Statistical Investigation of State of Data Science and Machine Learning 2020

Jiaqi Feng and Yuqian Wang

PART 0: Contribution Statement

Jiaqi Feng: Part I, Part II 2.1, 2.2, 2.5

Yuqian Wang: Part II 2.3, 2.4, Part IV

Both: Part III

PART I: Introduction

Background

The purpose of this study is to investigate the state of the data science and machine learning industry in 2020 based on a survey conducted by Kaggle concerning the current state of the overall data science industry and a subset of the data science community. From the result of this survey, we want to have a better understanding of the gender, age, education level difference of these respondents and these features' relationship with their work, salary, and tool choices.

Questions & Outline

The main goal of this analysis is to explore the answer of this survey in order to get a better understanding of the general trend as well as the current workers in the data science and machine learning industry. We will do an investigation in the following aspects:

1. To get a sense of general time spent answering the survey, make point and interval estimation of the time spent answering the survey questions
2. Do males have more coding experience (measured by years) on average than females?
3. The age, gender, education difference in each role in the field
4. Do the current roles of participants have an impact on their use of the programming language on a regular basis?
5. Are employees from larger companies (with more employees) more likely to have higher yearly compensation (greater than 150,000 USD) ? If yes, is there a linear relationship between the two?
6. Build a model with selected features to predict how many years have a respondent used machine learning methods.

Data

This dataset contains the result of a data science survey. There are 39+ questions asked and of 20036 responses are collected. The first column contains the time spent for each individual taking the survey and the remaining columns correspond to the answer of a specific question. Note that for answers that are allowed to have multiple choices, different choices are recorded in distinct columns. Each row corresponds to an individual who takes the survey.

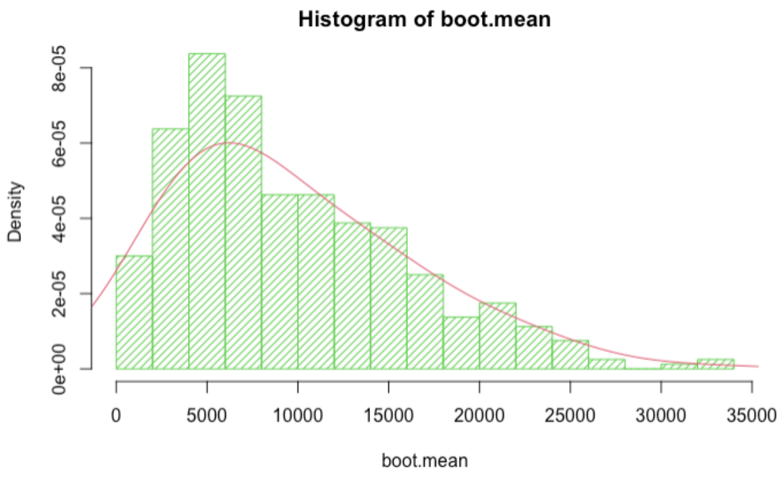
PART II: Basic Analysis

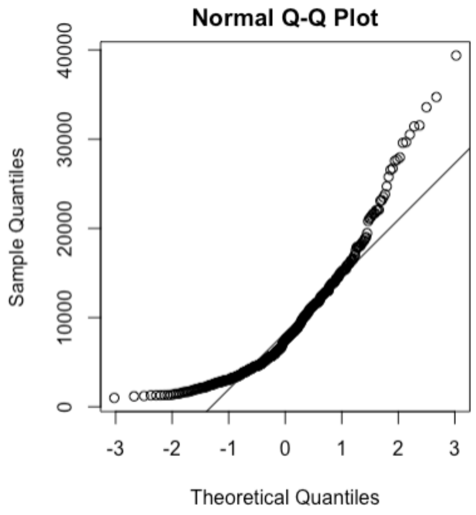
2.1 Point and interval estimation of the time spent answering the survey questions

Methods:

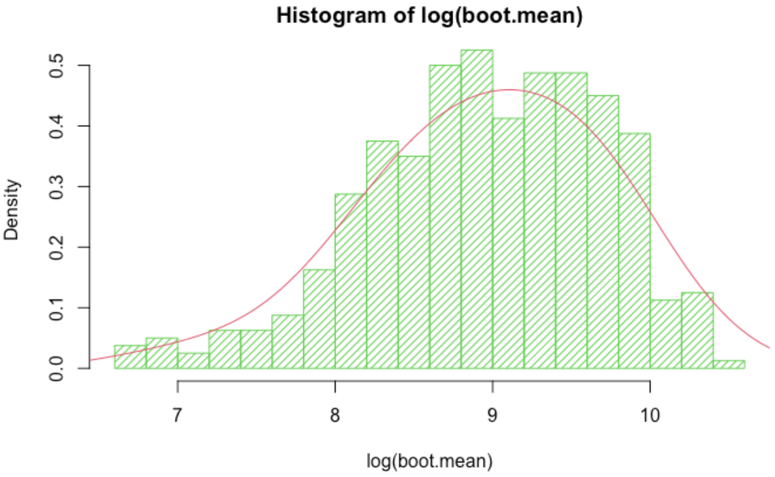
Compute the average amount of time spent answering the survey and construct a 95% confidence interval of the average time spent using bootstrap for 400 times.

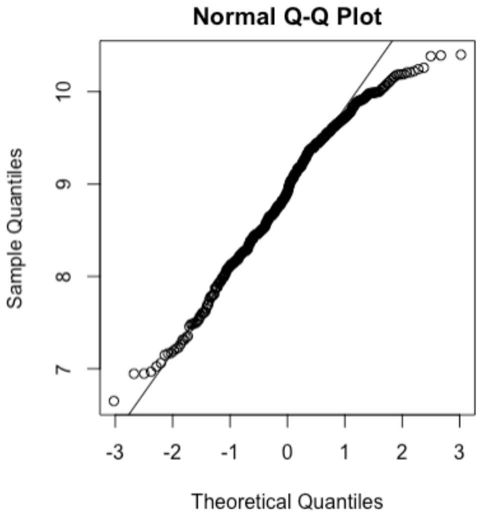
Visualize the bootstrap distribution on original time spent and the Q-Q plot:





After performing log transformation on time spent:





Make point and interval estimation for the log of time spent:

Point estimation for log(time spent): 6.64 seconds

Interval estimation for log(time spent): (5.10, 8.18) seconds

Analysis:

From the histogram of the original bootstrap distribution, we can see that the bootstrap distribution is right skewed, which does not follow a normal distribution. From the Q-Q plot, there are many points deviating from the Q-Q line, which also suggests that the bootstrap distribution is not normal. After taking log transformation on the data, the bootstrap distribution tends to be normally distributed. From the Q-Q plot, the points fit the Q-Q line quite well, so we conclude that the transformed data comes from a normal distribution and make the point estimation for the log of time spent, which is 6.64 seconds. And then construct a 95% confidence interval for the log of time spent, which is (5.10, 8.18) seconds, roughly 165 to 3560 seconds. This means that we are 95% confident that the average amount of the log of time spent answering the survey questions is (5.10, 8.18) seconds. Since the time spent has a great variance, we may expect some variance in the precision of the answers collected in this survey.

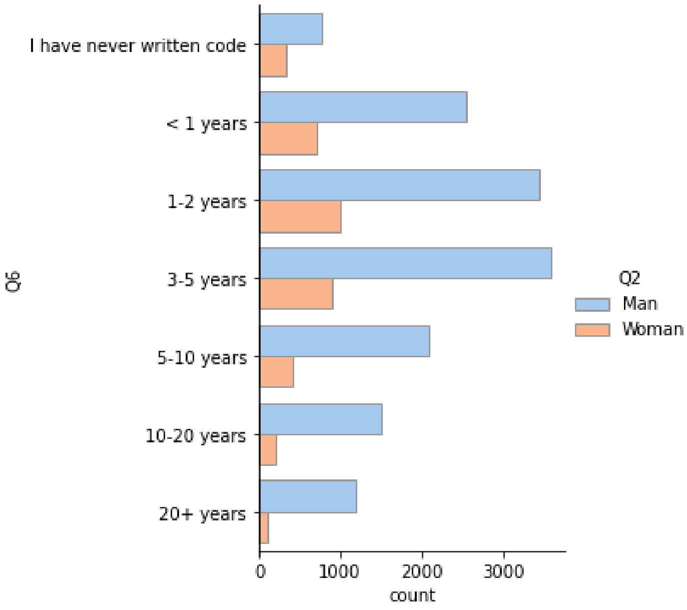
Conclusion:

We can conclude from our bootstrap estimation that the average of the log of time spent answering the survey is about 6.64 seconds based on the point estimation. The interval estimation indicates a log of time spent of 5.10 ~ 8.18 seconds, roughly 165 to 3560 seconds.

2.2 Do males have more coding experience (measured by years) on average than females?

Methods:

First, visualize the distribution of coding experience of males and females:



From the distributions of the two groups, we can see that each one of them follows the normal distribution approximately. Also, there is a slight difference between the two distributions from a close look, as males are more likely to have more coding experience than females.

