

For office use only

Team Control Number

For office use only

T1 _____

0000

F1 _____

T2 _____

F2 _____

T3 _____

Problem Chosen

F3 _____

T4 _____

A

F4 _____

2019
MCM/ICM
Summary Sheet

Research on graduate early warning model

Summary

To solve this problem, we use the method of principal component analysis to analyze the time series. We take two factors in consideration: the number of SCI, EI and papers respectively, and the number of students, average score, number of students under the guidance of tutors; Finally, we get the related formulas and give the reliable prediction results. The conclusion of sensitivity test shows that the model is robust.

Keywords: PCA; Factor Analysis; Time Series

Research on graduate early warning model

December 16, 2019

Summary

To solve this problem, we use the method of principal component analysis to analyze the time series. We take two factors in consideration: the number of SCI, EI and papers respectively, and the number of students, average score, number of students under the guidance of tutors; Finally, we get the related formulas and give the reliable prediction results. The conclusion of sensitivity test shows that the model is robust.

Keywords: PCA; Factor Analysis; Time Series

Contents

1	Introduction	3
2	Analysis of the Problem	3
3	Calculating and Simplifying the Model	4
4	The Model Results	6
5	Evaluate of the Mode	10
5.1	Strengths	10
5.2	Weaknesses	11
	Reference	12
	Appendices	12
	Appendix A Code	12
	Appendix B Data	13

List of Figures

1	Factor1	6
2	Factor2	7
3	Result1	8
4	Result2	9
5	Delay Time	9

List of Tables

1	Result	5
2	Data	14

List of Questions

Question 1	4
Question 2	4
Question 3	4

1 Introduction

In today's world, the graduation problem of college students has become increasingly serious. Many college students can't even finish their studies on time, so they are eliminated. Therefore, it is of great significance to analyze the graduation problems of college students.

Doctoral education is the highest level of graduate education. The level of doctoral education not only reflects the level of national higher education, but also reflects the level of national scientific research. In recent years, with the rapid growth of doctoral students, the phenomenon that doctoral students can not graduate on time is becoming more and more common, which has become the focus of social attention and discussion in recent years. However, there are various and complex reasons that affect the overdue graduation of doctoral students. How to find out the key factors and the degree of influence that affect the overdue graduation of doctoral students, so as to take effective measures to strengthen management, cultivate doctoral students in line with social needs within the specified time, minimize the number of overdue graduates and timely transport high-level talents for the country, is what this study should explore. It is the core theme of this study. According to the doctoral data of a university as the supporting data of the whole research, we find out the key factors influencing the doctoral delay through factor analysis, and on this basis, we calculate the score coefficient of the factors, and take the graduation indicators of the University as the required data, which are not mentioned in all policies. We use the standard value for processing, and succeed This paper proposes a model of overdue warning. According to the theory of sampling, the number of graduates who may graduate in the next year and the number of those who may delay their graduation can be predicted from the sample to the whole.^[1-3]

2 Analysis of the Problem

Colleges and research institutes in China recruit a large number of graduate students every year. After several years of study, these graduate students should graduate from their own colleges and research institutes. At present, the graduate situation of each unit is not optimistic, and the number of graduate students who can not graduate smoothly for various reasons is increasing year by year, which gradually attracted the attention of relevant units and the Ministry of education.

Please establish the relevant mathematical model according to the study conditions of graduate students (Master's degree and Doctor's degree) of Nanjing University of Science and Technology (NJUST) over the years, so as to provide decision-making reference for graduate school to send graduation warning to some graduate students who may not graduate successfully in normal time. The specific problems are as follows:

Question 1

Establish a mathematical model, and give the main factors and proportion that affect graduate students to graduate smoothly in normal time.

Question 2

Establish a mathematical model, give an early warning one year in advance, and give a suggestion to extend the study (extend the time) and an evaluation standard to terminate the study.

Question 3

Establish a mathematical model to predict the graduate students who will graduate smoothly and need to extend their studies (including the extension of time) in 2020 in Nanjing University of Science and Technology.

3 Calculating and Simplifying the Model

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. Factor analysis aims to find independent latent variables.

The theory behind factor analytic methods is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Factor analysis is commonly used in biology, psychometrics, personality theories, marketing, product management, operations research, and finance. It may help to deal with data sets where there are large numbers of observed variables that are thought to reflect a smaller number of underlying/latent variables. It is one of the most commonly used inter-dependency techniques and is used when the relevant set of variables shows a systematic inter-dependence and the objective is to find out the latent factors that create a commonality.

