Introduction

This report aims to address three key questions proposed by NTEU through the assessment of the RMP website. The report will also conduct an evaluation of the NTEU's claims in accordance with the question analysis.

The claims include that the student evaluations should not be used as a reliable measure for staff performance. Also, that the RMP profiles have been collected by the universities to build a predictive model for professors' ratings without profiles. Factors such as the student perceived difficulties, professor's demographics, discipline of university and study could influence the result of professor's rating. However, the quality of education delivered and choosing professors, or their classes based on the degree of difficulties would impair the academic process.

To ensure the rationale of the analysis, a sensible approach to inspect the overall number of od observations was taken. The result is shown in Table 1. The number of null value data is different for each feature. For example, the grades variable contains 15414 null value data while the student_difficult variable only has 13 missing data. It is not appropriate to remove 15414 missing values from the entire dataset since the impact of dropping these data would lead to an insufficient sample size and then undermine the effectiveness of analysis. Thus, the methods of replacing the null values with "Unknown" and dropping the missing value according to the usefulness and the quantity of the null value data are applied for better data conservation. In addition, Spearman's correlation test and Chi-square Table test were utilized to gain an overview of relevance for analysis (Table 2). Correlation test identifies whether there is a significance of relation between two variables while Chi-square Table test reveals the statistically significant difference between two frequencies. According to Table 2, variables student_difficult and would_take_again have a relatively stronger correlation with ratings, but all features have a relationship with the professor's rating since the p_value is too small to support the null hypothesis that there is a significant difference between the two variables.

```
data.isnull().sum().sort_values(ascending = False)
```

would_take_agains	16161
grades	15414
stu_tags	14710
attence	14453
for_credits	14156
gender	1540
age	1540
comments	101
student_difficult	13
student_star	13
post_date	13
department_name	0
professor_name	0
school_name	0
name_onlines	0
local_name	0
state_name	0
prof_id	0
help_not_useful	0
help_useful	0

Table1. The list of null value data

Variable	Correlation_Coefficient	P_value
gender	0.013	0.001
age	-0.026	0.0
student_difficult	-0.419	0.0
help_useful	-0.08	0.0
help_not_useful	-0.072	0.0
attence_Not Mandatory	-0.013	0.0
attence_Unknown	-0.022	0.0
for_credits_Yes	-0.017	0.0
for_credits_Unknown	0.019	0.0
would_take_agains_Unknown	-0.074	0.0
would_take_agains_Yes	0.249	0.0

Table 2. Statistical Test outcomes

Q1. What affects a professor's rating on RMP?

In response to this question, three suggested factors reported difficulty, professors' demographics, and discipline of school and study are examined step by step. Attribute 'student_star' displays the professor's rating given by student would be a dependent variable to any reported difficulty as the techniques will be used for identifying the influences on the professor's rating based on RMP dataset. 'student_difficult' indicates the level of difficulty student perceived regarding the selected course, which is considered as the top priority factor that closely associated with a professor's rating. Under the pre-condition of difficulty, 'attence'

reported whether the attendance is required for a course is also included in the determinator list, as students may either taken many intensive courses at one time or be struggled to make a balance between their jobs and the study at school, in case of that student are more likely to select a non-mandatory class or rating the professor negatively. Besides, whether a student took the course for credits is also relevant to the professor's rating. Credits oriented students are sensitive to the perceived difficulty of the course. In terms of demographic factors, 'gender' and 'age' tied up with the professor's rating. Students may have gender preference with the professor or professors at one age period are more popular in the school. When it comes to the discipline, the universities under the attribute 'school_name' represent the schools where professors are teaching, and the departments in the 'department_name' category represent the departments where professors are working in are relevant to the rating. University with a good reputation or sufficient resources may attract high-quality teaching staff to back up the overall academic performance or even advancing the rank for the school. Likewise, Professors with high motivation tend to deliver better administrative and academic works.

Apart from above factors, the effects of 'grades', 'state_name', 'would_take_again', and 'stu_tags' have been translated into the professor's rating. In specific, the course score has a noticeable impact on student's sentiment, student received lower mark may ascribe the failure to the teaching team. In addition, a state with a famous university may generate an academic-friendly environment for better study while 'would_take_again' associated with the easiness and acceptableness of the course and teaching Moreover, the stu_tags is the feedback to the professor reflecting the feelings and expectations of the students.

1.1 Exploring the relation between reported difficult and a professor's rating

To gain an insight of the relationship, the numeric variables: prof_id, student_star, sdudent_difficult were selected. Table 3 shows the descriptive statistics. It clearly states that the average rating of professors is 3.7662, and the mean of student reported difficult is 2.8497. The mean of ratings is higher than that of reported difficult.

The graph below explains the relationship between student reported difficult and the professor's rating. The highest and lowest rating for a professor is 5.0 and 1.0 respectively while the level of difficult perceived by student range from 1.00 to 5.00. As shown in the graph, the A student who considered the course is easy to learn rated the professor high while a professor received a low score provided a tough course to student. This reveals a negative relationship between the rating and reported difficult. The student reported difficult should be comprised as a key attribute to the predict model for the professor's rating. As the easiness of a course has a significant impact on the professor's evaluation, it could be interesting to consider whether students do not like difficult courses, or courses are difficult because they are being poorly taught! However, it is not confident to claim that a student enrolled in a lecture with reference to the easiness of the course would impair the lecture quality delivered. The lecture quality metrics have not been stated in the data set provided. A questionnaire or survey is recommended to research on what affects the lecture quality.

	student_star	student_difficult
count	18052.000000	18052.000000
mean	3.787974	2.841181
std	1.402525	1.303649
min	1.000000	1.000000
25%	3.000000	2.000000
50%	4.500000	3.000000
75%	5.000000	4.000000
max	5.000000	5.000000

Table 3. Statistical insights to the relation between reported difficult and the professor's rating

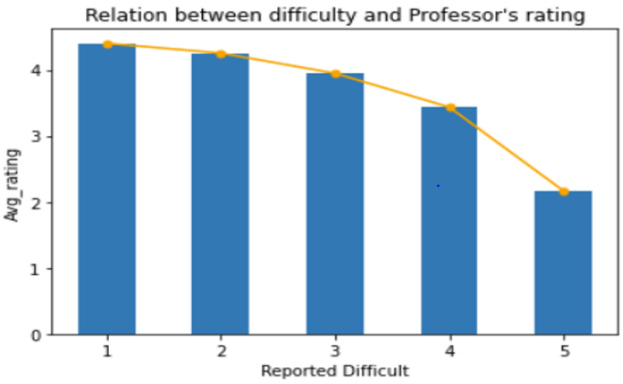


Fig1. Relation between reported difficult and the professor's rating

1.1.1 Exploring the relation between attendance and the professor's rating

Three bars in figure 2 demonstrate the relation between the attendance and average rating of the professor. The average professor's rating by attendance including Mandatory requirements and Unknown is quite similar but the average rating of Not Mandatory attendance is lower than that of Mandatory and unknown, which means the requirement for attendance is also a significant factor that affects the professor's rating. However, the visual output might differ if we drop the Unknown variables. Also, the sample size would have an impact on the visual output, which implies a need of investigating the impact difference.

