Report on Indeed Dataset

The Analysis of Influential Factors of Salary in the U.S.

Ingrid Wijaya, YueLong Zhang, Dong Hyun Chun, Arman Bazak

1 Introduction

Salary is one of the most important factor that candidates considers when they are seeking for a job, and it is much related to job satisfaction. According to the survey conducted by the Society for Human Resource Ma-gement, since 2002, salary has been one of the top five contributors to job satisfaction. More specifically, 63% of the employees view salary as an important aspect of job satisfaction [1].

The motivation behind this report is to find the factors that influences salary in the United States. Furthermore, for the population that places greater importance on higher salary, how can they aim to earn more? With that being said, we will answer the following research questions:

- What are the factors that influences salary?
- Which candidates are offered a higher salary?

In the following section, we discuss the dataset and the data cleaning that we performed.

2 Data

The data set that we chose to use is the Datafest 2018 Indeed data set which contains the data on the job listings posted. The data set has 14586035 data points and 23 variables. A description of each variables is provided below.

Variables	Definition
date	Unixtime date when events occured.
companyId	-
jobId	-
country	Country of job-posting.
stateProvince	-me of the state or province of the job posting.
city	-me of job posting city.
avgOverallCompanyRating	Average rating of the company (1-5 stars), with 0s for non-rated companies
numOfRatings	Total number of reviews the company had.
industry	Industry associated with the company.
normTitle	The normalized / canonical job title.
normTitleCategory	The category (similar to occupatio-l category) of the norm title
descriptionCharacterLength	Number of characters in job description.
descriptionWordCount	Number of words in job description.
experienceRequired	Minimum experience required for the job in years.
estimatedSalary	Estimated annual salary (0 when estimate not available)
salaryCurrency	Salary currency.
jobLanguage	Job language.
supervisingJob	Whether this job is classified as a supervising job
licenseRequiredJob	Whether this job is classified as requiring a license
educationRequirement	The jobś education requirement.
jobAgeDays	Age of job in days, based on job create date and on central time-zone.
clicks	The total number of clicks on the job on the date.
localClicks	The total number of clicks on the job from a local user (same city and country) on the date. Resets when 'refreshed'.

Table 1: Description of each variables in Indeed data set

A brief summary statistics that includes the number of unique values, data type of each variable, mode, percentage of data points missing, variance, mean and median is provided below. It is to be noted that non-applicable values such as variance, mean and median of categorical variables are indicated by "-".

Variables	Number of unique values	Data type	% of missing values	Mode	Variance	Mean	Median
companyId	150993	string	0	[company02436]	-	-	-
jobId	520434	string	0	[job0210510, job0246932]	-	-	-
country	3	string	0	[US]	-	-	-
stateProvince	87	string	0	[CA]	-	-	-
city	17747	string	2.32	[New York]	-	-	-
avgOverallRating	42	float64	0	[0.0]	3.55	1.71	0
numReviews	630	float64	56.55	[53.0]	635866.41	187.28	69
industry	186	string	89.43	[HEALTH_CARE]	-	-	-
normTitle	6391	string	8.93	[retail sales associate]	-	-	-
normTitleCategory	57	string	8.93	[ma-gement]	-	-	-
descriptionCharacterLength	11075	int64	0	[1366.0]	2909253.32	2367.60	2005
descriptionWordCount	2242	int64	0	[280.0]	59761.60	383.84	333
experienceRequired	42	float64	64.24	[2.0]	8.95	3.29	2
estimatedSalary	2060	int64	0	[0.0]	801900218.02	43668.53	33500
salaryCurrency	3	string	85.04	[USD]	-	-	-
jobLanguage	3	string	0	[EN]	-	-	-
supervisingJob	2	float64	13.98	[0.0]	0.13	0.16	0
licenseRequiredJob	2	float64	13.98	[0.0]	0.23	0.36	0
educationRequirements	3	string	13.98	[None]	-	-	-
jobAgeDays	102	int64	0	[0.0]	666.41	30.78	23
clicks	1827	int64	0	[26.0]	1265.54	22.44	18
localClicks	764	int64	0	[0.0]	109.82	3.71	1

Table 2: Summary statistics of variables before data cleaning

Referring to table 2, we noticed that the data set contains job listings taken from three different countries, that are, the United States, Ca-da, and Germany. We decided to focus our analysis to within the United States, so we removed the all the job listings that are not in the U.S. from the data set. We then handle - values in the data set. For numerical variables, we replaced - values with its mean. For categorical variables, we replaced - values with its mode. With that being said, the final data set contains the job listings in the U.S and it leaves us with 11547904 data points.

