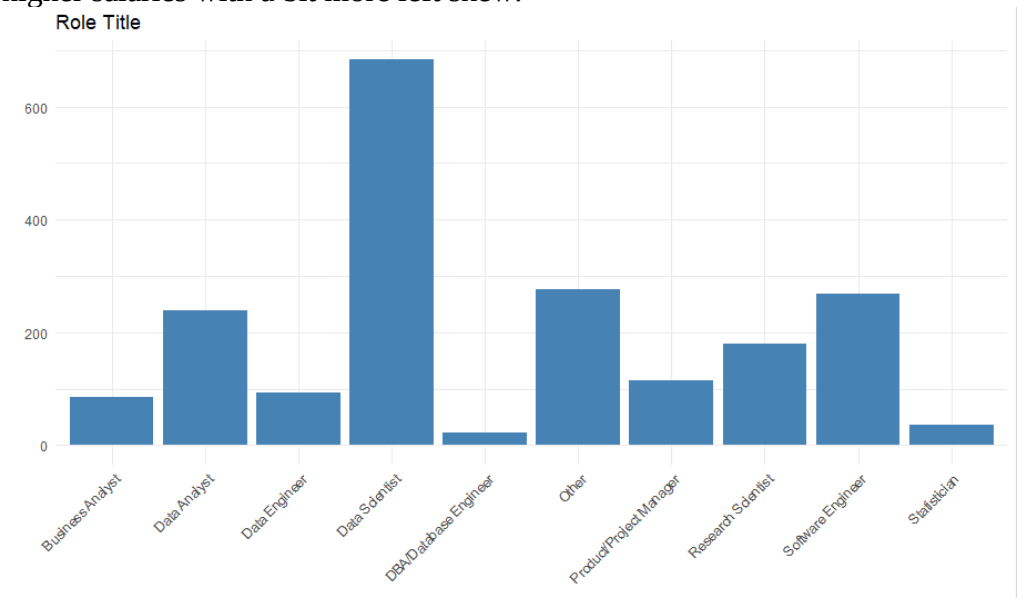
The first step in EDA process was to obtain summary statistics on the data. This was done for the salary variable, shown below, as well as salary by several variables such as role title, gender, and education (see appendix).

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17500   85000  112500  134073  175000  500001
```

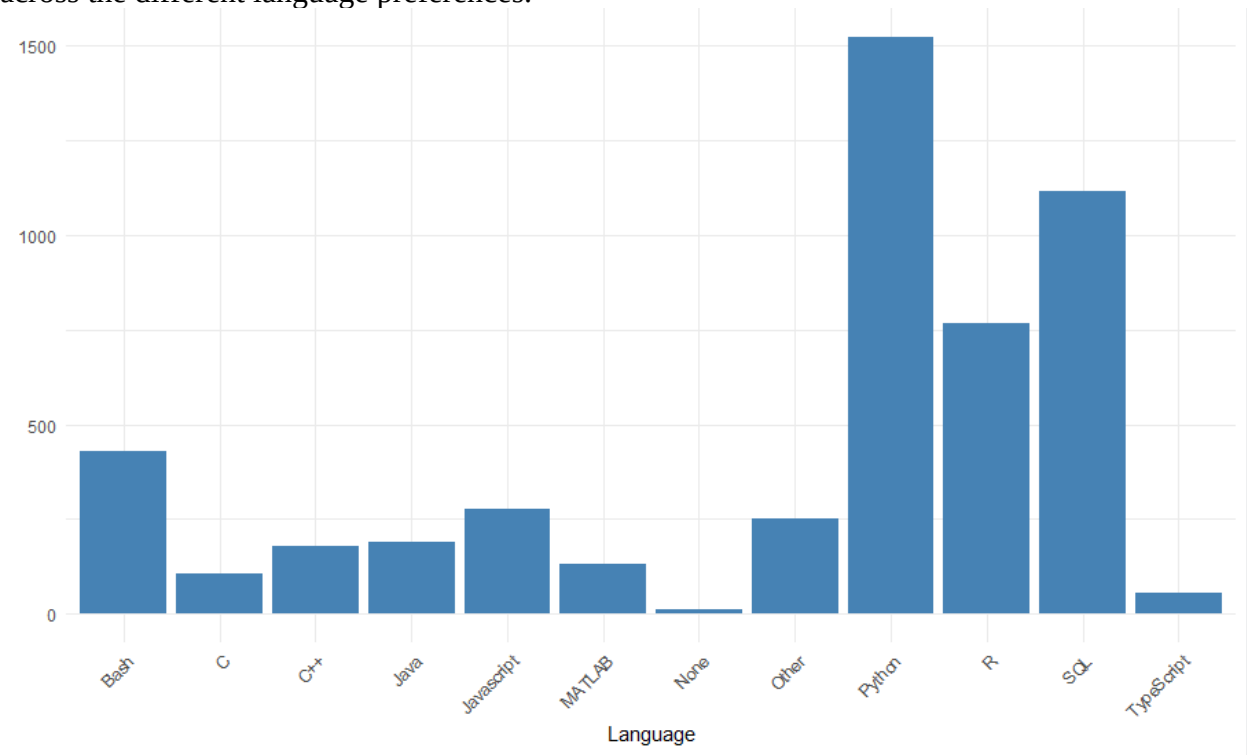
Next, visualizations were done to view distributions of salary, age, gender, education, role