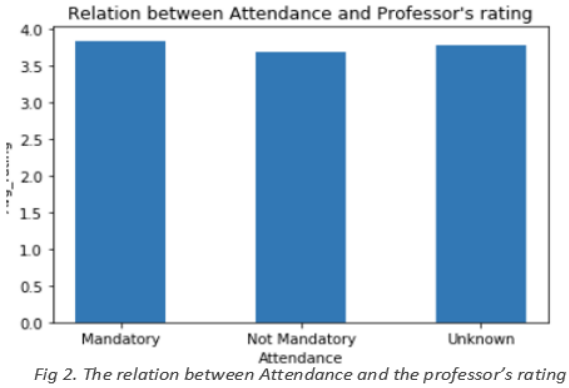


Fig 2. The relation between Attendance and the professor's rating

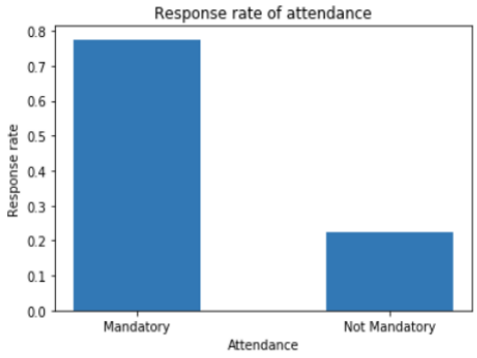


Fig 3. The comparison of impact difference regarding attendance

Figure 3 compares the impact difference between Mandatory and Not Mandatory. According to the chart, the number of ratings received from mandatory courses made up more than 75% of the total number of ratings while the number of rating given by students with non-mandatory course merely accounted for around one-fifth. This means students studied in mandatory classes make more voice than that with the non-mandatory lectures. The NTEUs should take the skewness into account when assessing a professor's performance.

1.2 Exploring how demographics of the professor is related to the professor's rating

Table 4 summarises the statistics of the demographic variables with the professor's rating. The professors' mean age is 47, they have been given a moderate rating with the mean rating of 3.79. The mean of gender is 0.51 with the mean rating of 3.79.

	gender	age	student_star
count	18052.000000	18052.000000	18052.000000
mean	0.505540	47.152393	3.787974
std	0.499983	19.644672	1.402525
min	0.000000	5.000000	1.000000
25%	0.000000	33.000000	3.000000
50%	1.000000	54.000000	4.500000
75%	1.000000	62.000000	5.000000
max	1.000000	104.000000	5.000000

Table 4. Statistical insights to the relation of professor's rating and demographics of the professor

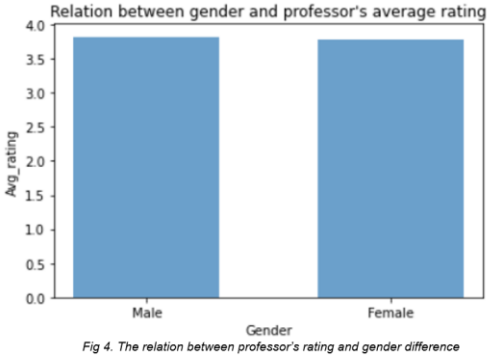


Fig 4. The relation between professor's rating and gender difference

1.2.1 Exploring the relation between gender and professor's rating

The bar chart (Figure 4) visualized the data to give a comparison of the professor's rating by gender difference. The average rating for females is 3.77, slightly lower than that of males (3.82). Also, the mean rating of gender is higher than female professor's rating. The different scores imply the impact of gender is related to the fairness of a professor's rating. This should be included when evaluating a professor's work even it is a small distinction as this may increase when doing skewed distribution.

1.2.2 Exploring the relation between the professor's age and the average rating of the professor.

Figure 5 states the relationship between the professor's average rating and the professor's age. Professors at the age of 80 received the highest rating score of 5.0, following by the age of 75 and 44 with a second-highest rating score of 4.5. The rating scores down to 1.0 when the professor is 73. The average rating scores fluctuate with the vary of age, which is relatively low when the professors' age is in the range of 67 to 72 and 20 to 30. Professors aged 49 to 66 received a moderate score of 3.75 to 4.0. There is not sufficient evidence to rule out the age variables from the attribute factors of the professor's rating. There is not sufficient evidence to rule out the age variables from the impactive factors of the professor's rating. For the phenomenon such as professor aged 80 got the highest score and professor aged 73 with the lowest score requires further investigation to figure out the root of these facts.

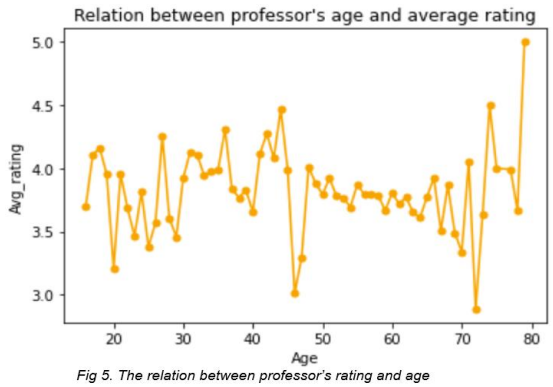


Fig 5. The relation between professor's rating and age

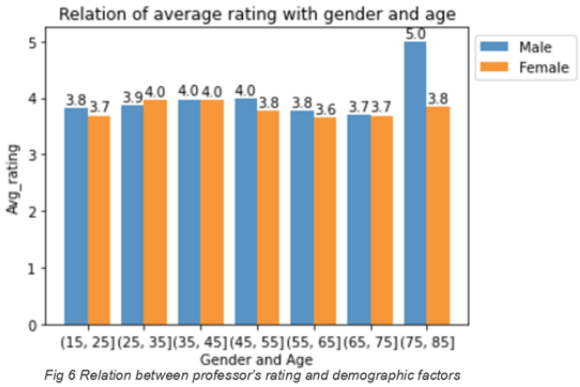


Fig 6 Relation between professor's rating and demographic factors

1.2.3 Relation between demographic factors and rating

The listed age variables range from 15 to 85 is numerous, which need to be transferred to different age groups for the purpose of identify the correlation between demographics and the average rating. Figure 6 describes the relation between the professor's rating and demographic variables. The age variables were divided into 7 different groups and incorporated into gender

and ratings. According to the statistical table, the mean of rating for male professors reaches the highest point (rating=5.0) in the age of 75 to 85 while the highest rating that female professor received is 4.0 with the age range from 25 to 45. With reference to the stacked bar chart (Figure 6), male professors have higher ratings in the age group 15-25, 45-55,55-65 and 75-85 while female professors were only rated a bit higher in the age of 25 to 35. It is obvious that the gender impact has penetrated almost all age groups, except those aged 35 to 45 and 65 to 75. The NTEU should develop further research on the impact of demographic factors to understanding the underlying correlations among different groups, this would help to reduce the demographic biases of the ratings.