The summary statistics for each variables, that includes the number of unique values, data type of each variable, mode, variance, mean and median is provided below. Once again, it is to be noted that non-applicable values such as variance, mean and median of categorical variables are indicated by "-".

Variables	Number of unique values	Data type	Mode	Variance	Mean	Median
companyId	109002	string	[company02436]	-	-	-
jobId	413028	string	[job0210510, job0246932]	-	-	-
country	1	string	[US]	-	-	-
stateProvince	58	string	[CA]	-	-	-
city	11687	string	[New York]	-	-	-
avgOverallRating	41	float64	[0.0]	3.55	1.91	2.70
numReviews	626	float64	[70.0]	351675.20	134.16	70
industry	175	string	[HEALTH_CARE]	-	-	-
normTitle	5099	string	[retail sales associate]	-	-	-
normTitleCategory	57	string	[ma-gement]	-	-	-
descriptionCharacterLength	10944	int64	[1389.0]	3195489.61	2471.67	2111
descriptionWordCount	2219	int64	[197.0]	65146.96	399.56	349
experienceRequired	42	float64	[2.0]	4.10	2.52	2
estimatedSalary	2060	int64	[22100.0]	857109013.53	46836.26	36600
salaryCurrency	1	string	[USD]	-	-	-
jobLanguage	3	string	[EN]	-	-	-
supervisingJob	2	float64	[0.0]	0.13	0.16	0
licenseRequiredJob	2	float64	[0.0]	0.23	0.36	0
educationRequirements	3	string	[None]	-	-	-
jobAgeDays	102	int64	[0.0]	672.02	30.98	23
clicks	1731	int64	[26.0]	1304.54	22.81	18
localClicks	691	int64	[0.0]	97.58	3.66	1

Table 3: Summary statistics of each variables after data cleaning

Referring to Table 3, we observed a high variance of 857109013.53 for the estimated salaries in the job listing. This peaks our interest to learn more about which variables has a big influence on the salary in the job listing, which we will further explore in the exploratory data analysis in the next section.

3 Exploratory Data Analysis

In the EDA, we would like to explore the method of analysis that would be appropriate to predict estimated salary. Hence, we will take a look at the correlation and the normality of the numerical variables.