Factor analysis is related to principal component analysis (PCA), but the two are not identical. There has been significant controversy in the field over differences between the two techniques (see section on exploratory factor analysis versus principal components analysis below). PCA can be considered as a more basic version of exploratory factor analysis (EFA) that was developed in the early days prior to the advent of high-speed computers. Both PCA and factor analysis aim to reduce the dimensionality of a set of data, but the approaches taken to do so are different for the two techniques. Factor analysis is clearly designed with the objective to identify certain unobservable factors from the observed variables, where as PCA

does not directly address this objective; at best, PCA provides an approximation to the required factors. From the point of view of exploratory analysis, the eigenvalues of PCA are inflated component loadings, i.e., contaminated with error variance.

We take two factors in consideration:

1. SCI, EI, papers;
2. Undergraduate school level, average score, number of students under the guidance of tutors;

Factors that influence whether a doctoral student can graduate are divided into scientific research factors and platform factors. Factor analysis was performed on the relevant data obtained. Factor analysis is based on correlations between variables, and based on these correlations, variables are combined to form the smallest factor, representing the total change in the variables, simplifying the variables, Clarify why. Scientific research factors are clearly divided into the number of SCI publications, the number of EI publications, and the number of general articles published. The platform elements are the undergraduate level, the average grade of the course, the instructor-led students Is divided into a number of. The calculation method for the doctoral degree evaluation index is based on the establishment of the above evaluation index set matrix and the standardization of data, and the correlation coefficient matrix R is calculated according to the following formula.

$$(r_{jk})_{m \times m} = \frac{1}{N} \sum_{i=1}^N Y_{ij} Y_{ik} (j, k = 1, 2, \dots, m) \quad (1)$$

The results of the factor analysis are as table 1.

Table 1: Result

S	F_1	F_2
S_1	0.2703	0.0752
S_2	0.2677	0.0310
S_3	0.1096	-0.2773
S_4	0.3009	0.0233
S_5	0.0699	0.3431
S_6	0.0865	0.3393

The results of factor analysis indicate that among the factors in scientific research, the SCI factor and EI factor are the main factors. At the platform factor, students' courses are divided into key factors. The total number of published articles, fresh graduates, and subject opening hours has a secondary effect on whether a doctoral student can graduate.

At the same time, we also obtained the following results: First, the number of doctoral students who published the SCI and EI papers had a significant impact on their ability to end time, while overall the number of doctoral students who publish ordinary papers has little effect on their ability to finish time. . Second, the level of students in undergraduate schools, the average course scores and the number of students their teachers teach have a significant impact on their ability to graduate, but opening hours. on the topic and whether new doctoral students are having less impact.

The SCI factor and the EI factor, the level of primary school for doctoral students, the average number of study points and the number of students supervised by the tutor have a greater influence on the successful completion of doctoral students. The influence of time and doctoral degree on new students is relatively small. It illustrates the need to offer courses at the doctoral stage. The more scientific papers published by doctoral students during their studies, the more they can complete their studies in good time, which also confirms the universities' requirements for the publication of scientific research results by doctoral students. The smaller the number of students led by the tutor, the more an individual student can be advised, and the more likely it is that the tutor will offer inexpensive instructions and various practical aids for the completion of the doctoral students.