1.3 Exploring how discipline of study and university is related to the professor's rating

With reference to Appendix 1, a large propotion of universities and departments are at the rating range 3.0-4.5 and a small percentage of universities and departments with rating below 2.0. There is not sufficient evidence to support a close connection between a professor's rating and the discipline of study and school, but it is necessary for the NTEU to develop the research on the departments and schools with the lowest rating to clarify the cause of that. The factors that affect a professor's rating may be hidden behind.

Given the lack of evidence to support the correlation between two variables (discipline factors vs rating), the research should first focus on universities and departments that have a substantial amount of data to work with. It is more likely that a school or a department has limited data to reflect the reality of their professor's rating, which would bring bias to the professor's evaluation(Cohen, 2013). Figure 8 identifies the top 10 Universities by the number of ratings they received. It found the majority of these schools are public universities, which conveys a message that the research outcomes may more applicable to public universities. According to Appendix 2, professors work in the department such as counselling&Career planning, Latin, Applied Linguistics, Ceramic, Industrial Technology, Development Studies, Guidance, Vocational Education, Media&Design received the highest rating. These departments may have up-to-date lecture design, may involve more interaction between the professors and students, may have a sufficient number of responses to the rating, or equipment with cutting-edge facilities(Agbetsiafa, 2010, Rosen, 2018). These potential causes may have an impact on the reliability of the ratings. However, It is hard to define the associations merely based on the ranking of the department or university, additional variables are required to link the department variables with the professor's rating.

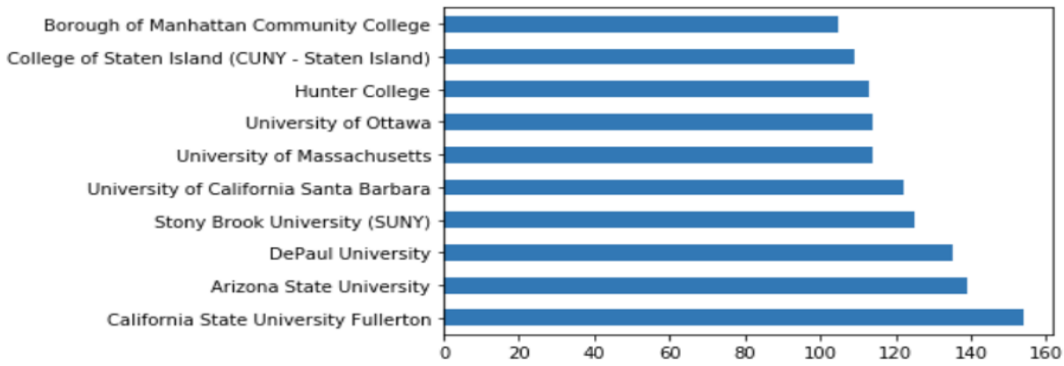


Fig 8. Universities with Substantial Data

1.4 Exploring other variables that may affect professor's rating

1.4.1 Credits oriented vs Professor's rating and Grades vs Professor's rating

From figure 9, it can be clearly observed that students who attended the class for credits give higher ratings for the professor than those who took the class for other purposes. This indicates that the purpose of taking a class has an association with the professor's rating. However, the rating of unknown values is just briefly lower than that of credits oriented, those unidentified values comprise both the rating of credit-oriented and the rating of other purposes. Thereby, it is not confident to confirm that the difference of purposes regarding taking a class would result in prejudice of the professor's rating.

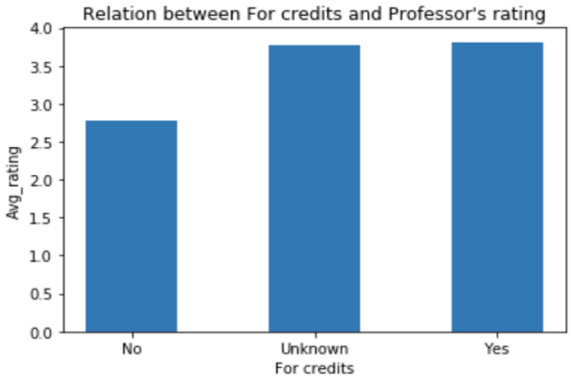


Fig 9. Relation between the professor's rating and student took the course for credits

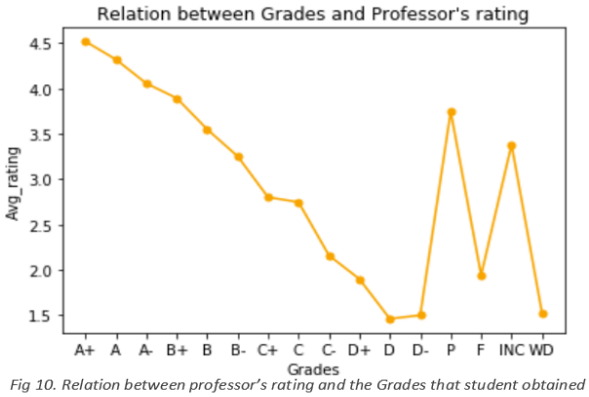


Fig 10. Relation between professor's rating and the Grades that student obtained

In relation to the line chart above (Figure 10), there is a possitive association between the Grades and the rating in the grades range of A+ to D- . In other words, in the range of that, the higher the student's grades, the higher the professor's rating and vise versa. However, students who withdrew the course or failed to pass the course rated the professor very low while students who did not complete the course or just reached the pass line gave higher ratings to professors. It is worth the NTEU invests time in these particular phenomena to detect the hinders of delivering quality teaching. A concusion that grades is one of the determinants of professor's rating can be made based on above observations.