3.1 Correlation

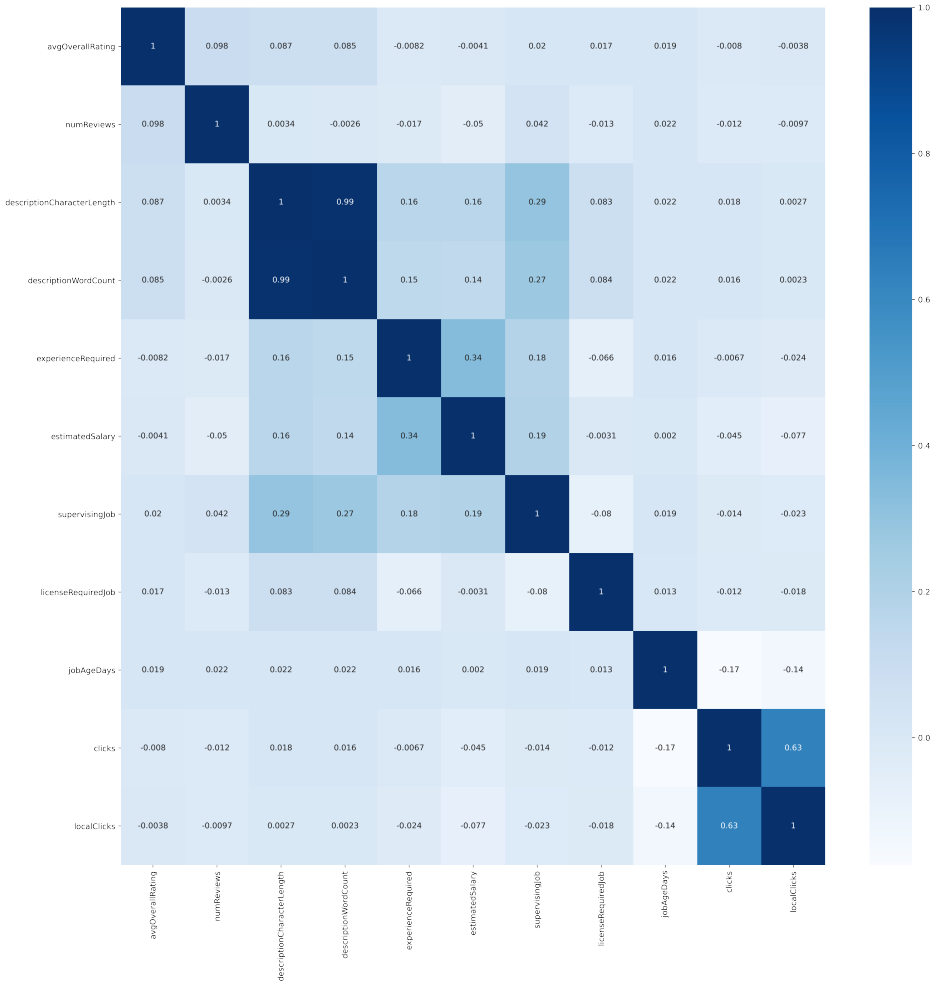

Figure 1: Correlation matrix for numerical variables

Referring to Figure 1, upon looking at the correlation between variables, there are no meaningful or interesting relationships found. More specifically, estimated salaries has very low correlation with all the other variables. We would like to use a more powerful type of analysis method to predict estimated salary, especially since this data set has a lot of categorical variables.

3.2 Visualization of numerical variables

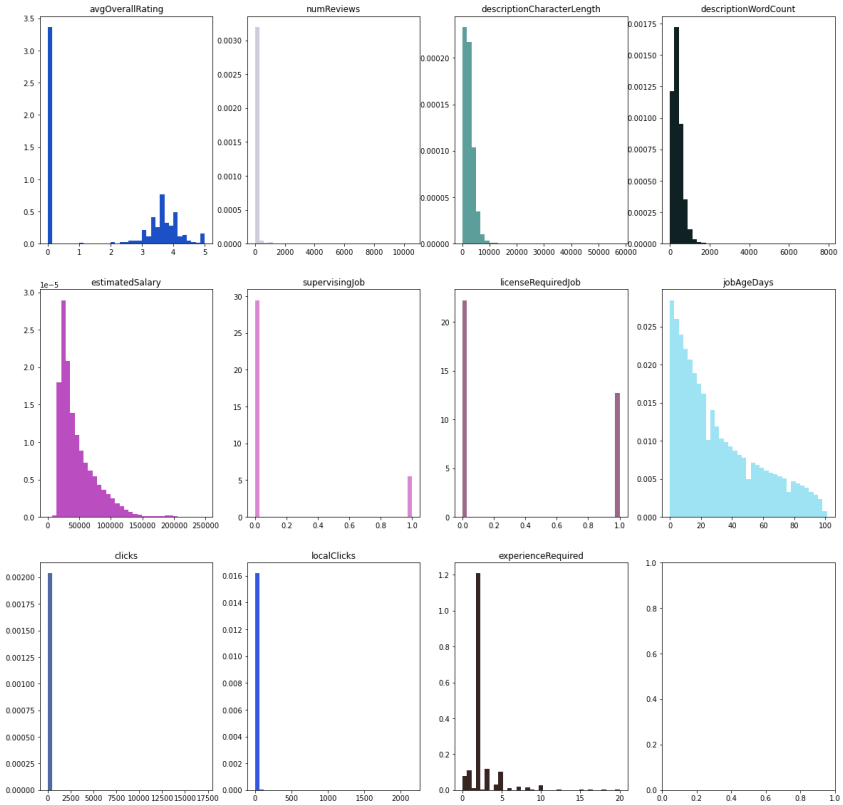

Figure 2: Histogram of variables

From figure 2, we can observe the density of the different numerical variables, to check for normality for future analysis. In the next section, we will discuss the method for our analysis.

