Analysis Of Individuals' Demographic and Technical Factors Along With Wage Potential

Author Contributions:
Matin Ghaffari: R-code, write ups and plots for: (graphical methods and analysis, analysis of missing entries and formal statistical testing), Introduction, conclusion, discussion, and contributions on data cleaning and regressions.
James Lu: R-code, write ups and plots for: (Multiple Regression, Model Transformation, Advanced Analysis), conclusion for mult. regression / model transformation / advanced analysis, python code for data cleaning and generation.

Introduction

In our study, we intend to analyze whether or not there is an association between various demographic and technical factors and yearly compensation amongst individuals in the data science and machine learning community. In our dataset (kaggle_survey_2020_responses.csv) there are 20,036 responses to an industry-wide survey that was live for 3.5 weeks in October, which asked 39 plus various data science and machine learning questions. Some questions range from basic questions such as age, region of residence, gender, work title/role, education level, and various demographics to more specific questions for more experienced individuals regarding coding languages, methodologies, frameworks, etc. In particular this survey was promoted on Kaggle.com and the Kaggle twitter page, and respondents had any window of time from 10/07/2020 to 10/30/2020 to complete the survey. Furthermore, respondents with the most experience were asked the most questions, and the participants from the 171 different countries and territories had their privacy protected, since free-form text responses were not included in the public survey dataset. Additionally, the order of the rows were shuffled to be non-chronological and countries or territories with less than 50 respondents are grouped into the nominal category of "Other" for further anonymity (2).

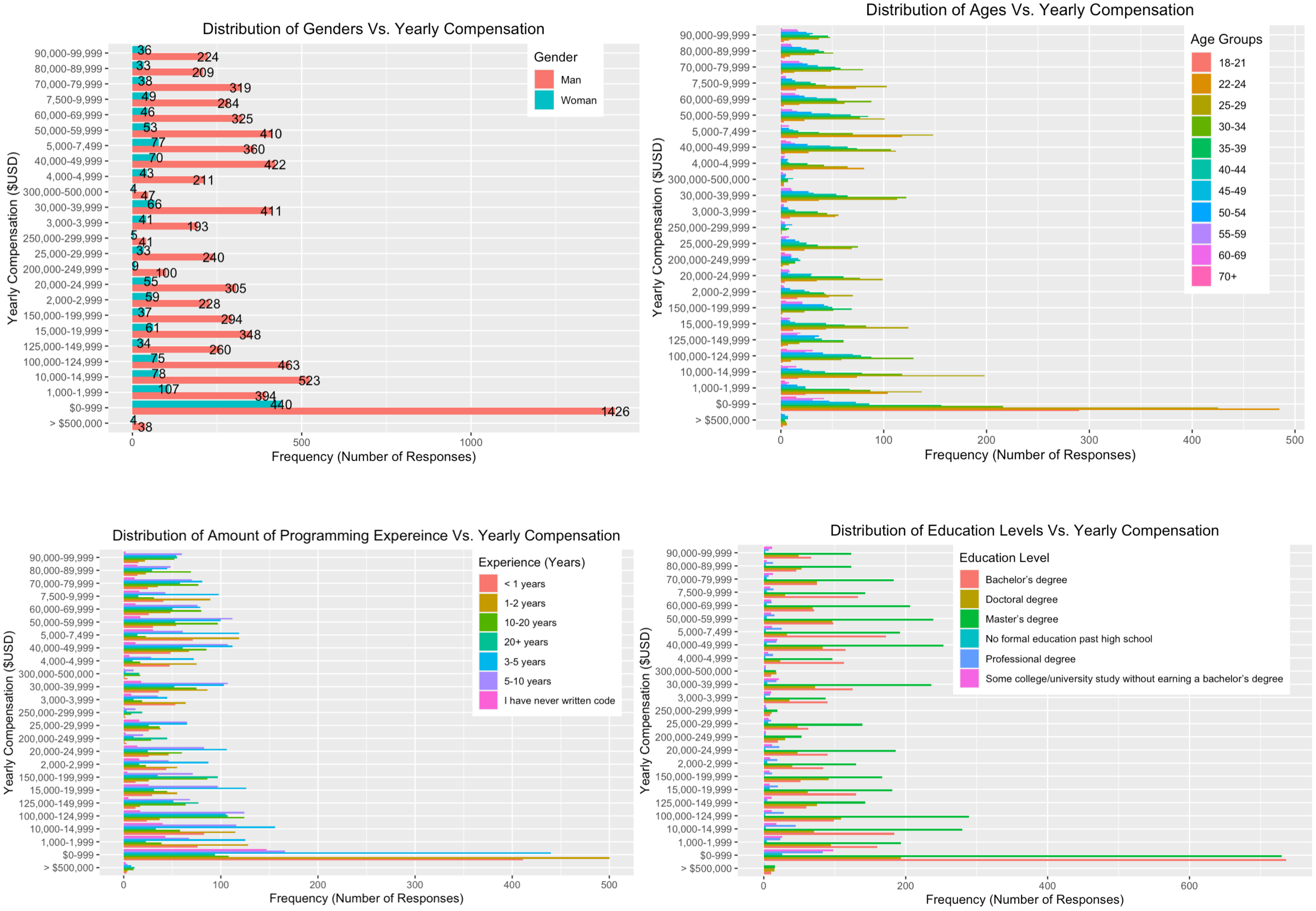
However, to answer our analysis question we used a subset of the original Kaggle dataset which includes only the relevant features to our question and with our data observations cleaned and tidied. We tidied our dataset by eliminating blank data entries and unimportant answer choices that were "Other" or "I prefer not to answer" since they are ambiguous and represented an unsubstantial distribution of the observations. Furthermore, we refined the dataset to include only the relevant features of the original dataset which are the categorical ordinal variables of "Q1" for age intervals, "Q4" for level of education, "Q6" for years of programming experience intervals, "Q22" for salary intervals. Additionally, we include the nominal categorical variables of "Q2" for gender and "Q3" for country of residence (1). However, for the purposes of our analysis methodologies, such as in our regression models, we needed to further alter these aforementioned categorical ordinal variables for the intervals by making them into continuous variables by replacing the observations with random values from the respective intervals. Subsequently our new tidied data set will have 9,628 entries.