title, and preferred programming languages. (See appendix for additional visualizations). The distribution of salaries can be seen below. As our summary statistics had shown, the majority of salaries for someone in the Data Science field centers in the $100,000-$200,000 range.

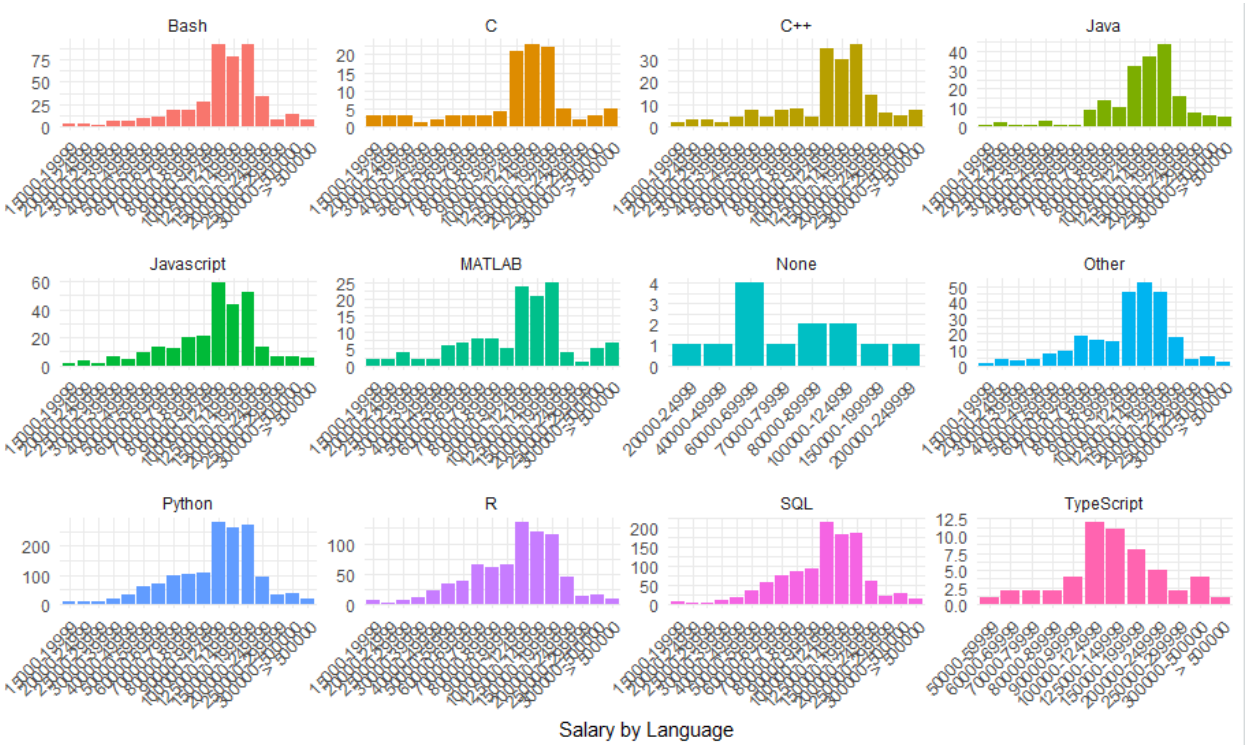


There are a wide range of role titles in the Data Science industry, but Data Scientist is clearly the most common. Most of the roles have similar distributions. Business Analyst and Data