4 Method

We use feature selection methods to find the important features relating the mean estimated salary, and then use statistical methods to analyze those variables.

4.1 Feature selection method

Given that the dataset contains 10 million data points and we have limited computatio-l resources, we only have a limited number of model for data mining. Since R^2 scores for regression and decision tree are low and those models are relatively simple, we finally use random forest model. After training the random forest, we analyze the important features given by the random forest.

4.2 Variable analysis method

Using the result from the feature selection method, we will choose the top three predictors of estimated salary and analyze the predictors further. With accordance to each predictors, we will perform an appropriate combiation of the analysis listed:

- Visualise the estimated salary of job listings between different categories using:
  - Box Plot
- Testing the equality in mean estimated salary of job listings between different categories using:
  - ANOVA

- Testing the equality in mean estimated salary of job listings between different pairs categories using:
  - Post Hoc Analysis using Tukey's method
- Testing the equality in variances of estimated salary of job listings between different categories using:
  - Levene's test

5 Results

5.1 Random Forest feature selection results

We split the data set into train and test data set, and train the model on train data set and check the R^2 on test data set. The R^2 is high, but MSE is still high. We believe the reason for this is that the variability of estimated salary is high. Even though we could explain most of variability in the test data set, the remaining variability still gives us huge uncertainty in prediction. Therefore, we could not use this model for prediction of mean estimated salary. However, since random forest gives us the impurity-based importance of each feature, we could use statistical methods to analyze those important features as shown in the table below.

Predictors	Importance Feature
normTitleCategory	0.2805
experienceRequired	0.1930
educationRequirements	0.1339
normTitle	0.1068
companyId	0.0537
descriptionCharacterLength	0.0414
jobId	0.0359
city	0.0322
descriptionWordCount	0.0303
stateProvince	0.0236
numReviews	0.0216
avgOverallRating	0.0154
supervisingJob	0.0109
licenseRequiredJob	0.0069
month	0.0063
industry	0.0037
jobAgeDays	0.0022
clicks	0.0008
localClicks	0.0005
year	0.0003
jobLanguage	0.0001

Table 4: Result from random forest model: Importance feature of each predictors

From Table 4, we can conclude that Job (Occupational) category, minimum experience required for the job, and the job's education requirement are three most important predictors of the estimated annual salary.

5.2 Variable analysis results

5.2.1 Job Category

Job categories is the leading important factor influencing the salary. In order to visualise of the estimated salary in each job category, we used a box plot.