4 The Model Results

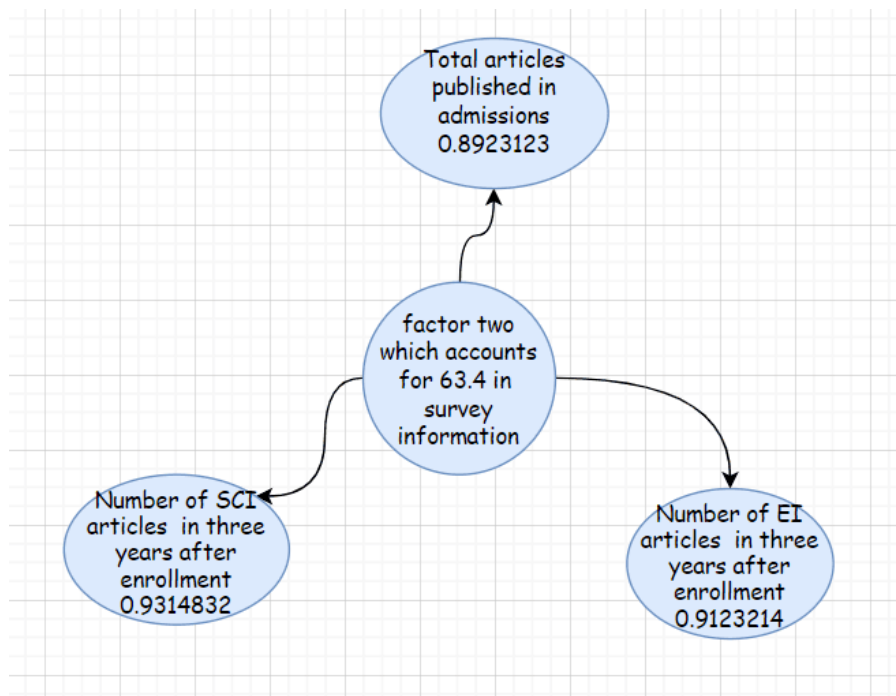


Figure 1: Factor1

After choosing two main factors, the larger the contribution rate, the more able

to reflect the effect of the main factor on the completion time. For each main factor, the closer the internal value of the main factor load matrix is to 1, the greater the impact of the corresponding index on delay graduation. The two main factors and related indicators are shown in Figures 1,2,and 3 as follows: From the main factor that accounts for the largest contribution rate in Figure 1, it can be seen that the three factors that affect the delay graduation time are total SCI published in three years, total articles published in admission and total EI published in three years, so we consider the main factor as a scientific research factor. This is also consistent with the current policy of various universities, that is, publishing several papers that meet the grade, such as professional journals such as SCI and EI. Of course, we can also see that the scores of factors after the rotation factor are not as influential as those of ordinary journals. Impact factor journals are more influential.

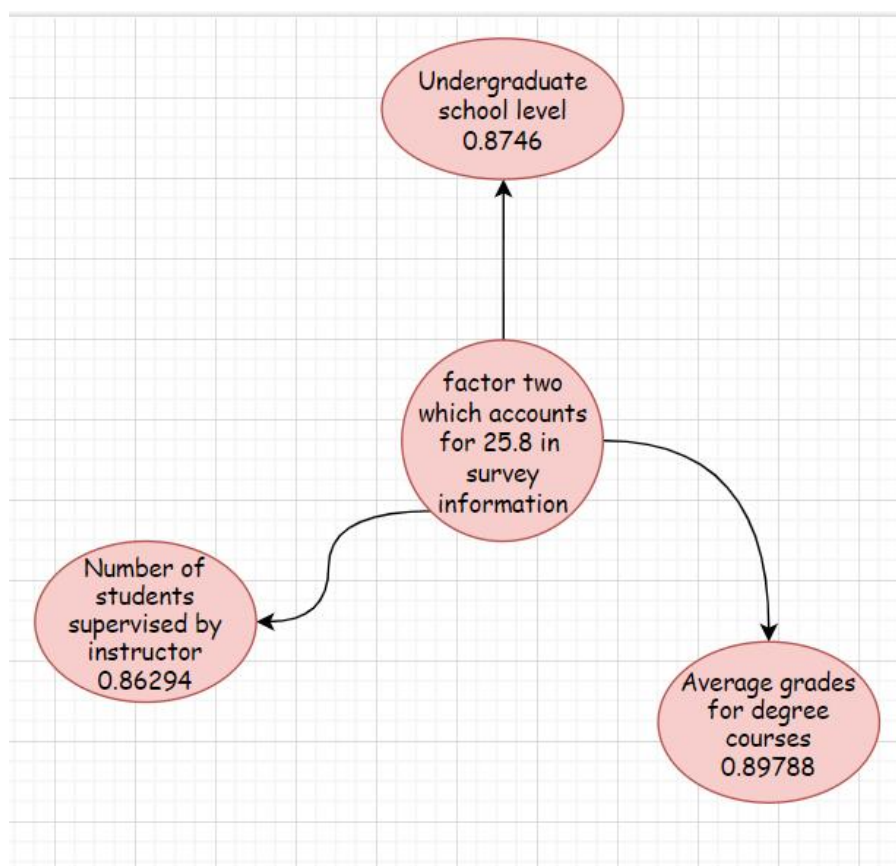


Figure 2: Factor2

As can be seen from the main factor 2 which accounts for 25.8% of the contribution rate in Figure2, the three factors affecting delay graduation time. Main factor two is named platform factor by us, which includes the grade of undergraduate school, the number of students under the guidance of tutors, and the weighted academic performance of normal courses. The main factor two reflects the impact of a good platform on the delay time. If the undergraduate receives more systematic discipline training, the higher the tutor's teaching level, coupled with the individual efforts of students, will greatly reduce the possibility of delay. Especially for personal factors, we can see that the rotated factors are very close to 0.9.