Analyst seem to have lower, more symmetrical distributions. Data Engineer, Data Scientist, Product/Project Manager, and Software Engineer also have similar distributions, centered around higher salaries with a bit more left skew.

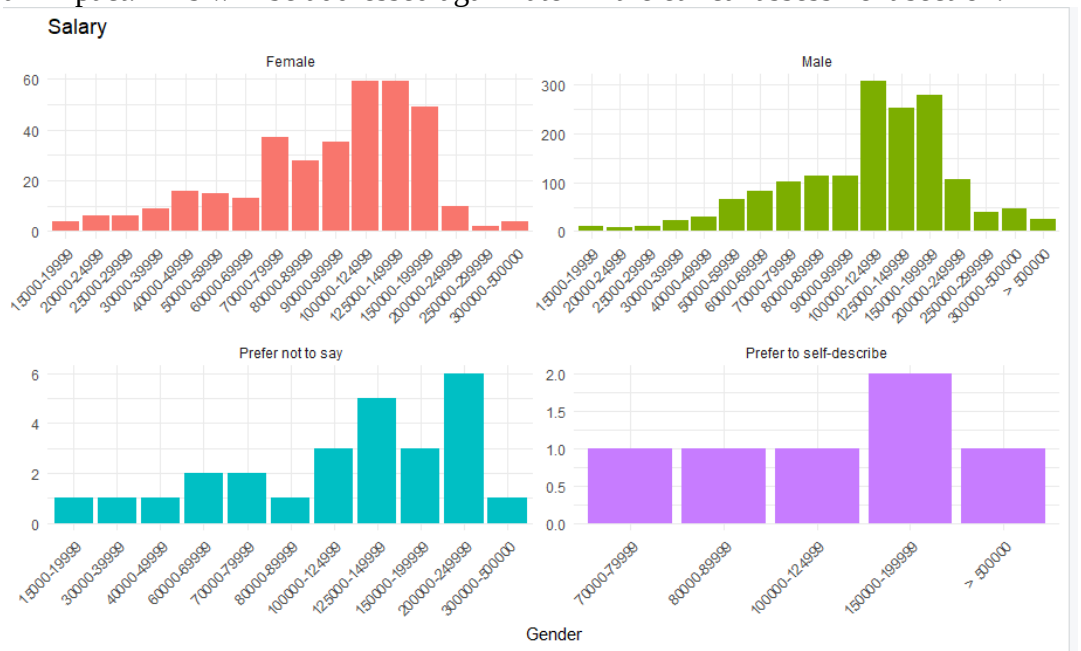


A wide range of languages are used in the Data Science field. According to the survey results, Python, R, and SQL are heavy favorites (or requirements). Salaries appear to be pretty similar across the different language preferences.