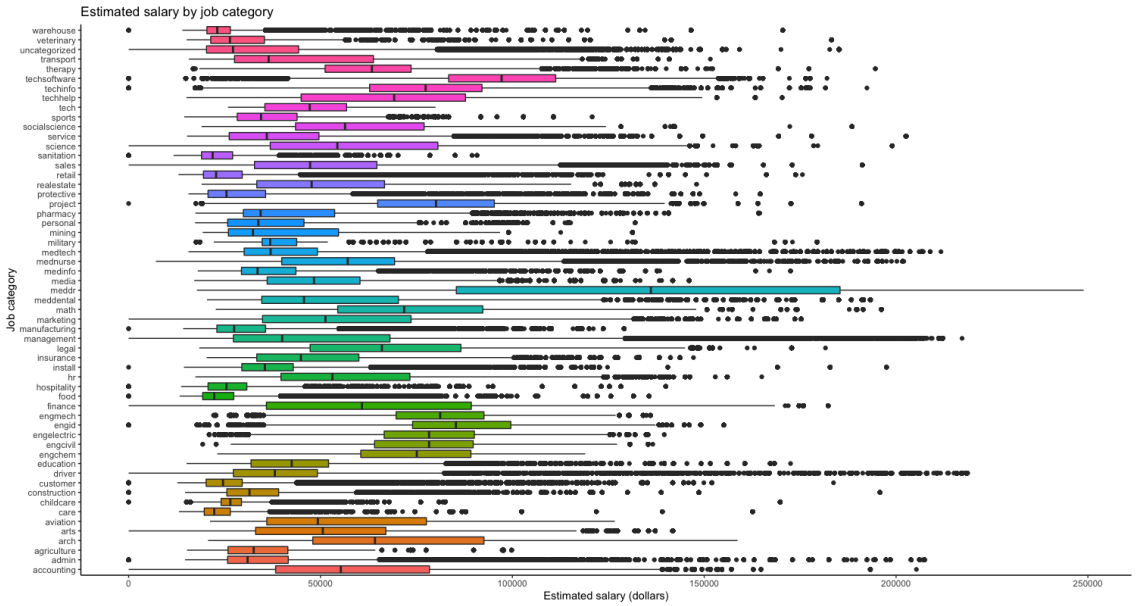

Figure 3: Boxplot of Estimated Salary relative to Job categories

Referring to Figure 3, there are significant differences of the mean and variability of the estimated salary. We performed ANOVA to test the hypothesis that there is a difference in mean of estimated salary in each job category, and obtained the following results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	$p < 0.05$
Job Category	56	4.358×10^{15}	7.782×10^{13}	214060	$< 2 \times 10^{-16}$	Yes
Residuals	11547847	5.540×10^{15}	4.798×10^{8}			

Table 5: ANOVA analysis for job category

Referring to Table 5, the test result is statistically significant at $\alpha = 0.05$ level. Hence, there is a difference in means of estimated salary in each job category. We also sort the job categories by the mean estimated salary, and the leading industries are medical, tech software, engineering, and math. This is probably because the demand for jobs in these industry are very high. Therefore, we conclude that STEM fields have highest mean estimated salary.

5.2.2 Minimum Experienced Required

The minimum experience required for the job is the second most important variable in determining salary. In order to get a clear visualisation of the amount of experience required related to the estimated salary, the plot below is spread into 4 distinct ranges from 0-5, 5-10 , 10-15, and 15-20 years of experience required.

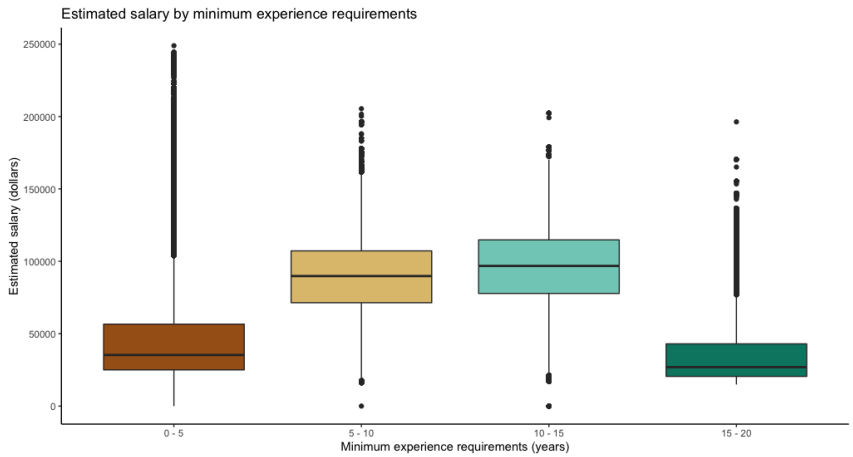

Figure 4: Boxplot of Estimated Salary relative to Experience Required

Referring to the boxplot in Figure 4, as the years of experience required increases, the estimated salary increases. However, after experience required exceeds 15 years, the average estimated salary. reaches its lowest compared to the other years. The reason for this inconsistency can be attributed for various reasons and outer factors that are not seen in the data. For example, the jobs that require 15 years or more of experience tend to be low paying jobs. Looking further into the data set, we found that jobs related to retail, management and food dominates the job listings in this group.