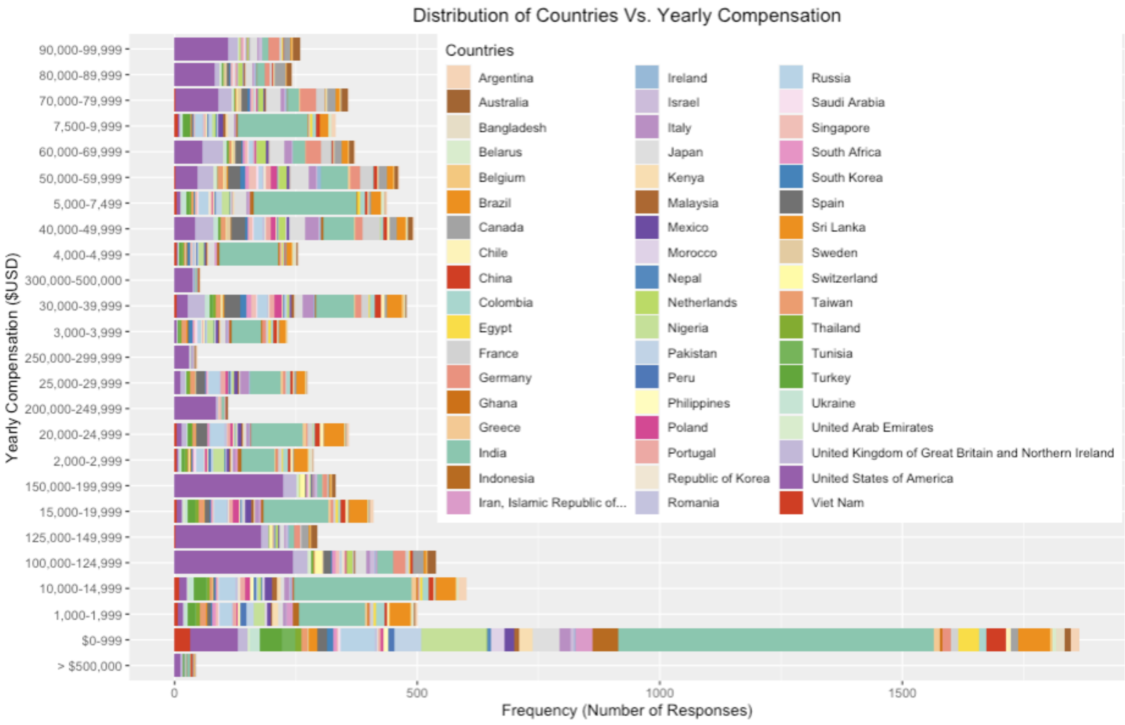
Therefore, in our analysis we will begin by first graphically analyzing the various features of gender, years programming experience, country of residence, and level of education against the amount of yearly compensation, and observe if there are any significant differences in the distributions and to identify potential biases in features. Next, we consider the missing entries that we initially filtered out for salaries in order to use formal statistical testing to help determine the potential influence of these missing points. Afterwards we explore finding possible correlations between these aforementioned variables and salary through the creation of a multivariate linear model and we then identify the characteristics that correlate most with yearly compensation to create a new improved transformed model. Ultimately, by doing this analysis we intend to provide insight on these various characteristics and if they associate or have relationships with the amount of yearly compensation of individuals which we anticipate would be useful for the purposes of data science and machine learning recruitment.

Graphical Analysis of Demographic Features vs. Yearly Compensation

We will begin our analysis by graphically analyzing the various features of gender, years of programming experience, country of residence, and level of education against the amount of yearly compensation. By doing the following, we may observe if there are many significant differences amongst the distributions of salaries by observing the frequencies of various salaries against these various features. In order to do this we use the tidied data groupings for the feature intervals and salary intervals in order to create grouped and stacked bar plots for these features using the fill color.

Figure 1: Grouped bar plots for the distribution of genders (top left), ages (top right), amount of programming experience (bottom left), education level (bottom right), and a stacked bar plot for the countries (bottom-most plot) against the amount of yearly compensation in US dollars.





We can see in Figure 1 above that there is a much greater frequency of respondents who make 0 to 999 US dollars. Additionally when observing the plots individually we see that across every salary grouping that the male gender was much more represented than female. Also when analyzing the distribution of the age groups we see that the oldest groups are the most infrequent and the most frequent age group in the distribution is green-brown for the ages 25-29 with the light green, dark green, and orange following for ages 30-34, 35-39, and 22-24. Furthermore we see that the higher green-brown and orange frequencies for the younger ages at lower salaries and especially higher at $0-$999, while the dark green and blue bars for the older age groups become more frequent at higher salaries. Also we see a similar trend with the amount of programming experience and level of education which similarly has much higher frequencies at lower salaries for respondents with either lower level of education or less years of programming experience and vice versa. Lastly we see from the stacked bar plot for the countries that India is the most frequent and is generally much more frequent than other countries most especially in the bottom 3 salaries from $0 to $14,999 where it was much more frequent than the other countries, however we see in the salaries that are $70,000 or more that US and followed by UK have the greatest frequencies, which may be explained by outsourcing and emerging markets.

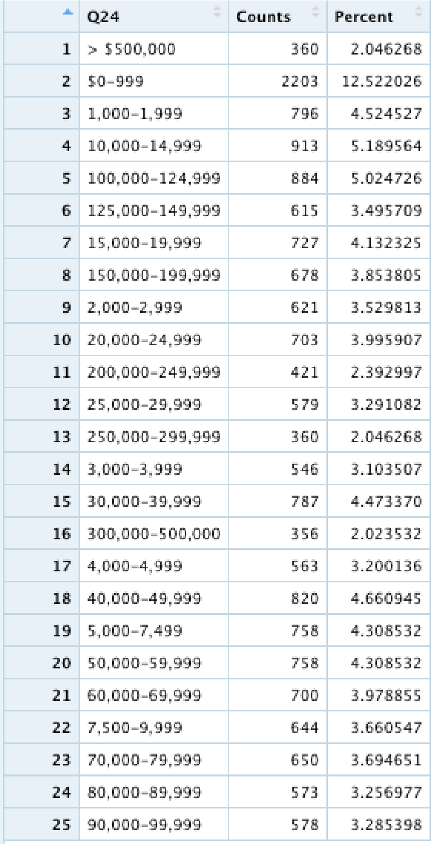
Therefore, this may suggest that various demographics of respondents may be underrepresented which we see based on our data that this is likely the case for women and older age groups who have much less observations which may make the data biased towards men or younger age groups who are instead unequally represented by much more observations (Please Refer to Appendix table 1.1 for frequency plots and count values for each of these features).