1.4.2 Would take agains vs Professor's rating and Locations vs Professor's rating

As shown in Figure 11, the rating given by the student who would like to take the course again is more than twice as high as those who are unwilling to take the class again. In the case of that, the willingness of a student to take a class again. However, students who did not indicate their willingness to re-take the class also gave relatively high ratings to the professors, thereby makes it hard to draw a persuasive conclusion on the effect of the factor. The reason that leads to the unknown values is critical for the NTEU to make the decision on whether the data should be used to evaluate the professors since the rating distinction between students who would re-take the lecture and those who are unenthusiastic to re-take the lecture may impair the reliability of the justification.

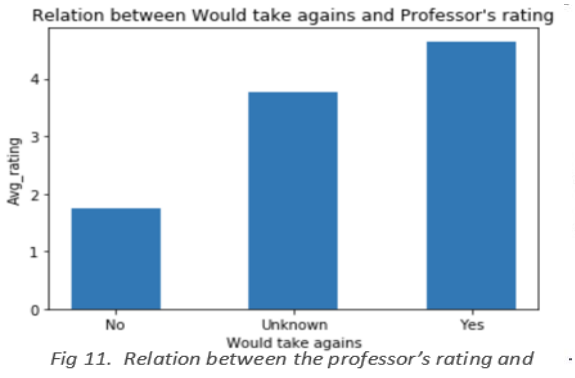


Fig 11. Relation between the professor's rating and whether the student would take the class again

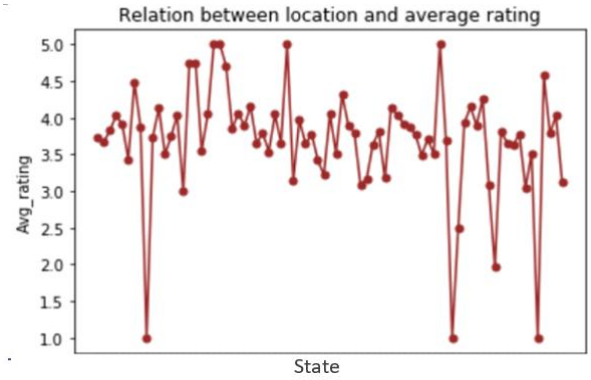


Fig 12. Relation between the professor's rating and the state

Figure 12 exhibits how location is connected to the professor's rating. Apparently, most states that the school is in have a moderate to high ratings (3.0-4.4), a small number of states present an extreme score of rating =5.0 or rating =1.0. Universities with higher scores such as the University of Oxford located at OXFORDSHIRE may also be high in world academic ranking (Appendix 3; Khosrowjerdi, & Zeraatkar, 2020). WARWICKSHIRE, PE (Prince Edward Island), and CARDIFF rank at the very bottom with a 1.0 rating score (Appendix 3). However, there is only one feedback(rating) from each of the three states. In the situation that, the ratings are too bias to be used as an indicator for performance administration. Also, it is hard to claim the location has a significant influence on professor's rating.

1.4.3 Student tags vs Professor's rating

The word could figure helps to identify the key themes of the students' descriptions. It is apparent that tough grader, gives good feedback, participation matters, skip class, grading criteria, amazing lectures, lecture heavy, won pass, and clear grading are top concerns that related to the professor's rating. These themes can be applied on a questionnaire basis for future data collection, which may go in-depth with qualitative analysis.

Fig 13. The description provided by students regarding the professors.

1.5 Summary of the findings

After running the correlation and Chi-square test, it indicated which variable may have a relationship with the professor's rating. Based on the statistical outputs, a combination of techniques (EDA) is used to explore the relevancy between the rating and other variables. It found that the reported difficult is negatively correlated with the professor's rating, the grades (range from A+ to D-) is positively associated with the rating, variables including the gender, age, for credits, would take again, attendance have effect on the professor's rating, and no sufficient evidence has been found in the exploration of the relationship between university

and rating, department and rating, and location and rating. It is also identified that the existence of unknown variables, lack sufficient sample size and the extreme phenomena would undermine the reliability of the professor's rating. Additionally, student tags revealed that the quality of feedback, the suitability of grading criteria, the attractiveness of the lecture, the quantity of lecture, the applicability of the grading approach are matter to the professor's rating.

Q2. Can a model be built to predict a professor's rating?

In this section, regression analysis is used as statistical modelling for estimating the relationship between a dependent variable and independent variables. In this case, the dependent variable is the professor's rating which refers back to the 'student_star' column. Besides, the rest of the variables could be considered as independent variables.

Before performing a regression analysis, it is necessary to have an overview of the data type for each variable. This is because some of them are qualitative or categorical data instead of quantitative. Additionally, the technique of regression analysis cannot be executed when there is a qualitative phenomenon. The solution for this limitation is to convert qualitative predictors by dummy variables.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 18052 entries, 0 to 19684
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   prof_id            18052 non-null  int64
 1   professor_name     18052 non-null  object
 2   gender             18052 non-null  int64
 3   age                18052 non-null  int64
 4   school_name        18052 non-null  object
 5   department_name    18052 non-null  object
 6   local_name         18052 non-null  object
 7   state_name         18052 non-null  object
 8   prof_id            18052 non-null  int64
 9   post_date          18052 non-null  object
 10  name_onlines       18052 non-null  object
 11  student_star       18052 non-null  float64
 12  student_difficult  18052 non-null  float64
 13  attence            18052 non-null  object
 14  for_credits        18052 non-null  object
 15  would_take_agains  18052 non-null  object
 16  grades             18052 non-null  object
 17  stu_tags           18052 non-null  object
 18  help_useful        18052 non-null  int64
 19  help_not_useful    18052 non-null  int64
 20  comments           18052 non-null  object
dtypes: float64(2), int64(6), object(13)
memory usage: 3.7+ MB
```

Based on the above information, there are 8 quantitative data and 13 qualitative data. Among these non-numerical predictors, it is practical to convert would_take_agains, attence and for_credits variables into dummy variables since each of them only have three unique values. Also, it is crucial to consider multicollinearity by defining k-1 dummy variables when there are k values. The reason is that the information provided by k dummy variables is redundant. Therefore, it should be removed by adding drop_first = True argument. Also, the rest of the categorical data need to be removed from the dataset for analysis purpose.