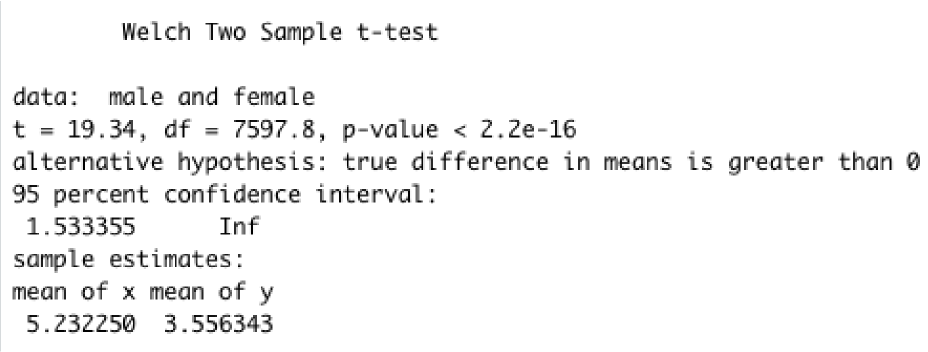
We then perform a two-sample t-test on the average coding experience of males and females. Since coding experience is a categorical variable in our data, we transform each category into a numerical value as a randomly generated integer within the range of each category.

Then we make the hypothesis test:

H0: Males and females have the same amount of coding experience (measured by years) on average.

H1: Males have more coding experience (measured by years) on average than females

Test result:



```
        Welch Two Sample t-test

data:  male and female
t = 19.34, df = 7597.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.533355      Inf
sample estimates:
mean of x mean of y
 5.232250  3.556343
```

Analysis:

From the above test result, we can see that males have about 5.23 years of coding experience on average, and females have 3.56 years on average. We get a p-value of 2.2e^-16 which is extremely small. Therefore, we reject the null hypothesis that males and females have the same amount of coding experience (measured by years) on average, and find a statistically significant difference between the average coding experience of males and females. This aligns with our observation from the plot that the distributions of coding experience of males and females are slightly different as males tend to have more coding experience on average.

Conclusion:

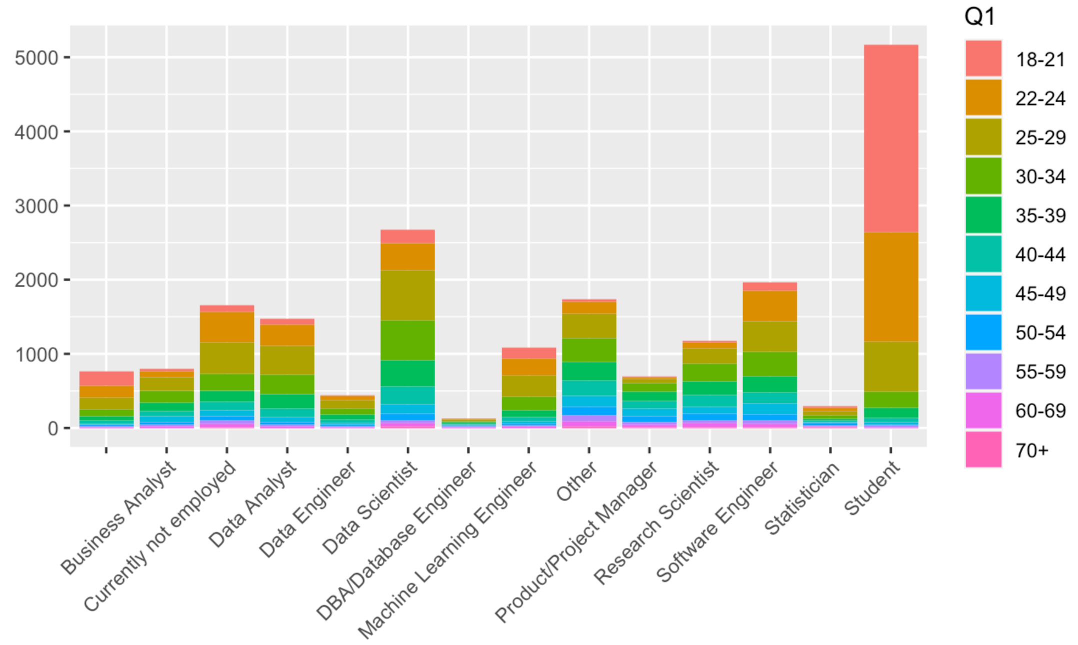
Based on our data, on average, there’s a statistically significant difference between the coding experience of males and females, and males tend to have more coding experience on average.