One interesting characteristic is salary by gender. In the survey results, male have the clear majority of responses. At first glance, the salary distributions seem similar. Males might have slightly higher salaries. What makes these this variable interesting is an employer should not take gender into account when assessing pay. Perhaps multicollinearity exists between gender and other variables that could lead to a higher salary. However, by law, gender itself should not have an impact. This will be addressed again later in the ethical assessment section.



Model Data Analysis

After training the model we get the following summary:

```
Residual standard error: 74310 on 1543 degrees of freedom
Multiple R-squared:  0.2101,    Adjusted R-squared:  0.1819
F-statistic:  7.46 on 55 and 1543 DF,  p-value: < 0.00000000000000022
```

With F=7.46, we can conclude there is a relationship between our predictor and response variables. However, the model is not a great fit with an adjusted R-squared value of 0.1819 and an very RMSE. Below are the coefficients, many of which are statistically significant at p=0.05.

```
                                                      Estimate Std. Error    t value        Pr(>|t|)
(Intercept)                                        80772.7184  38897.402  2.0765582 0.0380075322842
Age_22-24                                          28092.7374  25560.994  1.0990471 0.2719189589657
Age_25-29                                          47822.7691  24683.138  1.9374671 0.0528705565923
Age_30-34                                          63953.5673  24727.673  2.5863156 0.0097916704833
Age_35-39                                          84838.1734  24855.624  3.4132385 0.0006585692554
Age_40-44                                          84166.5403  25033.544  3.3621503 0.0007922924991
Age_45-49                                         109014.6044  25179.225  4.3295457 0.0000159037676
Age_50-54                                          92162.7182  25423.377  3.6251171 0.0002982120468
Age_55-59                                         104731.7764  25876.680  4.0473422 0.0000543683297
Age_60-69                                         100063.6933  25969.378  3.8531417 0.0001213915317
Age_70+                                           107950.2594  28879.917  3.7379006 0.0001923354399
Gender_Male                                        16713.0282   5133.910  3.2554191 0.0011567972932
Gender_Prefer not to say                            2171.1715  16875.728  0.1286565 0.8976462541478
Gender_Prefer to self-describe                     26919.9980  31505.459  0.8544550 0.3929855573199
Education_High School                             -53703.6451  46221.037 -1.1618875 0.2454608576968
Education_Some college                            -77780.0987  32916.819 -2.3629288 0.0182545996980
Education_Professional degree                      -4740.9917  34950.747 -0.1356478 0.8921174133988
Education_Bachelor's degree                       -59609.8028  31575.179 -1.8878690 0.0592310270723
Education_Master's degree                         -61111.1414  31514.506 -1.9391432 0.0526660040096
Education_Doctoral degree                         -49126.6017  31856.368 -1.5421282 0.1232475018450
Role_Product/Project Manager                       -5556.0161   9119.685 -0.6092333 0.5424594731541
Role_Data Analyst                                 -33193.7149   6709.589 -4.9472058 0.0000008352001
Role_Other                                         -5474.6014   6304.435 -0.8683730 0.3853252265994
Role_Research Scientist                           -31744.1198   7520.222 -4.2211678 0.0000257212776
Role_Data Engineer                                 -5983.3194   9373.598 -0.6383162 0.5233625793195
Role_Business Analyst                             -18062.7829  10301.929 -1.7533398 0.0797422444335
Role_Software Engineer                             -3065.1249   7175.600 -0.4271594 0.6693228598649
Role_Statistician                                 -13067.3918  15195.135 -0.8599720 0.3899380434243
Role_DBA/Database Engineer                        -41036.7213  18450.686 -2.2241298 0.0262839587496
Company_Size50-249 employees                       12491.1643   6755.016  1.8491688 0.0646246026061
Company_Size250-999 employees                       1296.9264   7106.136  0.1825080 0.8552080755605