```
# Drop all the variables which could not be convert into numerical forms.
num_data = data.drop(["prof_id","professor_name","school_name","department_name","local_name",
                      "state_name","post_date","grades","name_onlines","stu_tags",
                      "comments"], axis = 1)
num_data.head(5)
```

	gender	age	student_star	student_difficult	attence	for_credits	would_take_agains	help_useful	help_not_useful
0	1	64	3.5	2.0	Unknown	Unknown	Unknown	0	0
1	1	5	5.0	1.0	Unknown	Unknown	Unknown	0	0
2	1	5	5.0	1.0	Unknown	Unknown	Unknown	0	0
7	0	68	3.0	2.0	Unknown	Unknown	Unknown	0	0
8	0	30	1.0	4.0	Unknown	Unknown	Unknown	2	0

```
# Convert attence, for_credits, would_take_agains into dummy variables
num_data = pd.get_dummies(num_data,drop_first = True)
num_data.head(5)
```

	gender	age	student_star	student_difficult	help_useful	help_not_useful	attence_Not Mandatory	attence_Unknown	for_credits_Unknown	for_credits_Yes	would_take_a
0	1	64	3.5	2.0	0	0	0	1	1	0	
1	1	5	5.0	1.0	0	0	0	1	1	0	
2	1	5	5.0	1.0	0	0	0	1	1	0	
7	0	68	3.0	2.0	0	0	0	1	1	0	
8	0	30	1.0	4.0	2	0	0	1	1	0	

2.1 Multiple linear regression model

We would like to analyse the relationship of professor's rating against multiple variables such as gender, age and class difficulty etc. Therefore, it is better to use multiple linear regression which allows us to increase the accuracy of the model since more variable is included.

2.1.1 Train-vali-test split

To begin, we need to divide our data into three datasets in order to obtain 50% training data, 25% validation data and 25% test data. For example, this data contains a total number of 18,052 values. Accordingly, 50% training data will be equivalent to 9,026 values, 25% validation and test data are equal to 4,513 values representatively. Furthermore, the data can be separated into different proportions depends on the building model and the amount of available data. In this case, we assume 50% training data, 25% validation data and 25% test data is a reasonable distribution to carry out the modelling.

```
# Split the data into 50% training data, 25% validation data and 25% test data
X = num_data.drop(["student_star","attence_Unknown",
                   "for_credits_Unknown","would_take_agains_Unknown"],axis =1).to_numpy()
y = num_data["student_star"].to_numpy()

X_tv, X_test, y_tv, y_test = train_test_split(X, y, test_size=0.25,random_state = 1)
X_train, X_vali, y_train, y_vali = train_test_split(X_tv, y_tv, test_size=1/3,random_state = 1)


print("Test:", X_test.shape)
print("Train and vali:", X_tv.shape)
print("Training:", X_train.shape)
print("Validation:", X_vali.shape)
```

```
Test: (4513, 8)
Train and vali: (13539, 8)
Training: (9026, 8)
Validation: (4513, 8)
```

Additionally, these three different datasets are used for distinct purposes. The training dataset is used for computing and training the model. The validation dataset is used for testing the model with various hyperparameters. After determining the optimal hyperparameters, the model will be trained again in the combined dataset before being evaluated on the test data. Moreover, this approach allows us to choose the optimal hyperparameters and prevent overfit the data.

2.1.2 Train the model and compute parameter

```
# Build and fit the multiple linear regression model
linear_reg = LinearRegression()
linear_reg.fit(X_train.reshape(-1,8), y_train)

# Print model parameters
print("beta 0: {:.4f}".format(linear_reg.intercept_))
for i in range(8):
    print("beta {}: {:.4f}".format(i+1, linear_reg.coef_[i]))
```

```
beta 0: 5.1376
beta 1: 0.0522
beta 2: -0.0022
beta 3: -0.4480
beta 4: -0.1015
beta 5: -0.0706
beta 6: -0.1554
beta 7: -0.3891
beta 8: 1.2801
```

```
betai_name = pd.Series(linear_reg.coef_,
                           index = num_data.drop(["student_star","attence_Unknown","would_take_agains_Unknown",
                                                  "for_credits_Unknown"],axis =1).columns)
betai_name
```

```
gender                   0.052218
age                     -0.002191
student_difficult       -0.448000
help_useful             -0.101462
help_not_useful         -0.070609
attence_Not Mandatory   -0.155373
for_credits_Yes         -0.389121
would_take_agains_Yes    1.280142
dtype: float64
```

In this section, we train the model in the training dataset and compute the parameter. As we can observe, beta 0 represents the intercept term and the rest of beta signifies the average increase in y associated with a one-unit increase in x. Therefore, the formula for fitting multiple linear regression models is given by:

student_star = 5.1376 + 0.0522 × gender – 0.0022 × age – 0.4480 × student_difficult – 0.1015 × help_useful – 0.0706 × help_not_useful – 0.1554 × attence_ Not Mandatory – 0.3891 × for_credit_Yes + 1.2801 × would_take_agains_Yes + ϵ

2.1.3 Predict with the regression model and calculate MSE

1. Training data & Validation data

```
# Predict with the multiple linear regression model
pred_train = linear_reg.predict(X_train)
pred_vali = linear_reg.predict(X_vali)

# Calculate the MSE
mse_train = mse(pred_train,y_train)
mse_vali = mse(pred_vali,y_vali)

print("Train mse: {:.4f}".format(mse_train))
print("Validation mse: {:.4f}".format(mse_vali))
```

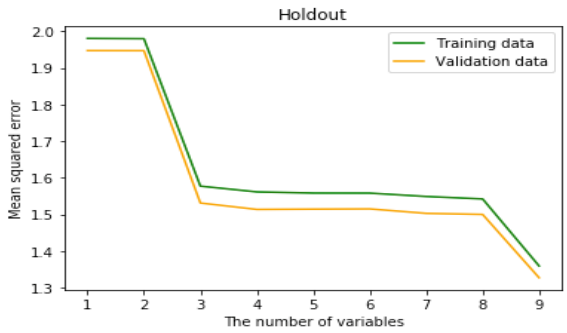
```
Train mse: 1.4296
Validation mse: 1.4061
```

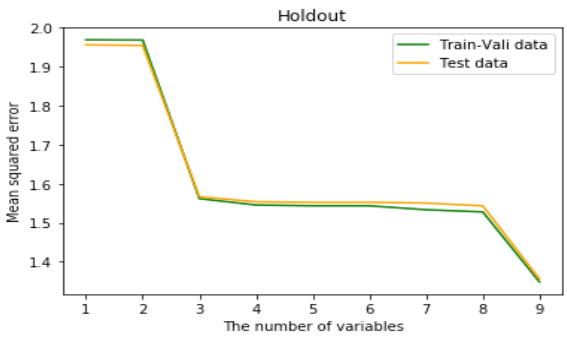
2.Train-Vali data & Test data