Analysis of Yearly Compensation When Considering Missing Data Entries

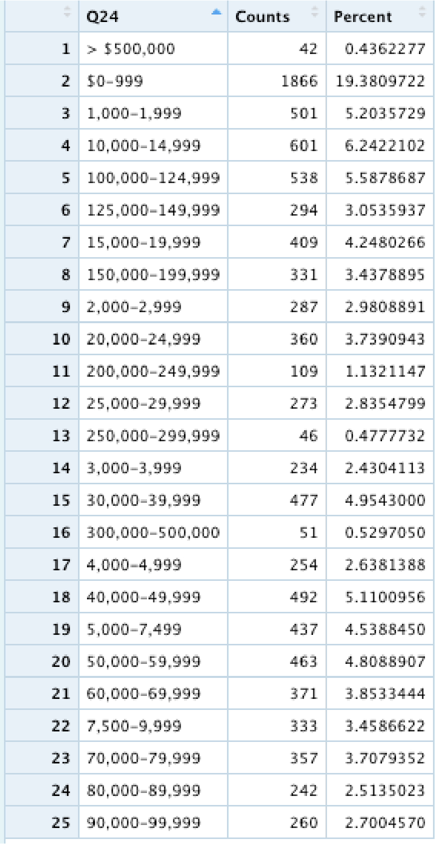
In this portion of the analysis we will explore the potential effects that the missing data entries may possibly have on the distribution of our yearly compensation. In order to accomplish this we will utilize another data set that does instead include the empty non-respondent observations for Q24, for the yearly compensation. We may then permute with random uniform values from the 25 possible categories of the various possible salary intervals. By doing this we see that the number of observations increase significantly in the altered data set from 9,628 to 17,593 entries, which we use in order to observe the new altered proportions of the data amongst the 25 categories of salary and how it differs from the tidied data set.

Table 1: Observed counts and percentages for the various intervals for yearly compensation in US dollars, with the right- hand side table being the counts and percentages without the missing entries while the graph on the left-hand side is when the missing entries are accounted for.

Adjusted Salaries: *17,593 entries* *Tidy Salaries:* *9,628 entries*



	Q24	Counts	Percent
1	> $500,000	360	2.046268
2	$0-999	2203	12.522026
3	1,000-1,999	796	4.524527
4	10,000-14,999	913	5.189564
5	100,000-124,999	884	5.024726
6	125,000-149,999	615	3.495709
7	15,000-19,999	727	4.132325
8	150,000-199,999	678	3.853805
9	2,000-2,999	621	3.529813
10	20,000-24,999	703	3.995907
11	200,000-249,999	421	2.392997
12	25,000-29,999	579	3.291082
13	250,000-299,999	360	2.046268
14	3,000-3,999	546	3.103507
15	30,000-39,999	787	4.473370
16	300,000-500,000	356	2.023532
17	4,000-4,999	563	3.200136
18	40,000-49,999	820	4.660945
19	5,000-7,499	758	4.308532
20	50,000-59,999	758	4.308532
21	60,000-69,999	700	3.978855
22	7,500-9,999	644	3.660547
23	70,000-79,999	650	3.694651
24	80,000-89,999	573	3.256977
25	90,000-99,999	578	3.285398



	Q24	Counts	Percent
1	> $500,000	42	0.4362277
2	$0-999	1866	19.3809722
3	1,000-1,999	501	5.2035729
4	10,000-14,999	601	6.2422102
5	100,000-124,999	538	5.5878687
6	125,000-149,999	294	3.0535937
7	15,000-19,999	409	4.2480266
8	150,000-199,999	331	3.4378895
9	2,000-2,999	287	2.9808891
10	20,000-24,999	360	3.7390943
11	200,000-249,999	109	1.1321147
12	25,000-29,999	273	2.8354799
13	250,000-299,999	46	0.4777732
14	3,000-3,999	234	2.4304113
15	30,000-39,999	477	4.9543000
16	300,000-500,000	51	0.5297050
17	4,000-4,999	254	2.6381388
18	40,000-49,999	492	5.1100956
19	5,000-7,499	437	4.5388450
20	50,000-59,999	463	4.8088907
21	60,000-69,999	371	3.8533444
22	7,500-9,999	333	3.4586622
23	70,000-79,999	357	3.7079352
24	80,000-89,999	242	2.5135023
25	90,000-99,999	260	2.7004570

Visually based on Table 1 we see after accounting for the missing data values with random permutations that it appears the proportions of the data points change quite significantly across all the categories. This may suggest that the feature for the amount compensation may be affected and vulnerable to influence by the non-respondent values. With this being said based on the data, it appears that the amount of compensation yearly may not be a robust feature, which is a likely result of the large number of missing data points and its effect on the proportions from the large variability introduced. Thus, this will likely affect our previous conclusions made from figure 1 in the event of missing entries, due to the difference in fit of the proportions of categories. Furthermore, we can help support this analysis with more formal statistical testing with the Chi-Squared goodness of fit test which can help us determine this with the use of hypothesis testing. Since we meet the 3 assumptions that:

1. We have observed counts → We satisfy this assumption as seen in Table 1
2. The data comprising the counts is independent → We satisfy this since we know from the study that the respondents are essentially at random since we are surveying via the internet, which allows us to assume the data comprising the counts is independent.
3. We expect to see at least 5 in each category. Thus assuming H_0, $E_i \geq 5$ for all i. → We satisfy this rule since we have a sample size of 17,593 and our rarest outcome is 2.02% for \$300,000 - \$500,000 thus 17,593 * 0.0202 > 5, therefore because our lowest expected count is ≥ 5 then we know that all expected counts are large enough to meet this assumption.

Parameter: Let p_{o1} , p_{o2} , p_{o3}, … , , p_{o25} be the percentages for the tidied salary distribution, and let p_{n1} , p_{n2} , p_{n3}, … , , p_{n25} be the percentages for the adjusted with the non-respondents' salary distributions.

Null Hypothesis (H_0): $p_{o1} = p_{n1}$, $p_{o2} = p_{n2}$,...., $p_{o3} = p_{n3}$. In other words, the given percentages adjusted with non-respondents' salary distribution are a good fit for the tidied data proportions.

Alternative Hypothesis (H_1): The given percentages adjusted with non-respondents' salary distribution are **NOT** a good fit for the tidied data proportions.

Table 2: Observed results from the chi-squared goodness-of-fit test using R (chisq.test()).