An ANOVA model is run to see if there is a statistical difference between the mean salary compared to the different years of experience required. The results from the ANOVA analysis is as follows:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	$p < 0.05$
Groups	3	1.023×10^{15}	3.411×10^{14}	443836	$< 2 \times 10^{-16}$	Yes
Residuals	11547900	8.875×10^{15}	7.685×10^{8}			

Table 6: ANOVA analysis for groups of minimum experience required

After running this model, a p-value of approximately 0 is returned, which is statistically significant at $\alpha = 0.05$. Referring to Table 6, it shows that the four experience groups have different means and such are statistically significant (P-value is less than 0.05 significance level). In order to compare the means pairwise, we performed Post Hoc analysis using Tukey's method with the following results:

	diff	lower	upper	p val	$p < 0.05$
2-1	43863.705	43759.466	43967.945	0	Yes
3-1	50672.701	50357.279	50988.124	0	Yes
4-1	-6202.798	-6509.654	-5895.942	0	Yes
3-2	6808.996	6478.192	7139.800	0	Yes
4-2	-50066.503	-50389.149	-49743.857	0	Yes
4-3	-56875.499	-57314.504	-56436.493	0	Yes

Table 7: Post-Hoc analysis using Tukey's method for groups of minimum experience required

Based on the results in Table 7, it is clear that the mean salary of the experience group increases as the level of experience increases for the first three groups. However, the last experience group, which is the one requiring the greatest experience level, has the lowest salary. This shows that further there are other factors that affect the expected salary and indicates that further analysis is required.

Similar to the means, each experience group exhibits different variances in the expected salary. In a dataset in which there are more than two groups to compare, instead of the traditional F-test, levene's test will be used.

	Df	F val	Pr(>F)	$p < 0.05$
Groups	2	2143.7	$< 2.2 \times 10^{-16}$	Yes

Table 8: Levene's test for comparison of variances for groups of minimum experience required

Table 8 shows that differences in variances in each group is statistically significant with p-value near 0. Based on the average variances, it is shown that unlike mean value, the variance of salary increases consistently as the level of experience requirement increases.

Taking a closer look at the data set, for all groups of experience requirements, the industries with the ordered top salaries are in medical, tech, engineering field.

5.2.3 Education requirement

The job's education requirement is the third most important factor that influences salary. There are three categories of job education requirements in the data set, which are No Education, High School Education and Higher Education. To visualise of the estimated salary in each category of education requirement, we created a box plot.

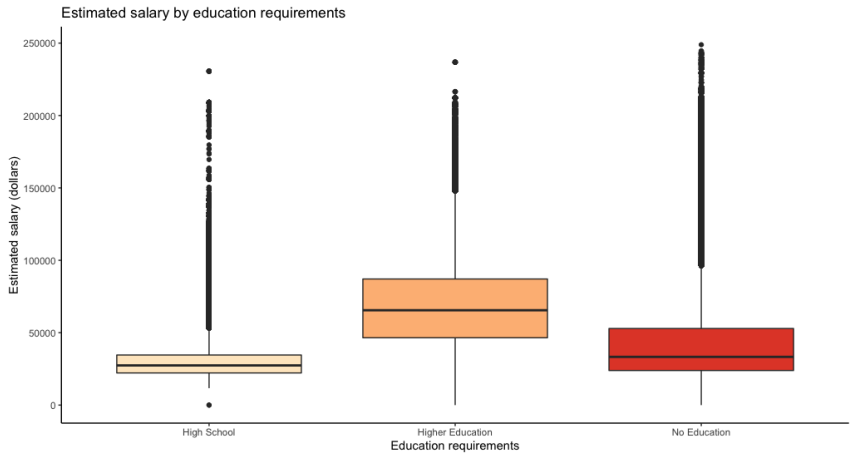

Figure 5: Boxplot of estimated salary relative to education requirements

Referring to 5, we can observe that surprisingly, the mean salary for jobs that does not require education are slightly higher than that for jobs that requires high school education.We will investigate this more closely later in the Post Hoc Analysis.