```
linear_reg = LinearRegression()
linear_reg.fit(X_tv, y_tv)

pred_tv = linear_reg.predict(X_tv)
pred_test = linear_reg.predict(X_test)

mse_tv = mse(pred_tv,y_tv)
mse_test = mse(pred_test,y_test)

print("Train-vali mse: {:.4f}".format(mse_tv))
print("Test mse: {:.4f}".format(mse_test))
```

```
Train-vali mse: 1.3481
Test mse: 1.3556
```

3. Holdout Validation





After predicting the regression model, mean squared error (MSE) is introduced to evaluate the model performance as it measures how close a fitted line is to the data point. In other words, how well the model predicts the observed data. Generally speaking, the smaller MSE value indicates better model performance. According to the result, the MSE is smaller when the model was built based on the train-vali and test dataset. This result reveals that the model will perform better when the number of training sample increases. Overall, the MSE value is relatively small, which means the data samples are close to the regression line.

2.2 Assumption and shortcoming

Multiple linear regression model is developed based on the following two assumptions. First of all, the relationship between target and predictor is linear. Besides, the error is independent and normally distributed with the same variance. Furthermore, this technique is limited to a linear relationship which sometimes is impractical and leads to erroneous and misleading results. Besides, it cannot be executed when there is a qualitative phenomenon.

2.3 Prototype - example

A 34 years old female professor is holding a class which attendance is compulsory. Most of the student takes this class for credit and the level of difficulty is 1. Additionally, students comment that this class is very helpful and would like to take again.

```
features = np.array([1, 0, 34, 1, 5, 0, 1, 1, 1])
betas = np.array([5.1376, 0.0522, -0.0022, -0.4480, -0.1015, -0.0706, -0.1554, -0.3891, 1.2801])

print(features @ betas)
```

```
4.8429
```

Based on the above information, the professor's rating is very high at 4.8429 which seems to be reasonable.

2.3.1 Text analytics – student_star vs comments

After building the model for the professor's rating against numerical variables, we believe it is worth examining whether we could develop another model for non-numerical variable such as 'comments'.

First of all, we extract out the comment based on the professor rating. This extraction is processed based on the following assumptions:

1. We assume the comments provide positive feedbacks if the professor rating is more than or equal to 4.
2. We assume the comments provide negative feedbacks if the professor rating is less than or equal to 2.

3. We assume any feedback falls outside the range of above two assumptions will be considered as median feedback.

```
good_corpus[:5]
```

```
['such a fun professor. really helpful and knows his stuff',
 "Such a easy class. It\\'s simple. Do your homework and pay attention and you will fly right by or be the person tha
t blames him for not leaarning. He wont let you fail. just ask for help....",
 'She was awesome! If you went to class and listened actively good grades were easy to get. She is extremely helpful
to anyone who actually goes out to seek help. My advice show up to class, ask questions and go to office hours if you
need more help.',
 'One of the best classes I took at UM.',
 'Steph was very helpful and cared about our experience. She wanted to be sure we learned and were comfortable with e
verything before our testing. Go to her office hours if you need help. I learned a lot.']
```

```
bad_corpus[:5]
```

```
['Stephanie ALWAYS had a mad or bored or "I could be doing so many better things with my time right now than babysitt
ing these college students" look on her face. When a student asked a question, she quickly breezed over a vague expla
nation and speededly moved to the next subject. She is extremely unapproachable and cold, and "tricky."',
 "Horrible. I went to class everyday and it still didn\\'t make a difference. Her tests do not reflect her level of t
eaching.",
 'No Comments',
 "she talks too fast, she doesn\\'t do anything in the class--she doesn\\'t make her own slides, she has her TA\\'s g
rade everything, so if you have a problem with a grade, don\\'t go talk to her... she has no say in the grades. print
off the slides for each lecture! they always stay the same.",
 'terrible!']
```

After dividing the comments into different categories, we use term frequency – inverse document frequency (TF-iDF) to emphasize the uncommon word which could be useful and put fewer weights on the common words. Then, develop a model based on TF-IDF representation. However, this model is difficult to interpret since corpus contains a numerous amount of token with unique beta. Therefore, a further advanced NLP analysis is required to improve the model interpretation.

1. Positive feedback vs Professor rating

```
X_pos = tfidf_df_pos
y_pos = pos[1]

Xpos_train, Xpos_test, ypos_train, ypos_test = train_test_split(X_pos, y_pos, test_size=0.20,random_state = 1)

linear_reg_pos = LinearRegression()
linear_reg_pos.fit(Xpos_train, ypos_train)

print(linear_reg_pos.intercept_)
print(linear_reg_pos.coef_)
```

```
4.628310208911243
[-0.00492927  0.27713214 -0.07105481 -0.03236784  0.24413769  0.3518435
 -0.10858896  0.14868819  0.02597738  0.11630651  0.16301682  0.02092884
  0.12163333  0.02314424 -0.20308612 -0.09475522 -0.12782592  0.0374296
 -0.02614004 -0.02651372 -0.01899899  0.06784327  0.02308787 -0.14999338
  0.12539389  0.02760339  0.03195337  0.01829729 -0.12551486  0.01236798
  0.00419326  0.12643552 -0.05874443 -0.22744594  0.02206866  0.12738901
  0.18734379  0.00700062 -0.0233654  -0.01841856  0.04482221  0.03379253
  0.07588541  0.11299298 -0.11931767 -0.0194286   0.0102659  -0.2597433
 -0.04221955 -0.1219271  -0.0456001   0.1686952   0.21845923 -0.04650168
  0.17423722  0.00137355  0.09106554 -0.02665632 -0.0233422   0.06843647
 -0.12798579 -0.27252132 -0.07710917  0.05193783 -0.28034909  0.02101236
  0.18614587  0.13029248 -0.04690892 -0.09161549 -0.1513748  -0.08917709
 -0.00583741  0.06714103  0.03335485  0.06745134  0.04353253  0.06727554
  0.04226663  0.11048609  0.05193647  0.11336928  0.15735165  0.11842814
 -0.02409478 -0.0987334   0.00106211  0.02944011 -0.07176026  0.04081555
  0.02374682  0.07305896  0.10024084  0.12124284  0.03163319  0.00722331
  0.11256862 -0.00947779  0.02385548  0.01843888]
```

2. Negative feedback vs Professor rating

```
X_neg = tfidf_df_neg
y_neg = neg[1]