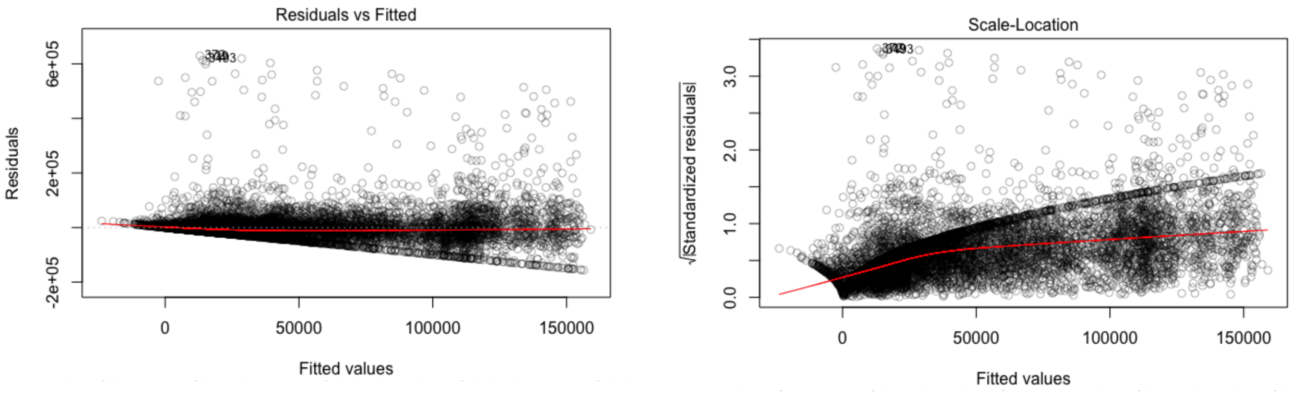
	X- Squared	***P-Value***	***Degree of Freedom***
Chi - Squared results	*3541.9*	*p-value < 2.2 E -16*	*24*

Therefore this formal statistical testing further supports this reasoning, since we have a very large X-squared which strongly suggests against the null hypothesis. And ultimately the formal testing yields a p-value less than 2.2 E -16 ≈ 0 < 0.05 (using 0.05, conventional significance level), thus the data suggests the p-value is so small that we can reject the null hypothesis in favor of the alternative hypothesis that the given percentages for the salaries with missing rows accounted for are NOT a good fit for the salary distributions without the empty observations.

Multiple Regression

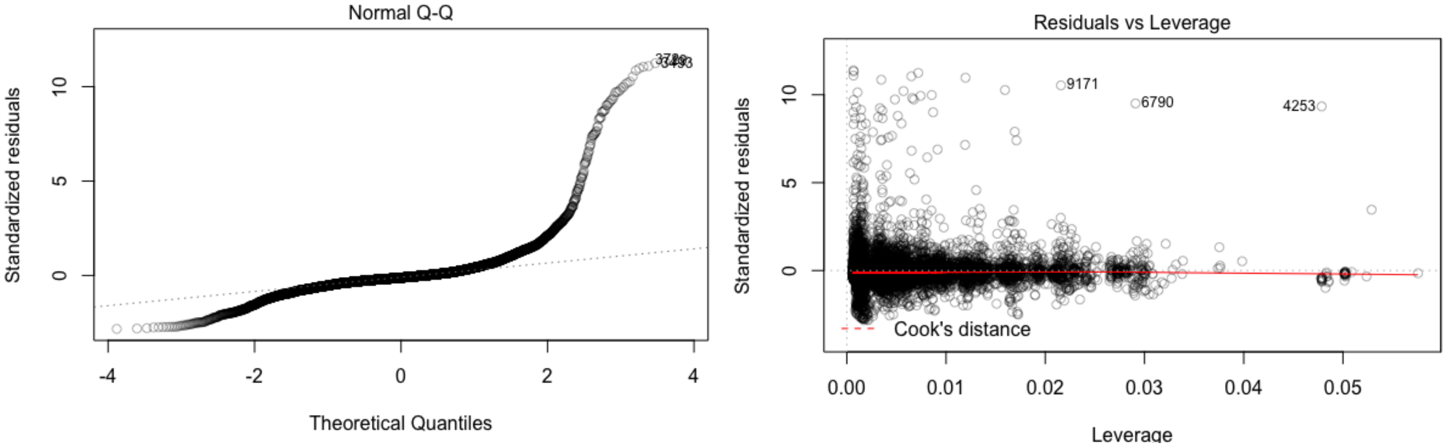
To begin our exploration on finding possible correlations between our chosen variables and wage we attempt to create a multivariate linear model. The variables that we intend to use are age, gender, country of residence, level of education, and number of years programming. One of the challenges with implementing linear regression with our chosen variables is the fact that nearly all of the survey answers were explanatory variables. To account for this we randomly choose a value in the given uniform range; we do this for the ages, years of programming experience, and salary to transform these explanatory variables into continuous. Although even after doing this we still had gender, country of origin, and level of education as categorical variables which we chose to represent as integers using the *factor()* function in *R*.

Figure 2: Residual vs. Leverage plot and Scale-Location plot for our multivariate linear model



After creating our linear model with our chosen variables, we can see in *Figure 1* that for both our residuals vs. fitted plot and our scale-location plot that in its current state our linear model does not meet the basic assumptions of a usable linear model which is normal residuals, constant variability, residual independence, and each independent variable is linearly related to the response variable.

Figure 3: Normal QQ plot and Residuals vs. Leverage from our multivariate model



From Figure 3 we can see that our residuals are not normal and our residuals vs. leverage indicate that there were some outliers with high influence in our data. Thus supporting what was mentioned previously, and ultimately our current model cannot be used as a reliable linear model, so a transformation of our model is needed.

Model Transformation And Selection

Our initial multivariate linear model was determined to be unfit to model wage potential and in this section we want to find which variables may need to be removed, as well as which variables have the best linear relationship with wage. To do this we analyze our summary statistics and p-values from our model and plot each suitable variable against wage to help determine which variables are best to include in our new model.

Table 1: Some examples of country variables and their corresponding p-values

Variable	Coefficient	P-Value
Australia	70875.89	< 2e-16
Bangladesh	-4755.22	0.648773
Belarus	3136.92	0.778256
Belgium	38941.03	0.000539
Brazil	6342.95	0.348716
Canada	57635.43	6.39e-15
….	….	….

After analyzing our coefficients and summary statistics from our linear model we found that representing each country in our linear model introduced a significant amount of noise as the inclusion of this variable introduced 54 independent variables for each country. Examining the values in Table 1 which demonstrates just how large the range of p-values are between countries. Our first step to modifying our model is to remove the countries variable from our model.

Table 2: anova() on the initial model against new model without countries

	Res. Df	RSS	DF	Sum of Sq	F	P(> F)
1	9566	2.9273e+13	-	-	-	-
2	9619	3.9537e+13	-53	-1.0264e+13	63.282	< 2.2e-16

After removing the countries variable from our model we found that this new model did have a better fit than our initial model shown by the p-value in Table 2, however the new model still did not match the assumptions for a usable linear model. To continue to simplify our model we chose to remove the gender variable and instead choose between associating average age + level of education to wage, or the average number of years of coding experience + level of education to wage.