2.3 The age, gender, education difference in each role

2.3.1 Age Difference

Method:

We group the data by the different age groups and draw the barplot:



Analysis:

From the plot, we can see that most students are in the age groups 18-21 and 22-24, while in other employed roles, the main age groups are 25-29, 30-34, and 35-39. From this result, we can see that people in data science-related works are mostly younger than 40, indicating the data science and machine learning field is young and vigorous in general. However, the role of product/project manager is an exception, in which more than half of the people are over 40, so that these roles may have more requirements on working experience

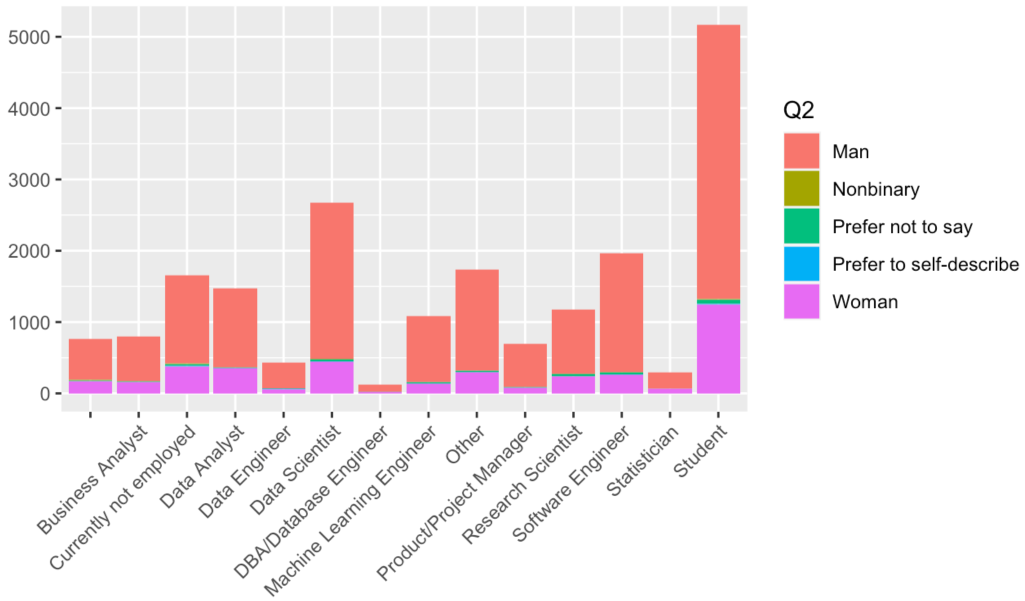
Conclusion:

Except for the data science students who are mainly under 24, most people currently in other employed data science-related roles are of age 25-39, and there are differences in workers' ages among some roles.

2.3.2 Gender Difference

Method:

We group the data by the different gender and draw the barplot:



Analysis:

From the plot, we can see that in each role, the proportion of males is greater than that of females. From EDA, we have found that the gender proportion for all respondents is 81.9% for males and 16.4% for females, which corresponds to our finding. However, we also find that the proportion of females in students is 24.5%, which is higher than that in other roles. This is a good signal since when these students graduate, we may expect an increase of female proportion in data science related fields in industry, so as to mitigate the gender gap.