	gender	Is a freshman	Whether to start the question on time	Undergraduate school level	Number of students supervised by instructor	Average grades for degree courses	Number of SCI articles published in three years of enrollment	Number of EI articles published three years after enrollment	Total articles published in admissions
0	male	no	yes	985	1	84.519772	0	1	2
1	female	no	yes	211	9	77.872854	2	4	7
2	female	yes	no	general	9	75.856038	0	4	5
3	female	yes	no	211	12	77.218129	0	2	3
4	male	no	yes	general	2	84.311015	3	1	5
5	female	no	no	211	14	78.445607	0	4	5
6	female	no	yes	211	4	74.776042	1	2	4
7	male	no	no	211	10	84.120066	0	0	1
8	female	no	no	985	7	77.323060	3	2	6
9	male	no	no	general	8	76.187739	3	2	6
10	male	no	yes	general	2	81.121091	1	0	2
11	female	no	no	general	13	78.284035	3	0	4
12	male	yes	yes	211	4	80.356020	1	4	6
13	female	no	no	211	2	77.223041	1	0	2
14	female	no	no	211	12	86.679415	2	4	7
15	male	no	yes	985	0	74.737380	2	1	4
16	female	no	no	211	10	68.857326	2	1	4
17	male	yes	no	211	2	78.066110	4	0	5
18	male	no	yes	general	4	79.421984	4	2	7
19	male	no	no	general	2	80.291050	4	0	5
20	female	no	yes	985	12	75.845347	0	0	1
21	female	yes	no	general	10	80.588750	2	3	6
22	female	yes	yes	211	11	80.729490	0	2	3
23	male	no	no	general	7	76.362930	2	1	4
24	male	no	no	211	4	77.482237	3	2	6
25	female	no	no	985	4	66.419019	4	0	5
26	female	no	yes	985	7	72.068672	4	1	6
27	female	yes	yes	985	11	78.261155	0	1	2
28	female	yes	no	211	5	81.200907	1	4	6
29	male	no	no	general	8	63.111150	4	1	6

Figure 3: Result1

Figure 3 show that the data source is based on a questionnaire survey from the academic research center of a university. Because there are a lot of data with missing values, it will affect the distribution of the overall variable. So in this process we filter out some missing values and follow After processing the data, a 238 rows and 10 columns DataFrame table was obtained. The table serves as a quantitative criterion for evaluating a doctoral student, including whether it is gender, Is a freshman, Whether to start the question on time, Undergraduate school level, Number of students supervised by instructor, Average grades for degree courses, Number of SCI articles published in three years of enrollment, Number of SCI articles published within 4 years of enrollment, Number of EI articles published three years after enrollment, Total articles published in admissions. In order to analyze the distribution of data, we calculate some basic characteristics of the data set, including min, max, std and other attributes.

As figure 4, we use the relationship between the factor scores to successfully calculate the factor scores of the Ph.D. in three years, and obtain the minimum factor score corresponding to the actual extension, and use this value to determine whether the doctoral students can graduate normally. In this case, because the data we collected are all data one year before graduation, the implementation of the model may change due to changes in the specific policies of the school, so this model has limitations, but it is sufficient to roughly judge whether it is a certain period of time. The result of being able to graduate on time.

According to statistical analysis, we can conclude that when the factor score is greater than 99.017585, we must graduate on time. When the factor score is less than 78.173158, the possibility of postponing graduation will greatly increase. When

t[298]:

	gender	is a freshman	Whether to start the question on time	Undergraduate school level	Number of students supervised by instructor	Average grades for degree courses	Number of SCI articles published in three years of enrollment	Number of EI articles published three years after enrollment	Total articles published in admissions	factor scores
0	male	yes	no	211	9	75.048577	3	1	4	136.666038
1	male	yes	yes	985	7	72.340240	4	0	4	31.014875
2	female	no	yes	general	0	81.332792	3	0	3	80.254915
3	male	yes	yes	211	11	69.495320	1	3	4	63.520383
4	male	no	yes	general	0	70.215601	4	0	4	62.377773
5	female	yes	yes	general	0	70.391123	1	3	4	43.412088
6	female	yes	no	general	13	79.525192	3	2	5	116.802059
7	female	yes	yes	general	4	72.463704	1	0	1	119.729813
8	male	no	yes	211	11	73.940347	3	4	7	85.791169
9	female	yes	no	general	8	81.900059	4	3	7	124.415388
10	male	no	yes	general	12	68.149558	0	0	0	77.557279
11	male	yes	no	211	2	71.365892	2	2	4	113.780147
12	male	no	yes	211	6	74.349398	3	2	5	83.850924
13	male	no	yes	211	14	69.097505	1	0	1	75.146180
14	male	no	no	985	9	75.427781	3	0	3	94.368964
15	female	yes	no	985	3	82.029226	2	2	4	61.274854
16	male	yes	no	211	6	73.100779	3	2	5	123.890765
17	male	yes	no	general	11	77.476316	4	1	5	117.930846
18	male	yes	no	211	12	74.340304	2	1	3	119.268623
19	male	no	yes	general	4	76.368179	2	4	6	125.506688

Figure 4: Result2

in between, it should be noted that the possibility of delay will greatly increase.

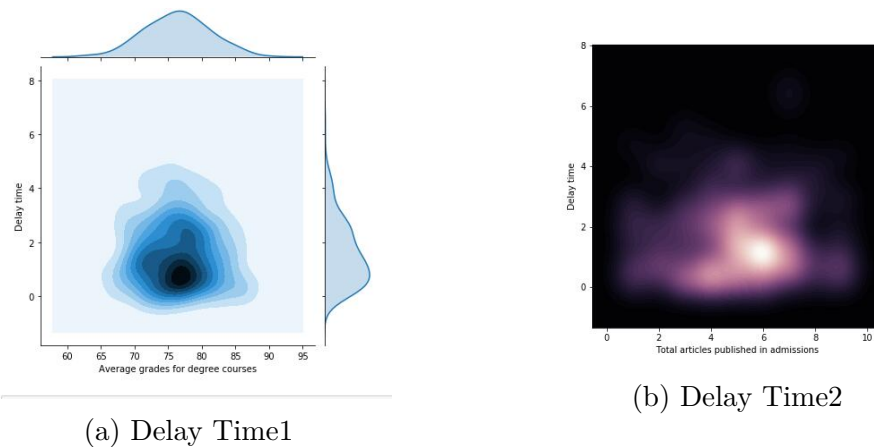


Figure 5: Delay Time

We can see from the data set that there are three types of classification attributes, namely gender, the level of the undergraduate school, whether it is a freshman, and whether the question is opened on time. According to my own understanding of the whole problem, so I made a corresponding pivot table to provide more relevant information about the data. We can also provide more input data sources for our factor analysis based on this pivot table, and more Data combination form.

5 Evaluate of the Mode

5.1 Strengths

Factor analysis is an extension of multiple regression analysis, including some modifications that make the approach less likely to be influenced by the personal biases of the researcher, and less dependent on the assumption of independence among variables. Technical explanations of factor analysis at various levels of difficulty can be found in the books by Noruis (1994), Kline (1994), Yotopoulos and Nugent (1976), and Mulaik (1974).

Ideally, in multiple regression analysis, one would observe a dependent variable on the left hand side of an equation that is explained by a set of independent variables on the right hand side. The problem, however, is that the connecting link between dependent and independent variables needs a strong theoretical basis. Furthermore, many of the so called independent variables may not really be independent at all, but rather change or move together in response to some other unknown variables or factors. Hence the researcher usually encounters criticism of both the proposed theoretical framework and the method of untangling the mutual dependence among the presumably independent variables.

In contrast, factor analysis is used to determine the underlying determinants of many variables, without the need to postulate causality. This is especially useful because economic and social indicators are closely intertwined, making it nearly impossible to find a set of economic and social variables that are not correlated in some way. This interdependence makes normal regression analysis problematic but, surprisingly, does not adversely affect factor analysis.

Factor analysis, then, is a formal mathematical procedure that estimates the unobserved independent factors (or components) that characterize the various so-called independent variables. Because they can be constructed to be independent, the factor estimates (called factor scores) can be used in regression analysis to find the correlation between the factors and the dependent variable. The dependent variable may or may not be incorporated directly into the procedure that estimates the independent factors. If one is looking to estimate a missing observation for a dependent variable, such as per capita GDP for a particular country (as is the goal of this study) then this variable is not made part of the procedure.

The mathematical procedure is less likely to be influenced by the personal biases of the researcher because he or she is supposed to include as many variables as possible, allowing it to dictate how the variables are grouped together into factors. It is also usually found that a much smaller number of factors than variables is able to explain most of the variance of any of the variables included in the procedure, or later used as a dependent variable.