Figure 4: Average Age vs Average Wage grouped by education level

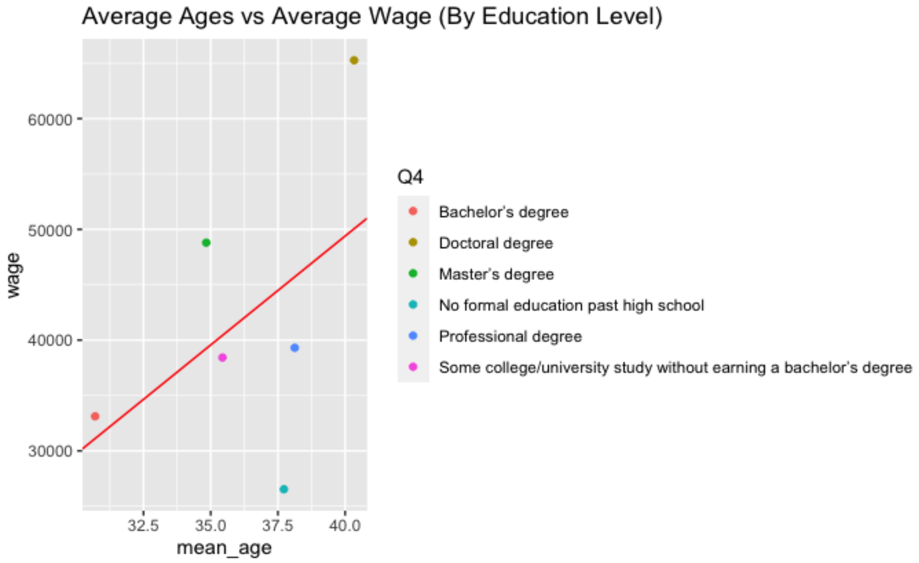


Figure 5: Average Coding Experience vs Average Wage grouped by education level

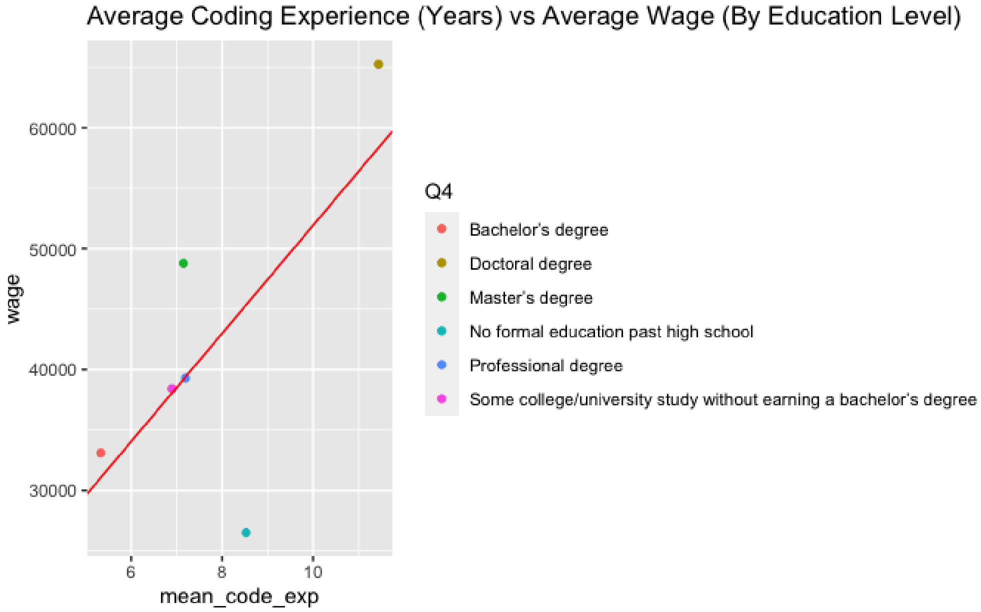


Table 3: R ^{2} values from both regression models

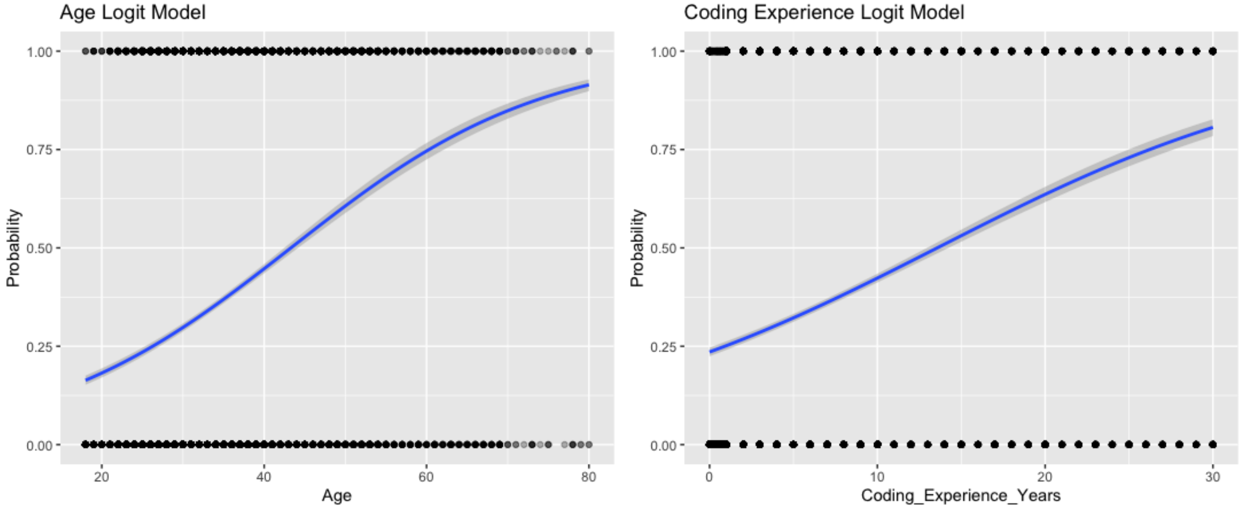
Regression Model	R^2
Average Age vs Average Wage	0.233
Average Coding Experience vs Average Wage	0.4619

After generating our two models shown by Figure 4 and Figure 5, we found that on average the youngest data scientists were those with bachelor's degrees while the oldest were those with doctorate degrees. From Figure 4 we also saw that an individual with a bachelor's degree can be expected to make more than someone with no formal education past highschool, although interestingly enough those with some college experience tended to be older but also had a higher wage than those with bachelor degrees. Another interesting discovery we had was that those with some college experience made roughly the same amount as those who achieved a professional degree. It is difficult to make any conclusions with these findings as there are numerous confounding socioeconomic variables that can affect the relationship. From Figure 5 we can see that those with no formal education past highschool had on average the second most years of experience with coding but by far made the least amount of money. We can also see that generally with more years of experience coding and higher the level of education we can expect a higher wage on average. The idea behind creating these regression lines was to see if we could find a correlation between average age (or average coding experience) and average wage based on level of education and attempt to predict future wages, however after examining the plots and the R^2 values for each model we found that neither model was a good fit for the data. Comparing the R^2 values of average age vs average wage (0.233) and average coding experience vs average wage (0.4619) it is clear that coding experience has a better fit for the data and thus would be a better predictor for wage, however the fit is still too poor to be a reliable predictor for wage. Overall we could find some association between age and education level vs wage and coding experience and education level, but we could not generate a regression line that could reliably predict wage from our chosen independent variables.