Xneg_train, Xneg_test, yneg_train, yneg_test = train_test_split(X_neg, y_neg, test_size=0.20,random_state = 1)

linear_reg_neg = LinearRegression()
linear_reg_neg.fit(Xneg_train, yneg_train)

print(linear_reg_neg.intercept_)
print(linear_reg_neg.coef_)
```

```
1.3925011637454663
[ 0.12340768 -0.02497677 -0.08504896  0.04206883 -0.21151256 -0.05312308
  0.12445794 -0.01541264  0.11852075  0.01517665  0.1740709   0.05387394
  0.10981867 -0.07729682  0.09845796  0.00209138 -0.07210158  0.00800429
  0.04314618 -0.10609547 -0.22675455  0.01628642  0.29378859  0.00882465
 -0.19096194 -0.10945913  0.05918207  0.17010151 -0.0147684  -0.06001047
  0.23809911  0.04683895 -0.04125792 -0.02694754 -0.01068204 -0.0071459
  0.2078215  -0.02207352  0.0869953   0.0538272  -0.24885062  0.00760763
 -0.08112961 -0.01663189 -0.0011923   0.1388626  -0.01008394  0.22399582
  0.25743973  0.34683292  0.071712   -0.22950116 -0.04399876  0.00738398
  0.27549401  0.10503882  0.03487    -0.01976694  0.02540496  0.08667904
  0.02071817  0.2342855  -0.00873482 -0.09585589 -0.04796236  0.06249888
  0.15657051  0.14156214  0.24014172  0.2867108  -0.05800118 -0.20286191
 -0.09414827 -0.05983985 -0.15536112 -0.08052924  0.35289691 -0.06091499
  0.04478997 -0.10380061 -0.12151156 -0.02588945 -0.19166395 -0.2572762
 -0.15985153  0.04821504  0.0410843   0.2073082  -0.06099131 -0.098132
 -0.03739316 -0.19680246 -0.12471575 -0.0041419  -0.05137749 -0.01057042
  0.03670106 -0.01059023  0.14414577 -0.42864028]
```

In conclusion, a multiple linear regression model can be built to predict a professor's rating. To be specific, some information is indispensable to anticipate the professor's rating. For example, gender, age, the difficulty of class, attendance requirement, the usefulness of class and whether the student took the class for credits, as well as whether the student would take the professor again. Some of the information is hard to collect without accessing the profile on RMP. Therefore, this model could not be used to anticipate the rating for professor without a profile. Alternatively, the university or college could develop their profile system by collecting the data from student surveys.

Q3: What are the social and ethical issues involved?

3.1 Ethical Issues

There are many unique challenges when dealing with data in this age of information. The ones with the most direct and severe impact on society are the ethical issues concerned with dealing with this data. As data can be transformed into information, it should be treated as intellectual capital especially when the data is about human beings. Of the many and varied ethical issues that can arise from dealing with this kind of data, it is helpful to focus on the four pillars proposed in the PAPA Framework. These include:

Privacy: The privacy pillar considers the issues related to the exposure of sensitive data. In the case of RMP, it is critical that the data collected by them is by mutual consent with the professors as it contains specific information about their demographics as well as their job profiles. Professors may choose to declare certain information "A" and "B" about themselves and withhold the rest. The ethical implication here is that disclosing information "A" and "B" may lead to revealing information "C" about the professors that may then be used by the RMP website, even though this information was intended to be kept private. This is called exposure to minute description. Hence, there is possibility that RMP is not honouring the right of professors to their privacy and revealing unauthorised information. When Universities use this deduced data, they too breach the rights of the professors.

Accuracy: The accuracy pillar addresses the issues regarding misinformation that arises from inaccurate data. It is one of the most important issues to address when assessing any kind of information since even a small proportion of inaccurate data can skew the results drastically. The RMP website consists of a large number of profiles that have missing data in them. A large number of values are missing from various columns. This skews the available data in this field

heavily and us of this data to model professor ratings could lead to misrepresentation of data. It is important to note that data such as gender and age is inferred using 'ageforname' Python package. This raises serious concerns when these values are considered in predicting professor ratings. As the accuracy of this data cannot be verified, universities using this information will produce inaccurate results and possibly give wrong predictions of professor ratings.

Property: The property pillar focusses on issues of ownership of data. As the data being collected is of the professors, they should ultimately be entitled to its ownership, unless there is an agreement between the RMP website and the professors which gives RMP the ownership. Defining the ownership of this intellectual property is of utmost importance in terms of the law and prevention of misuse of this data. If the data is owned by the professors and is being taken by universities for unintended use of data, this could raise major ethical concerns and even be deemed unlawful. Similarly, if RMP owns this data and decides to sell the data for a purpose which the professors did not agree to. Data sets are costly to produce at the source, but the cost of reproduction is negligible. Therefore, it is unethical for the owner this intellectual property to not be compensated fairly its use.

Access: This pillar is concerned with the issues regarding accessibility of data. If any individual can access the data on demand from the RMP website and use it for various purposes, it violates the ethics of property and privacy. Even though it may not be unethical for universities to access data regarding professors and their ratings, what they do with the data can be. Hence, ethical concerns here are that even though unauthorized retrieval of data may be prevented, harmful consequences can arise with misuse of data by authorized personnel.