First, we will be performing mixed ANOVA first to test if there is a difference in mean salaries between the three categories of education. In the ANOVA, we did blocking on job category and experience required to capture their effect on salary since we have come to learn that they influence the estimated salary greatly. The results from the ANOVA analysis is as follows:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	$p < 0.05$
Job Category	56	4.358×10^{15}	7.782×10^{13}	214060	$< 2 \times 10^{-16}$	Yes
Experience Required	41	9.2×10^{14}	2.244×10^{13}	61727	$< 2 \times 10^{-16}$	Yes
Education Requirements	2	4.223×10^{14}	2.111×10^{14}	580778	$< 2 \times 10^{-16}$	Yes
Residuals	11547804	4.198×10^{15}	3.635×10^{8}			

Table 9: ANOVA analysis for education requirements

Referring to Table 9, at a significance level of 0.05, since P_{val} is less than 2e-16, which is less than 0.05, education requirements are still statistically significant after blocking by job category first and the by experienced required. This implies education requirements are significantly different at 0.05 level. We will now investigate whether the mean estimated salary of job listings that does not require any education is statistically different from those that require high school education using Post-Hoc analysis with Tukey's method. The table below only comprise of results that will help with our analysis.

	diff	lower	upper	p val	$p < 0.05$
Higher Education-High School	38260.43	38210.09	38310.77	$< 2 \times 10^{-16}$	Yes
None-High School	13164.03	13119.75	13208.31	$< 2 \times 10^{-16}$	Yes
None-Higher Education	-25096.40	-25139.20	-25053.59	$< 2 \times 10^{-16}$	Yes

Table 10: Post-Hoc analysis using Tukey's method for education requirements

Referring to Table 10, there is a significant difference between estimated salaries based on education requirements at 0.05 level, which supports the analysis from the box plot in Figure 5.

We looked further into why the job listings that does not require education has a higher salary than those that require only a high school education through the data set. We first explored whether work experience could be a possible reason. However, we observe that both categories of education have relatively similar work experience requirement for each respective job category. Aside from that, we found that 50.1% of the job listing does not require an education and 23.7% of the job listing require a high school education. The vast difference could be an indication that the value of a high school diploma is not an important aspect that employers look for in a candidate. Furthermore, we believe that employers place more importance on the ability and skill sets than

the educational background of a candidate. Apart from jobs that requires one to have a higher education, employers probably do not want to place a requirement on education that can discourage those without an education to apply. Reason being that setting a education requirement would not impact the company positively either as it reduces the pool that candidates and pass on the opportunity to hire talented individuals that do not have an education background. Hence, we believe a possible reason why job listings that does not require education has a higher salary than those that require only a high school education through the data set could be due to the lack of value of high school education from the perspective of an employer.

For all categories of education requirements, the industries with the ordered highest salaries are generally tech, engineering, and then medical related jobs. This is possibly because a candidate's skills is more important than their level of education. Most of the time, candidate would need to undergo a technical interview or training to test their skills.

6 Conclusion

There are many other factors that job-seekers usually look for, such as work culture, benefits and salary. In this report, we focus on salary and the factors that influences salary. Given that we have a large data set to begin with, we wanted to capture as much of the data set as possible to make the analysis representative of the population. At the same time, we also encountered limitations in predicting salary using the random forest model as the remaining variability still provides uncertainty for prediction despite being able to explain most of the variability in the test data. Although the model cannot be used to predict the expected salary of a job, it can still effectively draw the three most important factors. With the results of the feature selection, we are still able to further analyse the top three predictors of salary. The main findings from our analysis, that are also the answers to the questions in Section 1 are:

- The three factors that influence the expected salary of a job in a order of significance are:
  1. Job Category
  2. Work Experience Requirements
  3. Education Requirements
- people in **STEM** fields with **higher education** are **more attractive** candidates and hence are offered a higher salary due to the the demand for jobs in these fields.

References

[1] Christina Lee et al. URL: https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/documents/2016-employee-job-satisfaction-and-engagement-report.pdf.

Acknowledgements

For the analysis, we accessed the Indeed data set via this link, as provided by Professor Vivian Lew under the week 6 module "Assignment: Final Dataset Selection Exploratory Data Analysis (EDA)".