Advanced Analysis

To further explore the idea of predicting wage with age and years of programming experience we attempt to transform wage into a binary variable by adjusting the wage in such a way that all values greater than or equal to 40,000 have value 1 and all those less than 40,000 have value 0. After doing this we intend to attempt logistic regression on the new response variable to see if we might be able to generate a better model to predict wage.

Figure 6: Logistic regression performed on wage, where wage is binarized in the following way: 0 if wage < 40,000, 1 if wage >= 40,000. Left: Age is the independent variable, Right: Years of



As we can see in Figure 6, generally speaking individuals that are older data scientists are more likely to make over 40,000 USD. Similarly, the more coding experience that an individual has the more likely they are to make over 40,000 USD. In the age model we can see that there is an abundance of data for ages 18 to ~70 and the S shape of the curve is fairly flat which might indicate that the growth of probability against age might be more gradual than the traditional S curve. In the coding experience model we can see that there is a small cluster of data from years of coding experience 0 to ~2. This cluster could be attributed to the recent growth of data science as a discipline and career choice so more individuals are just beginning to learn programming and coding. As for the shape of the curve in the coding experience model we can see that the shape is even flatter than the age model and could also suggest that the probability of making >= 40,000 USD increases gradually as the number of years of coding experience increases.

Table 4: Null deviance and residual deviance of each model including the degrees of freedom for each one as well

Model	Null Deviance, DF	Residual Deviance, DF
Age	12724, 9627	11651, 9627
Coding Experience	12724, 9626	11734, 9626

As a way to measure how appropriate our model is we can observe the residual deviance in Table 4, and we see that for both models the residual deviance is relatively close to the degrees of freedom which suggests that our model has been trained properly. To test the assumption that our model was trained properly and to test the accuracy of our model we intend to conduct point based estimates and compare them to our observed values.

Table 5: Estimates and observed values for the probability of making >= 40,000 for various ages

Age	Estimate	Observed
20	0.1819433	0.001454092
40	0.447089	0.2183216
60	0.746183	0.3560449

Table 6: Estimates and observed values for the probability of making >= 40,000 for various years of coding experience

Coding Experience (Years)	Estimate	Observed
2	0.2681474	0.06792688
12	0.4660025	0.243872
22	0.6751655	0.3202119

After comparing our point based estimates with their observed values (shown in Table 5 and Table 6) we found that all of our estimates from various regions of the distribution provided inaccurate estimates when compared to their observed values. To conclude the logistic regression model for both age and years of experience programming do not generate reliable or accurate estimations for our data, this could be because the relationship between age / coding experience and wage is sporadic and has high variance or other confounding factors such as differing costs of living around the world / currency conversion , etc.

Conclusion

Therefore, from our findings that we explored from these survey results from kaggle.com, to help us determine if there is an association between various demographic and technical factors and yearly compensation amongst individuals in the data science and machine learning community. In the beginning of our analysis, we first graphically analyzing the various features of gender, years programming experience, country of residence, and level of education against the amount of yearly compensation, and we observed that generally the younger ages, those with less years of programming experience, and those with lower levels of education often had lower salaries and vice-versa for higher salaries. Furthermore, we saw that there were much less women across all salary categories which may suggest that the data is unrepresentative and potentially biased for information in regards to women which can also be said for the older age groups which also appear under-represented. Lastly we plotted the distributions of various countries which appeared to show the trend of increased lower salaries in emerging markets such as India and the higher salaries were more distributed in western countries such as the USA or UK which may be explained by outsourcing or other potential economic factors.

Secondly, we then further analyze the yearly compensation categorical variable in order to explore the potential effects that the missing data entries may possibly have on the distribution of our yearly compensation. We see after performing basic and formal statistical tests that in the case of missing data values with random permutations that the proportions of the data points appear change quite significantly across all the categories. This may suggest that the feature for the amount compensation may be affected and vulnerable to influence by the non-respondent values.

For the third part of our analysis we attempted to generate a multivariate linear regression model based on age, gender, country of residence, level of education, and number of years programming to predict wages for data scientists. After generating our model we conducted basic tests to see if the assumptions for multivariate regression lines were met. We found that although the variance was relatively constant, the residuals were not normal and because of that we could not use the regression model as it was.

After creating our initial regression model we attempted to find the most influential variables so that we could potentially transform the model. We found that some of the variables we included such as countries and gender weren't suitable for this specific application of regression, so we instead simplified our model into basic linear regression models featuring the averages of age by education level, and averages of coding experience by education level to see if any associations and predictions could be found. Ultimately we came to the conclusion that our simplified models were not a good fit for the data and could not accurately or reliably predict future wages given age / experience / level of education. Perhaps a better model could have been created by using groupings of categorical / explanatory variables to account for all the noise in the data.

For our advanced analysis we approached regression again by attempting to binarize our response variable (wage) and perform logistic regression to predict the probability of earning >= 40,000 USD a year. We began by creating two logistic models for both age and years of experience coding. After generating our models we found that our curve was very flat and increased gradually. After examining the residual deviance and comparing it to the degrees of freedom we assumed that our model had an acceptable fit. We then performed some point based estimates using both models however we found that our models were extremely inaccurate and unreliable, this could have been caused by a number of confounding variables such as differing costs of living around the world which may impact salary, currency conversion, etc. Further exploratory analysis can be done and different methods of regression might create a better model that can more accurately predict wages from this dataset.