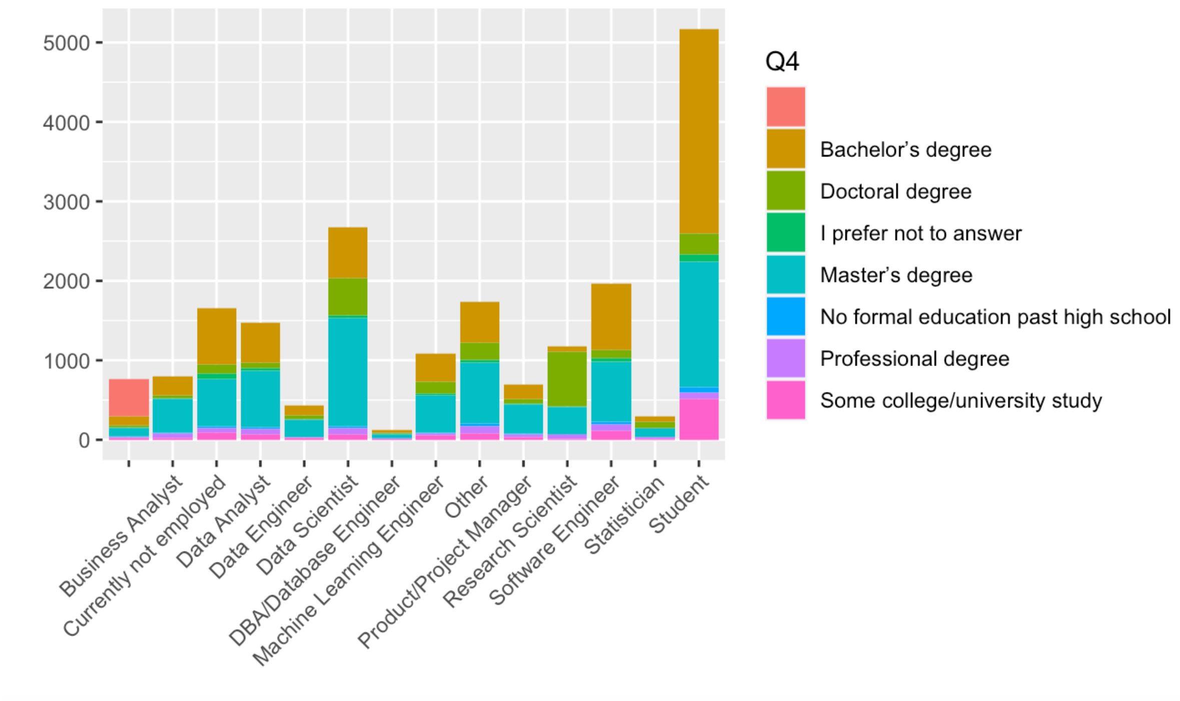
Conclusion:

In all roles, the proportion of males is greater than females, but the difference between these two proportions is smaller in the role of Students.

2.3.3 Education Level Difference

Method:

We group the data by the different education level and draw the barplot:



Analysis:

From the plot, we can see that the most respondents have Bachelor’s or higher degrees. For most roles, the largest proportion of education level is Master’s degree. More specifically, the role of Research Scientist has the largest proportion of Doctoral degrees, and the role of software engineer contains the largest percentage of Bachelor's degree respondents. From these findings, we can see that most works related to Data Science and Machine learning require a higher level of education, but the requirement varies in different titles.

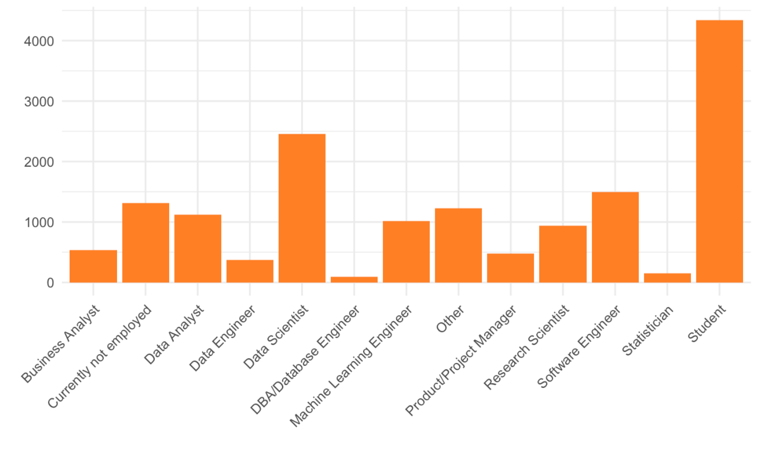
Conclusion:

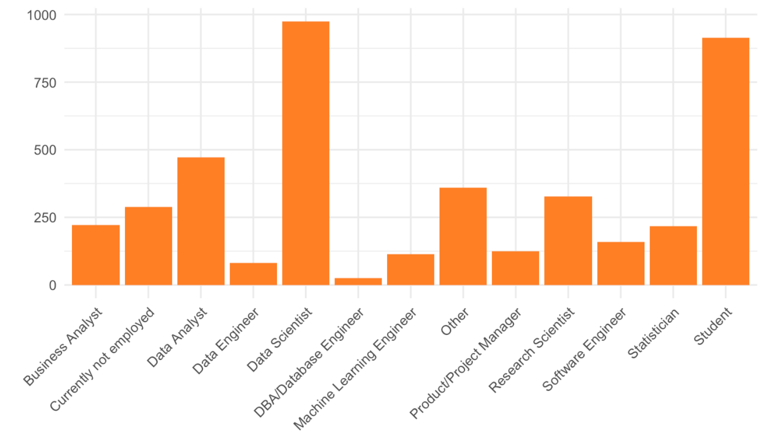
Most employed respondents in each role have an education level of Bachelor’s or higher degree, and the Master’s degree is the most common education level for most of the roles.