Company_Size1000-9,999 employees                   12063.5911   5958.860  2.0244798 0.0430927729348
Company_Size> 10,000 employees                     23204.9308   5676.991  4.0875405 0.0000458406866
Language_1No_answer                                -3878.6973   5424.010 -0.7150978 0.4746568144027
Language_2R                                       -10479.5248   4278.132 -2.4495562 0.0144137421172
Language_3SQL                                       4203.2874   4428.360  0.9491747 0.3426803781794
Language_4C                                       -11974.5492   9963.927 -1.2017902 0.2296292390941
Language_5C++                                      12977.9445   7737.458  1.6772879 0.0936887382005
Language_6Java                                      8227.0621   7013.611  1.1730137 0.2409713704123
Language_7Javascript                              -16270.6587   6161.090 -2.6408735 0.0083527153327
Language_8TypeScript                               35595.1297  12525.568  2.8417976 0.0045449812426
Language_9Bash                                      5012.3342   4938.895  1.0148696 0.3103270772331
Language_10MATLAB                                   9238.3505   7932.382  1.1646376 0.2443457492740
Language_11None                                   -42339.8529  26043.210 -1.6257540 0.1042062082438
Language_12Other                                    4207.0572   5887.754  0.7145437 0.4749991483999
Algo_1No_answer                                    14336.3680   5734.687  2.4999392 0.0125245826393
Algo_2No_answer                                      974.0485   5119.090  0.1902777 0.8491165873125
Algo_3Gradient Boosting Machines (xgboost, lightgbm, etc)  12964.1373   4723.352  2.7446899 0.0061270276704
Algo_4Bayesian Approaches                           8561.2247   4607.114  1.8582619 0.0633221569013
Algo_5Evolutionary Approaches                      21360.5644   8499.588  2.5131293 0.0120677582497
Algo_6Dense Neural Networks (MLPs, etc)            11283.3101   5657.132  1.9945284 0.0462700192703
Algo_7No_answer                                     7448.8379   5401.330  1.3790748 0.1680715511054
Algo_8Generative Adversarial Networks              -7378.3690   9260.618 -0.7967469 0.4257206048727
Algo_9Recurrent Neural Networks                     6992.7936   5948.367  1.1755821 0.2399432547073
Algo_10Transformer Networks (BERT, gpt-2, etc)     14491.4659   7979.841  1.8160095 0.0695629180806
Algo_11None                                       -14898.7833   8239.978 -1.8081097 0.0707840648311
```

After removing less predictive variables, a final model was reached. While the R-squared value

is lower, the F-statistic is higher, the diagnostics look better, and most of the variables are statistically significant. However, the model once again had a high RMSE. (Too high to even display.)

```
Residual standard error: 76260 on 1572 degrees of freedom
Multiple R-squared:  0.1525,    Adjusted R-squared:  0.1384
F-statistic: 10.88 on 26 and 1572 DF,  p-value: < 0.00000000000000022
```

```
Coefficients:
                                Estimate Std. Error t value             Pr(>|t|)
(Intercept)                      65441.9    25139.2   2.603             0.009323 **
Age_22-24                         2869.7    25617.5   0.112             0.910822
Age_25-29                        25307.9    24633.5   1.027             0.304400
Age_30-34                        42280.7    24612.3   1.718             0.086017 .
Age_35-39                        63723.0    24730.6   2.577             0.010066 *
Age_40-44                        63929.0    24920.5   2.565             0.010400 *
Age_45-49                        82592.4    25150.1   3.284             0.001046 **
Age_50-54                        69624.5    25305.5   2.751             0.006003 **
Age_55-59                        83651.2    25690.8   3.256             0.001154 **
Age_60-69                        76594.4    25856.3   2.962             0.003099 **
Age_70+                          87559.4    28860.7   3.034             0.002454 **
Gender_Male                      23076.3     5157.5   4.474 0.0000082170600616 ***
Gender_Prefer not to say         10553.5    17040.4   0.619             0.535796
Gender_Prefer to self-describe   54247.3    31800.6   1.706             0.088232 .