3.2 Social Impact

Universities are suspected to use the data from the RMP website, but this can come at severe societal costs. No matter what decision the universities plant to take based on this data could be flawed. This is because the collected data about professor's performance is subjective and only pertains to the student's perspective. As suspected by the NTU, it could be based on numerous factors like demographics, department of study and the difficulty of the course. Therefore, the question to be addressed here, is that should they really use a model that may be fundamentally flawed due to certain biases within it?

Even if the data was accurate and predictions of the were a true reflection of a professor's overall standard of teaching, the final purpose of this data needs to be evaluated. The primary use of this data would be made in hiring decisions. In such a scenario, decisions are being considering things like gender, department of study, age, location of professors and the difficulty of class. This would lead to discrimination due to biases in each of these areas:

a) Gender discrimination if a male professor is chosen over a female with the same credentials, but the model predicts males are more likely to be better.
b) Age discrimination if a younger professor is snubbed for an elder one only because of the model's prediction of a higher approval rating for them.
c) Bias toward hiring professors that make the class easier because they are likely to get higher ratings, resulting in poor quality of the courses they teach.
d) Bias against hiring professors from specific locations.

The universities may then go on to share the information they derive from modelling RMP website's data and share that information with other institutions. This means that if one professor is deemed to be bad based on their rating, they may never get a job again. This also means that when other institutions may want to hire the best academic professionals using

prediction form this model, they will often get a false result and may not end up with the right personnel.

3.3 Ethical issues in the team's analysis

Issues when during pre-processing and EDA:

a. Dealing with missing values: A number of columns that contained information regarding attendance, grades, credits and retaking the course were missing. Therefore, an analysis on the remainder of the data is not the most accurate representation of the true scenario, thereby compromising on the quality of analysis.

b. Assumption of age and gender: As the age and gender are predicted using the 'ageforname' python package and not true values, any result derived using this data is just an estimate. This may portray certain age groups and genders in a better light unintentionally.

c. Analysing university and department data: There were insufficient sample of data points to analyse professor ratings belonging to a particular university or department. Therefore, a single positive or negative rating was enough to determine performance based on these two factors, which leads to misrepresentation of data.

Issues during modelling:

a. Variables used: The variables used in modelling of data were selected by performing statistical analysis on the data provided to the team. This analysis may produce distorted the results as: the accuracy of data cannot be verified; certain assumptions are made about the data and results may have skewed pre-processing of data. This could result in inclusion of unwanted variables in the model, which may affect the final prediction.

b. Training, validation and test data: During model formulation, 50% data was used to train, 25% data to validate and 25% data to test the model. This goes against the rule of thumb which requires 60% data to train, 20% data to validate and 20% data to test the model. This could lead to undertraining of the model since 50% data may not be enough to achieve the best fit model.

Conclusion

The exploratory analysis revealed that the data on RMP website was not of high quality due to which the university's model may have suffered. The team's modelling revealed that the professor ratings were dependent on various factors which allowed biases to creep into the results, thereby leading to misinformation. The NTEU's claims of professor rating being dependent on perceived difficulty and demographics but not enough evidence was found for university and department of teaching to have an influence over the ratings. All factors considered, the ratings are only subjective and must not be used as absolute truths about professors. Being dependent on models that predict ratings using these factors raise serious ethical concerns and can have a very negative social impact. NTEU must ensure universities are not following such a practice and device policies against this practice.

Appendix 1

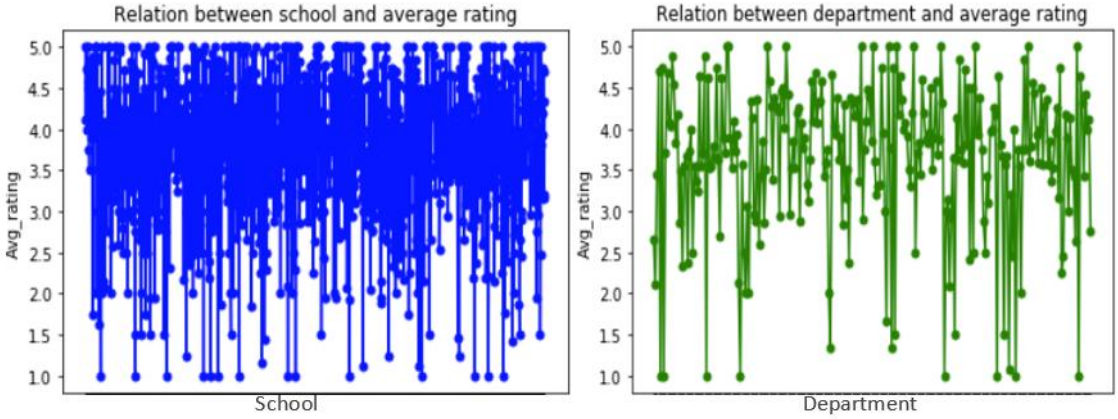
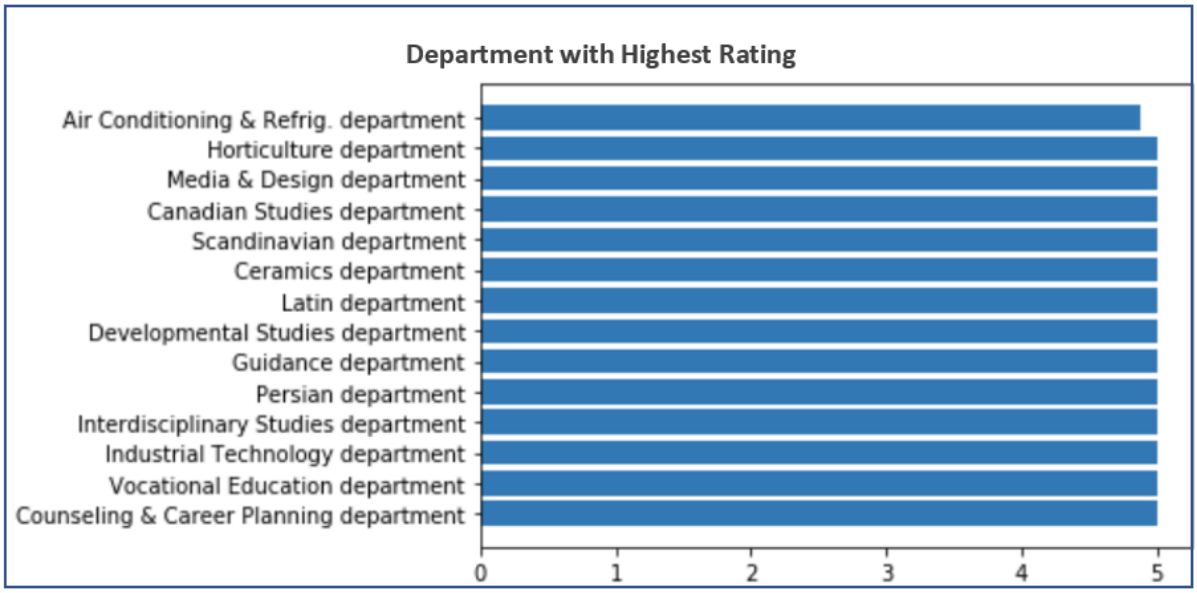


Fig 7. Relation between the discipline of study and university and professor's rating

Appendix 2



Appendix 3

State_name	Rating
MANCHESTER	5.000000
HAMPSHIRE	5.000000
GLASGOW	5.000000
OXFORDSHIRE	5.000000
EAST SUSSEX	4.750000
EDINBURGH	4.750000
HI	4.696970
WEST MIDLANDS	4.576923
Baltimore	4.486842
NB	4.314286

Table 5. Top 10 states by rating

State_name	Value counts
CA	3409
NY	2233
FL	773
TX	711
OH	621
...	...
PE	1
HAMPSHIRE	1
GLASGOW	1
MANCHESTER	1
CARDIFF	1

Table 6. Count the number of responses

State_name	Rating
WARWICKSHIRE	1.000000
PE	1.000000
CARDIFF	1.000000
Sacramento	1.960526
Pomona	2.500000
DURHAM	3.000000
VT	3.041667
SK	3.086957
NE	3.092105
WY	3.125000

Table 7. Bottom 10 state by rating