2.4 Does the current roles of participants have an impact on their use of the programming language on a regular basis?

Method:

In this question, we want to investigate the difference in current role distribution between people who use different programming languages, more specifically, we want to analyze the different distribution of roles between data scientists using Python and R.

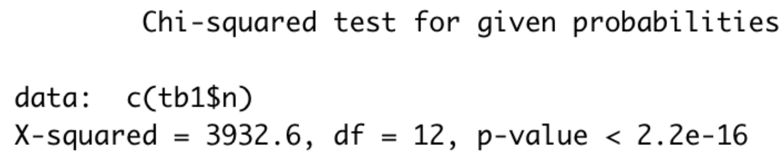




From the two plots, we can see that the distribution of roles for Python and R are not the same. To get a statistically significant conclusion, we conduct a chi-square test to see whether the probability distribution of R and Python are the same.

H0: The distribution of current roles for people using Python is the same as that of R

H1: The distribution of current roles for people using Python is different from that of R



```
	Chi-squared test for given probabilities

data:  c(tb1$n)
X-squared = 3932.6, df = 12, p-value < 2.2e-16
```

Analysis:

From the plot, we can see that the student is the most popular role for people who use Python on a regular basis, however, the most popular role for people who use R is data scientist. In the chi-square test, our p-value is very small, indicating that there's a statistically significant difference between the distribution of current roles among Python users and R users. There are many possible reasons for this difference. For example, Python has become popular in recent years while R is a useful programming language for a long time, and people who have the role of "data scientist" may be older, so more R users are likely to be data scientists. In this sense, although there's a difference between the current roles of different programming languages' users, we cannot say that the current roles of participants have an impact on their use of the programming language on a regular basis.

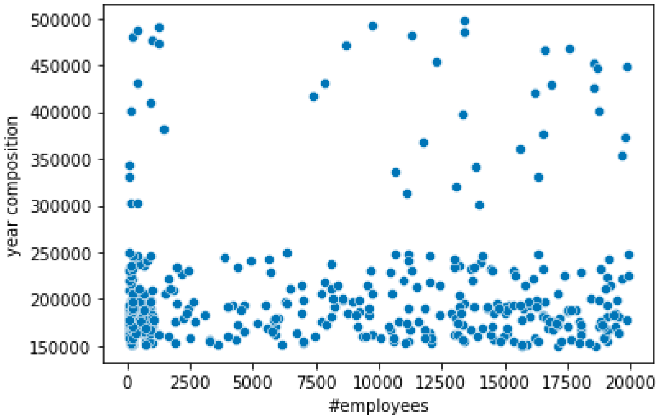
Conclusion:

There's a statistical difference between the distribution of current roles of people who use Python on a regular basis and those who use R, but we cannot conclude that the current roles of participants have an impact on their use of the programming language on a regular basis

2.5 Are employees from larger companies (with more employees) more likely to have higher yearly compensation (greater than 150,000 USD) ? If yes, is there a linear relationship between the two?

Methods:

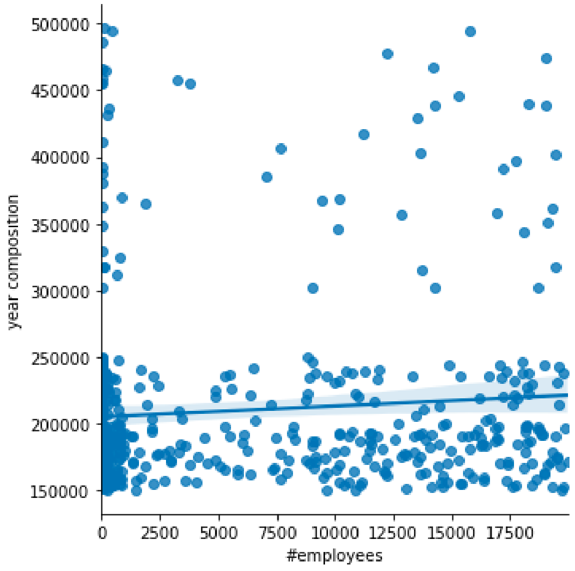
First, visualize the distribution of yearly composition of employees in different size of companies:

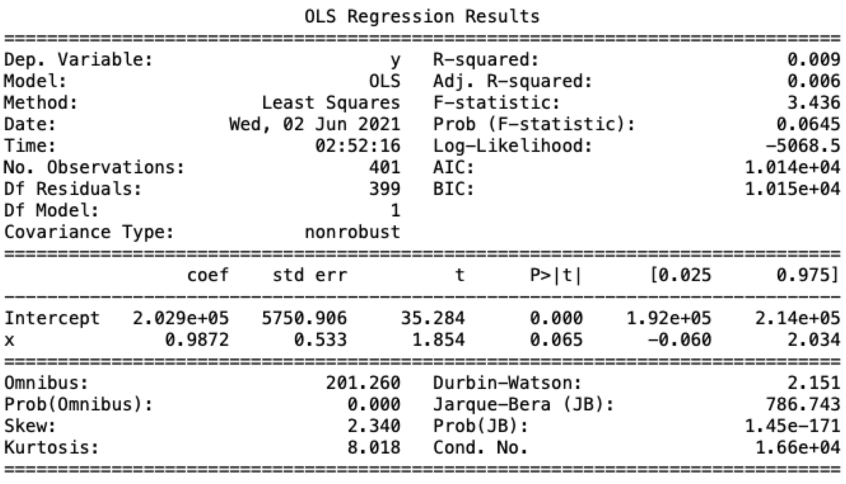


There is no trend observed from the scatter plot.

Next, we want to fit a linear regression model on the number of employees in those companies and the yearly composition of employees to further confirm our observation. Notice that both the number of employees and yearly composition are categorical variables in our data, we transform each category into a numerical value as a randomly generated integer in the range of each category.

Linear model plot and the regression result:





```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.006
Method:                 Least Squares   F-statistic:                     3.436
Date:                Wed, 02 Jun 2021   Prob (F-statistic):             0.0645
Time:                        02:52:16   Log-Likelihood:                -5068.5
No. Observations:                 401   AIC:                         1.014e+04
Df Residuals:                     399   BIC:                         1.015e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2.029e+05   5750.906     35.284      0.000    1.92e+05    2.14e+05
x              0.9872      0.533      1.854      0.065      -0.060       2.034
==============================================================================
Omnibus:                      201.260   Durbin-Watson:                   2.151
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              786.743
Skew:                           2.340   Prob(JB):                    1.45e-171
Kurtosis:                       8.018   Cond. No.                     1.66e+04
==============================================================================
```

Analysis:

We can see from the plot that the regression model we generated is underfitting the data. The regression line has a positive slope indicating that the more employees in the companies, more likely the employees will have higher yearly compensation. But most of the points are deviated from the regression line. Also, from the regression result, we have an R-squared around 0.009, so only 0.9% of the variability in the response variable can be explained by our model, suggesting that our model is very inaccurate. This is reasonable since we have a large set of data with high variability, so it is likely that they do not have a linear relation.

Conclusion:

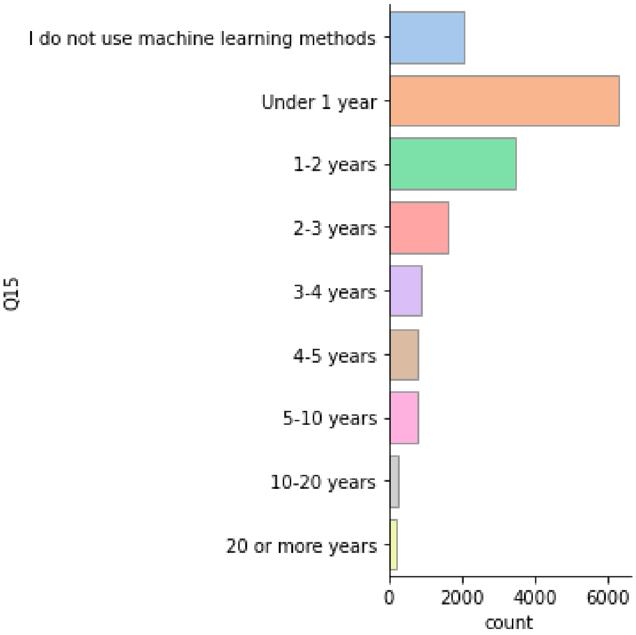
We cannot conclude that employees from larger companies will have much higher yearly compensations, and the two do not have a linear relationship.