Role_Product/Project Manager    -16960.2     8978.9  -1.889             0.059088 .
Role_Data Analyst               -49712.9     6463.1  -7.692 0.0000000000000254 ***
Role_Other                      -14109.0     6112.5  -2.308             0.021117 *
Role_Research Scientist         -25511.6     7091.8  -3.597             0.000331 ***
Role_Data Engineer              -12138.8     9329.3  -1.301             0.193401
Role_Business Analyst           -31535.3    10210.0  -3.089             0.002046 **
Role_Software Engineer           -8630.6     6286.3  -1.373             0.169972
Role_Statistician               -21170.2    15200.5  -1.393             0.163898
Role_DBA/Database Engineer      -51243.6    18441.8  -2.779             0.005523 **
Company_Size50-249 employees     13881.6     6822.9   2.035             0.042063 *
Company_Size250-999 employees      926.3     7154.3   0.129             0.897002
Company_Size1000-9,999 employees 14387.5     6020.9   2.390             0.016984 *
Company_Size> 10,000 employees   24245.7     5720.1   4.239 0.0000237907944077 ***
```

We can interpret our model to say a salary for an 18-21 year old, who is female, with a role title of Data Scientist, at a company with <50 employees would be $65,442. This salary would increase with age (likely associated with experience), being male, and at a larger company. Interestingly, any other role title would reduce the salary.

Conclusion

The summary statistics of the model are pretty disappointing. While this model may not be great at predicting an exact salary, I do feel it led to some insights. The role of Data Scientist appears to be the most well-paying job. This can help get a feel for potential differences in pay for different roles, especially where other variables are constant. While not unexpected, salary appears to increase significantly with age. Surprisingly, factors such as level of education and programming languages do not have as much impact on salaries. Finally, being male increases salary. While this could be a product of other factors, this could be a societal or legal problem.

Assumptions

Due to the nature of the data, a couple of assumptions had to be made. The first assumption is all responses to the survey were answered truthfully. Enough overinflated salaries could cause issues with the model. The second assumption, and the biggest, is using the midpoint of the salary ranges as the target variable. Assuming the salaries are normally distributed within each range, this should hopefully be representative of the true salaries.

Limitations and Recommendations

As stated earlier, salaries were given within previously specified ranges. To create a linear regression model, a numeric target variable was needed. Thus, the midpoint of each range was assigned as the salary for the respondent. Another challenge is the lack of numerical variables. While the type of data doesn't lend itself to having a lot of numeric variables, having nothing but categorical variables made it difficult.

Future Use and Implementation

There could be several uses for this model. The first would be the primary reason for creating the model, for personal use. Applicants could input known variables into the model to help provide an expectation for salaries when applying for positions or to understand where their current salary might be. The model could also be used to help consult organizations on appropriate salary ranges for roles and candidates. Finally, the model could be used on sites like LinkedIn. LinkedIn could use the description of the role and its requirements to determine a predicted salary to display to potential applicants.

Ethical Assessment

The largest ethical concern is potential disparate impact when using the model. If factors such as age and sex/gender are included in the data and these groups are disproportionally affected, there could be significant legal issues. While it would not guarantee an absence of disparate impact, it would be better not to include these variables in the model.