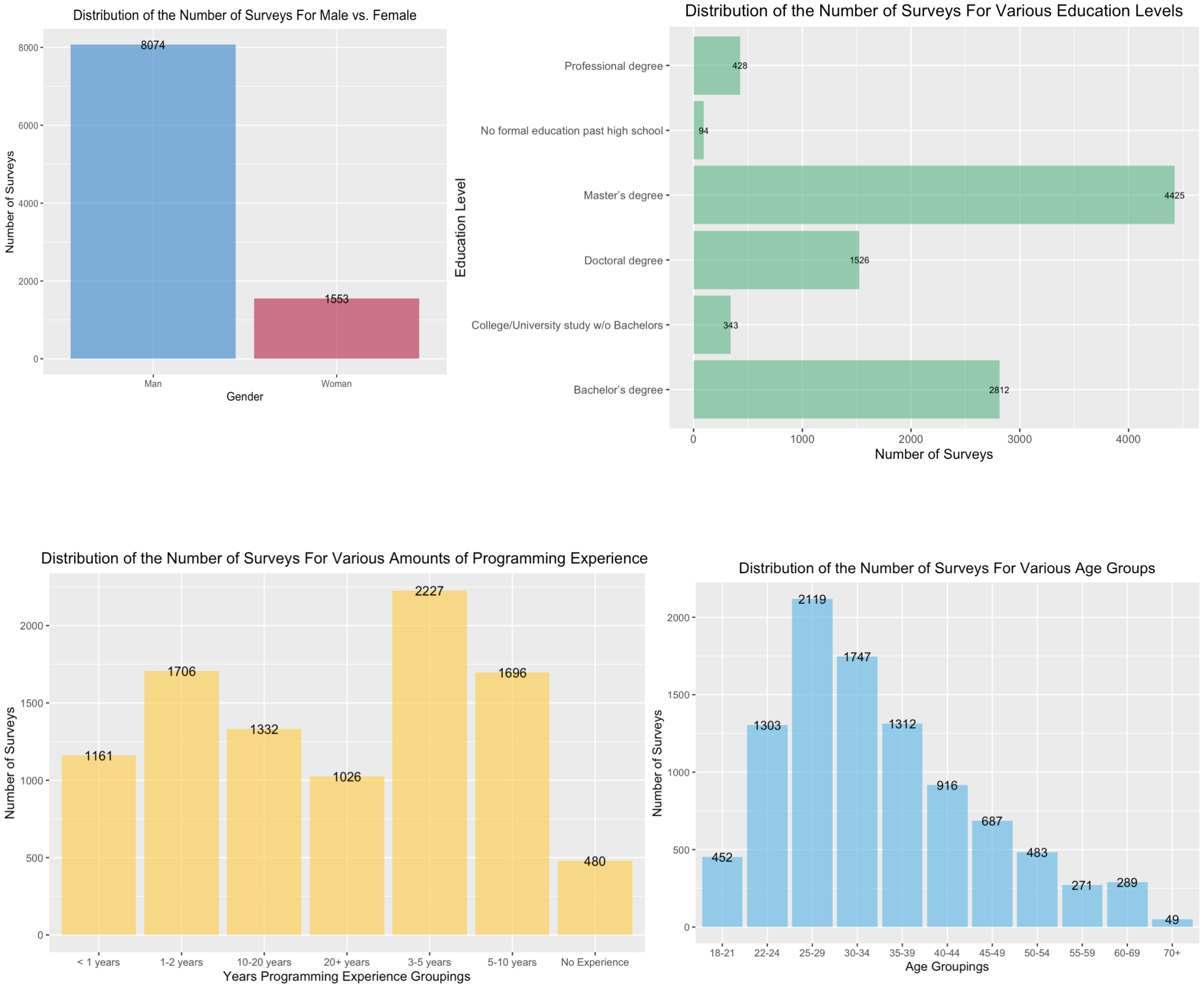
In further works we would like to build a better regression model for us to make predictions with given the following features by more adequately accounting for the large amount of noise within the data. Furthermore, another application of this analysis that we would be interested in further building our analysis for would be to see if the following data helps determine any relationships or provide any useful insights regarding the gender and wage gap. However, this leads into the limitations of this data since as we saw earlier and can be seen in Appendix Table 1 is that some of the categories for these features may be under-represented since for instance we see more than 4 times more men respondents than female and substantially less observations for older age groups which may introduce potential bias in their interpretations. Another limitation of this data is that the survey was only advertised on kaggle.com and the Kaggle twitter page, which can possible introduce dependencies amongst the data since it is possible that a large portion of kaggle users share certain characteristics and relationships that may be unrepresentative of certain features that we are analyzing.