PART III: Advanced Analysis

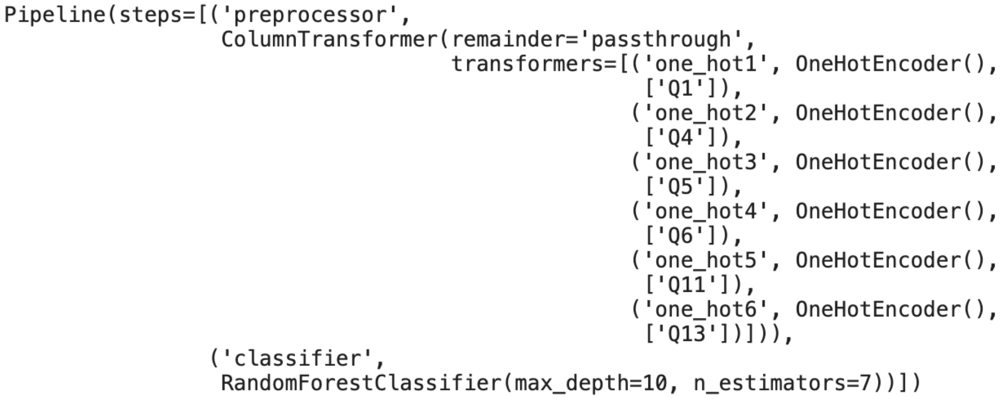
In this part, we will try building a model using the random forest classifier to predict how many years have a respondent used machine learning methods.

Method:

First have a visualization of the distribution of years of machine learning experience of the respondents:



We want to select the features used to build the model. Here, we use the Random Forest Classifier. We select those questions that may be relevant to the experience of machine learning and have only one answer as features (Q1, Q4, Q5, Q6, Q11, and Q13). Since the features are all categorical, we do one-hot encoding to transform the features and generate the model as below:



```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(remainder='passthrough',
                                   transformers=[('one_hot1', OneHotEncoder(),
                                                  ['Q1']),
                                                 ('one_hot2', OneHotEncoder(),
                                                  ['Q4']),
                                                 ('one_hot3', OneHotEncoder(),
                                                  ['Q5']),
                                                 ('one_hot4', OneHotEncoder(),
                                                  ['Q6']),
                                                 ('one_hot5', OneHotEncoder(),
                                                  ['Q11']),
                                                 ('one_hot6', OneHotEncoder(),
                                                  ['Q13'])])),
                ('classifier',
                 RandomForestClassifier(max_depth=10, n_estimators=7))])
```

We then split our dataset into training data and test data, and fit the model on the training set for 100 times. Each time, we use the fitted data to make predictions on the test set and calculate the R-square. We get the average R-square of 0.438.

Analysis:

The average R-square of 0.438 indicates that our model can explain 43.8% variability of the response variable, which is not very accurate. It may be due to the high variability of the data, or the features we selected cannot adequately predict the years of experience of machine learning in reality. Also, some answers to Q15 are close in categories, for example, 1-2 years and 3-4 years are very close, so that predictions of these categories may not be very precise.

Conclusion:

The model we generate can explain 43.8% variability of the response variable, which is not very accurate in predicting the experience of machine learning of respondents. However, due to great variability in real world data, this may be expected to happen.

PART IV: Conclusion

Conclusion Summary:

In Part II, we start with a look at the time spent of respondents answering the survey and make the point and interval estimation of the log of time spent, which is 6.64 seconds and (5.10, 8.18) seconds, roughly 165 to 3560 seconds. We then focus on the features of the respondents and their work from their answers. First, we examine the difference in coding experience between males and females and find that males have more coding experience measured by years than females on average. Then, we go on to investigate the differences in gender, age, and education level in each title of role. For gender differences, we find that males have a larger proportion in each title, but there's a smaller gap in the group 'Student.' For age differences, we find that in most employed titles, respondents are generally from 24 to 39, indicating that the main group in the Data Science field are young people, but age distributions are slightly different in each title. And for education level differences, we find that most people for all the roles have Bachelor's degree or higher, while the most common education level is different in each role, and some even have Doctoral degrees as the most common education level. Besides, we find that the current role distribution for respondents who use R and python is significantly different, but there's no direct causal relationship. We also investigate the linear relationship between company size and yearly compensation and do not find a significant linear relationship between them. Lastly, we build a model using the random forest classifier to predict how many years have a respondent used machine learning methods. Due to great variance in the data, however, our model does not have a robust performance.

Discussion:

Our data is collected from a survey conducted by Kaggle, it's a voluntary sample and is not generated randomly, so there may be bias in this sample. From Part 2.1 of our analysis on time spent answering the survey, we can see that most of the respondents answer the survey with the time spent ranging from roughly 165 to 3560 seconds, which may indicate some variance of precision in their answers. In this sense, we cannot generalize our results to all the people who have work related to data science and machine learning. To get a more accurate and general result, we need to do random sampling on the population of people in the data science and machine learning field at a larger scale and take some measures to better ensure the precision of the answers collected.