References

1. Kaggle Data: https://www.kaggle.com/c/kaggle-survey-2019/data

Appendix

Table 1

Q5	mean	median	q1	q3	range
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Business Analyst	114765.	95000	65000	137500	72500
Data Analyst	92311.	85000	65000	112500	47500
Data Engineer	132823.	112500	95000	175000	80000
Data Scientist	146310.	137500	112500	175000	62500
DBA/Database Engineer	120435.	112500	85000	156250	71250
Other	138628.	112500	65000	175000	110000
Product/Project Manager	146087.	112500	95000	175000	80000
Research Scientist	129693.	112500	75000	175000	100000
Software Engineer	140604.	112500	95000	175000	80000
Statistician	135068.	112500	85000	137500	52500

Summary statistics of salary by role

Table 2

Q2	mean	median	q1	q3	range
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Female	111591.	112500	75000	137500	62500
Male	138605.	112500	85000	175000	90000
Prefer not to say	144712.	137500	77500	212500	135000
Prefer to self-describe	187084.	143750	91875	175000	83125

Summary statistics of salary by role

Table 3

Q1	mean	median	q1	q3	range
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
18-21	84038.	75000	55000	85000	30000
22-24	82087.	75000	65000	95000	30000
25-29	105168.	95000	75000	137500	62500
30-34	126367.	112500	85000	175000	90000
35-39	146011.	137500	95000	175000	80000
40-44	150081.	137500	112500	175000	62500
45-49	161206.	137500	112500	175000	62500
50-54	153574.	137500	112500	175000	62500
55-59	166505.	137500	112500	175000	62500
60-69	157422.	137500	95000	175000	80000
70+	165431.	95000	65000	225000	160000

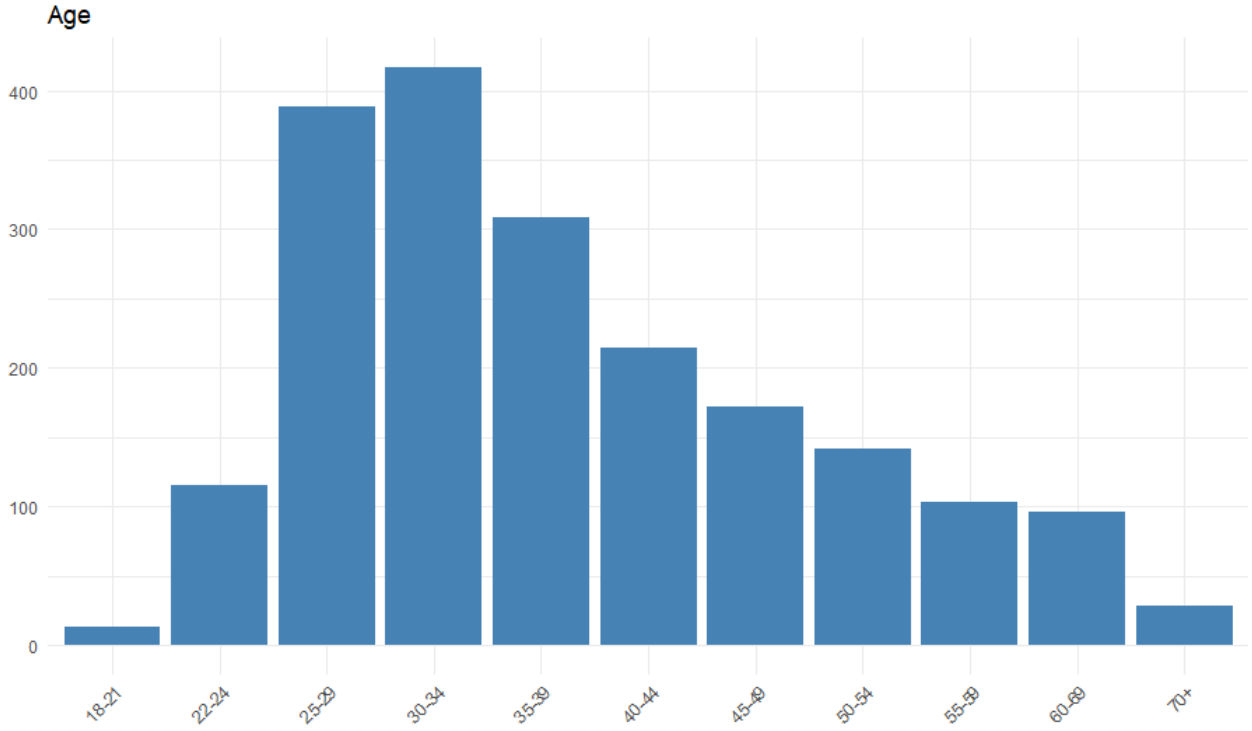
Summary statistics of salary by age

Table 4

Q4	mean	median	q1	q3	range
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
I prefer not to answer	158056.	137500	112500	175000	62500
High School	156667.	175000	128125	175000	46875
Some college	113000	90000	57500	175000	117500
Professional degree	190370.	175000	125000	225000	100000
Bachelor's degree	124649.	112500	75000	175000	100000
Master's degree	130462.	112500	85000	175000	90000
Doctoral degree	150749.	137500	95000	175000	80000

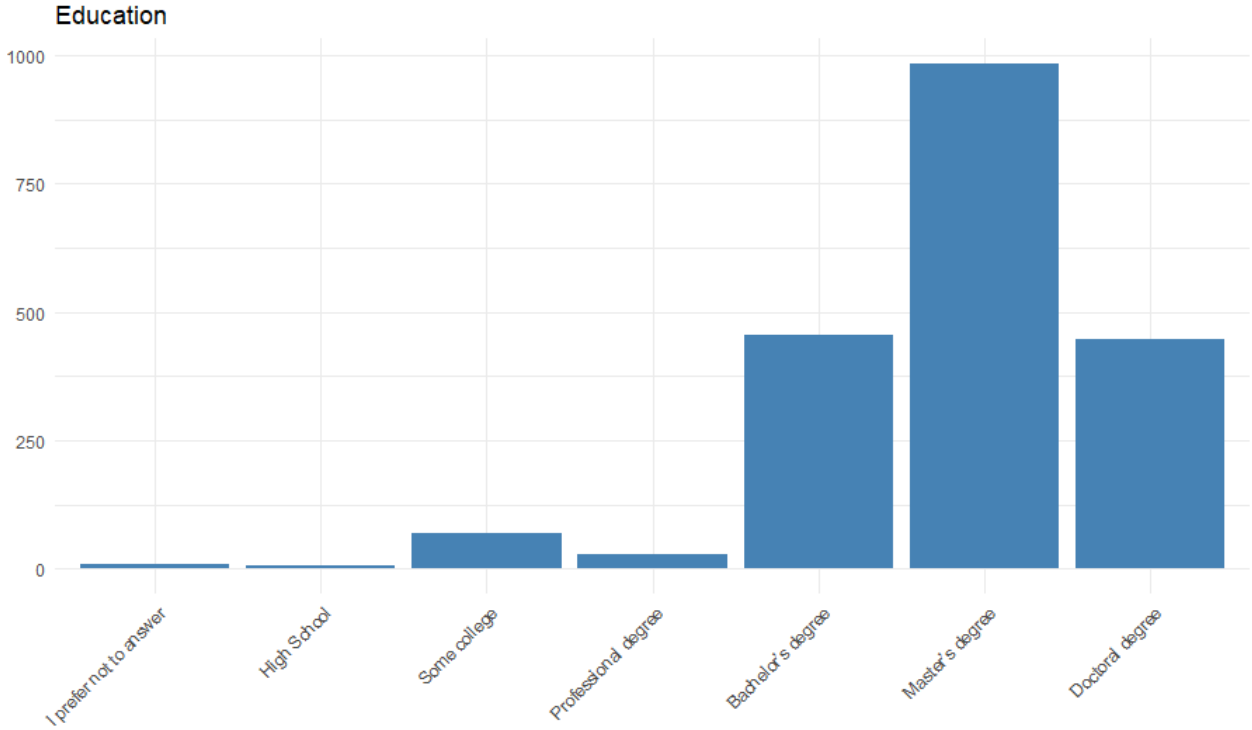
Summary statistics of salary by education

Graph 1



Frequency of respondents by age

Graph 2



Frequency of respondents by education