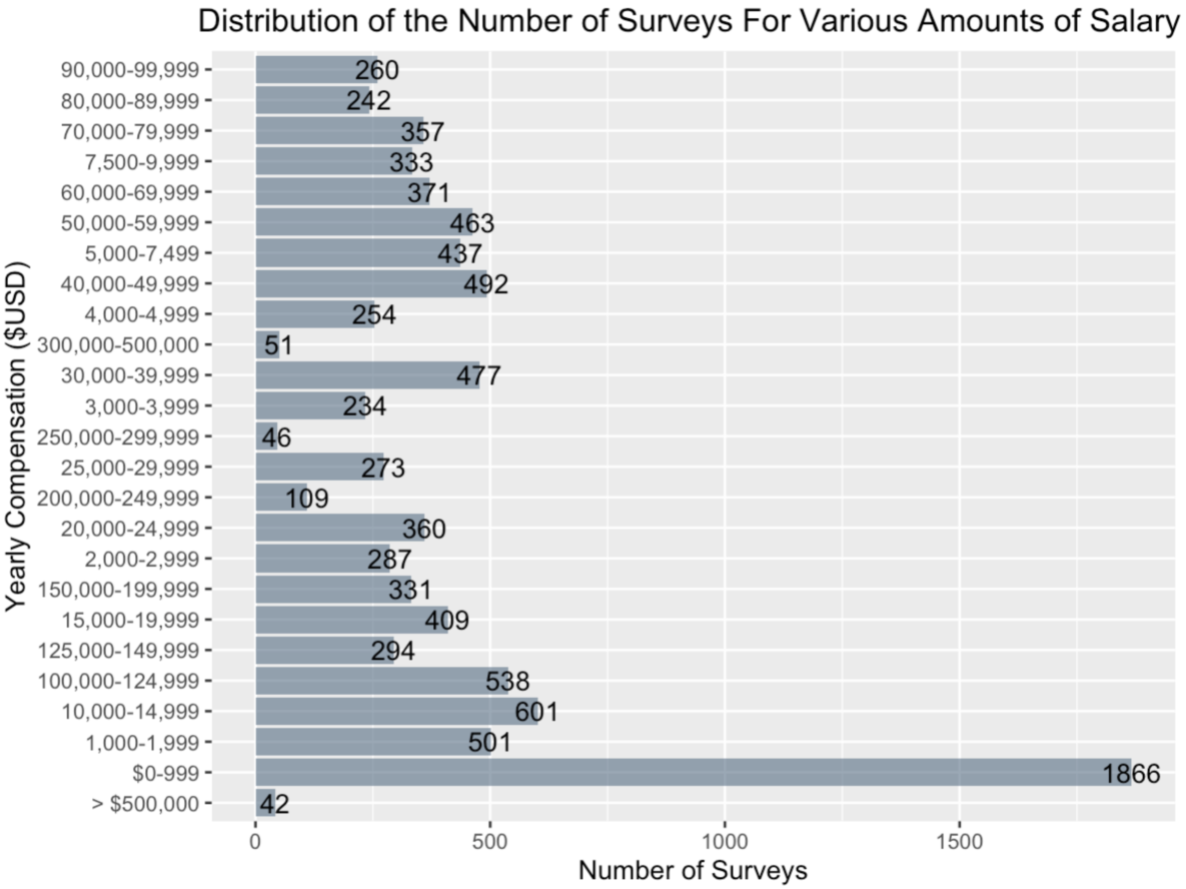
Works Cited

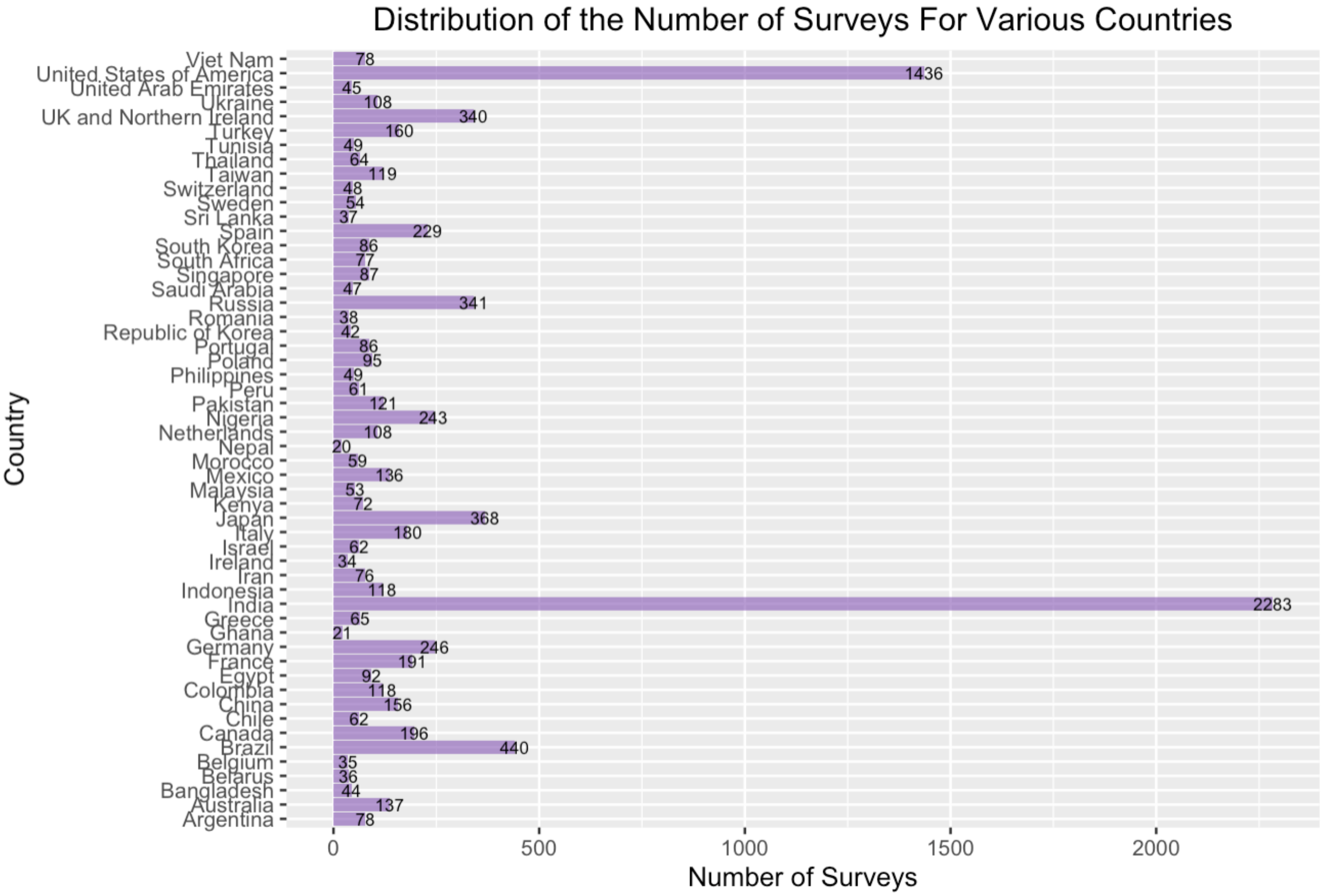
1) *2020 Kaggle Machine Learning & Data Science Survey*, www.kaggle.com/c/kaggle-survey-2020/overview.
2) "2020 Kaggle DS & ML Survey Methodology and Survey Flow Logic." Kaggle.com.

Appendix

Table 1.1: Frequency plots and count values for each of the features used in our graphical analysis in the first section of the analysis, which are for the following: gender, education level, amount of programming experience, age, yearly compensation, and country of residence.




Distribution of the Number of Surveys For Various Amounts of Salary
Yearly Compensation ($USD)
90,000-99,999
260
80,000-89,999
242
70,000-79,999
357
7,500-9,999
333
60,000-69,999
371
50,000-59,999
463
5,000-7,499
437
40,000-49,999
492
4,000-4,999
254
300,000-500,000
51
30,000-39,999
477
3,000-3,999
234
250,000-299,999
46
25,000-29,999
273
200,000-249,999
109
20,000-24,999
360
2,000-2,999
287
150,000-199,999
331
15,000-19,999
409
125,000-149,999
294
100,000-124,999
538
10,000-14,999
601
1,000-1,999
501
$0-999
1866
> $500,000
42
0
500
1000
1500
Number of Surveys


Distribution of the Number of Surveys For Various Countries
Country
Viet Nam
78
United States of America
1436
United Arab Emirates
45
Ukraine
108
UK and Northern Ireland
340
Turkey
160
Tunisia
49
Thailand
64
Taiwan
119
Switzerland
48
Sweden
54
Sri Lanka
37
Spain
229
South Korea
86
South Africa
77
Singapore
87
Saudi Arabia
47
Russia
341
Romania
38
Republic of Korea
42
Portugal
86
Poland
95
Philippines
49
Peru
61
Pakistan
121
Nigeria
243
Netherlands
108
Nepal
20
Morocco
59
Mexico
136
Malaysia
53
Kenya
72
Japan
368
Italy
180
Israel
62
Ireland
34
Iran
76
Indonesia
118
India
2283
Greece
65
Ghana
21
Germany
246
France
191
Egypt
92
Colombia
118
China
156
Chile
62
Canada
196
Brazil
440
Belgium
35
Belarus
36
Bangladesh
44
Australia
137
Argentina
78
0
500
1000
1500
2000
Number of Surveys