5.2 Weaknesses

There are still many shortcomings or limitations in research.

First, there are some shortcomings in model construction and support for variables in the study. Potential variables are the academic basis of PhD students and their research potential. Their effective measurement is an important part of the analysis of the reasons for the delayed completion of doctoral students. Due to the lack of necessary information, there is no research on the possible impact of this factor on graduation. Make direct estimates. In the study, although the number of meetings with the mentor were used as the time of advice, and various academic titles were used as indicators to measure the mentor's professional level, there was some rationality, but the service was simple and deviations may occur. Factors such as doctoral student fertility status, thesis opening time, mentor size, mentor status, and the work environment can also affect postgraduate studies. Due to data limitations, the study did not analyze these variables.


Secondly, the training of doctoral students under the tutor system has distinctive individual characteristics, and model-based interpretation is generalized. The mechanism of influence of different factors is still an issue that needs to be further explored in future research. Among the reasons for the delay, which are caused by factors in the education system, which are caused by other external factors, and the relationship between different factors, these are all factors that should be paid attention to when discussing the relationship between the factors and the doctoral students. Third, the research sample has limitations. The conclusions of the research are based on data from only one university. The relevant conclusions only partially verify some factors that affect the postgraduate completion of the doctoral degree. These conclusions are applicable to other doctoral education units. Sex has yet to be tested.

References

- [1] J. Wang and J. He, *Research on Tourism Economic Early Warning Model*. Springer Berlin Heidelberg, 2013.
- [2] J. Yan, J. Yang, and M. Weiyan, “Research on early warning model establish for electrical safety supervision system in china,” pp. 448–451, 2009.
- [3] H. Z. Wang, K. Liu, X. L. Shen, and Z. F. Tan, “Research on forecast and early-warning model of energy-economy-environment system,” in *First IITA International Joint Conference on Artificial Intelligence, Hainan Island, China, 25-26 April 2009*, 2009.

Appendices

Appendix A Code

```
1: main.py 
import numpy as np
import pandas as pd
np.random.uniform(0.85,0.90,(6,1))
index=np.arange(0,238)
columns=[]
columns_name=['gender','Is a freshman','Whether to start the question on
→ time','Undergraduate school level','Number of students supervised by
→ instructor','Average grades for degree courses','Number of SCI
→ articles published in three years of enrollment','Number of SCI
→ articles published within 4 years of enrollment','Number of EI
→ articles published three years after enrollment','Total articles
→ published in admissions']
x=pd.DataFrame(index=index,columns=columns_name,data=np.nan)
rand=np.random.randint(0,2,(238,1))
h=np.array(x['Is a freshman'].map({1:'male',0:'female'}))
rand=np.random.randint(0,2,(238,1))
y=np.random.randint(0,2,(238,1))
x['Is a freshman']=y
z=np.random.randint(0,2,(238,1))
x['Whether to start the question on time']=z
q=np.random.randint(0,3,(238,1))
x['Undergraduate school level']=q
n=np.random.randint(0,15,(238,1))
x['Number of students supervised by instructor']=n
x['Is a freshman']=np.array(x['Is a freshman'].map({1:'yes',0:'no'}))
```

```
x['Whether to start the question on time']=np.array(x['Whether to start
→ the question on time'].map({1:'yes',0:'no'}))
x['Undergraduate school level']=np.array(x['Undergraduate school
→ level'].map({1:'211',2:'985',0:'general'}))
x['Number of SCI articles published in three years of
→ enrollment']=np.random.randint(0,5,(238,1))
x['Number of EI articles published three years after
→ enrollment']=np.random.randint(0,5,(238,1))
x['Total articles published in admissions']=x['Number of SCI articles
→ published in three years of enrollment']+x['Number of EI articles
→ published three years after enrollment']+np.random.randint(0,2)
del x['Number of SCI articles published within 4 years of enrollment']
x['Average grades for degree courses']=np.random.normal(76,5,(238,1))
x['gender']=np.random.randint(0,2,(238,1))
x['gender']=x['gender'].map({1:'male',0:'female'})
x.describe()
x['Delay time']=np.round(np.abs(np.random.normal(0,2,(238,1))),1)
x.columns
x.to_csv('dsadsa2.csv',index=True,header=True)
y=x.iloc[:, :-1]
y.head(30)
y.groupby(by=['gender','Undergraduate school level','Is a
→ freshman','Whether to start the question on time']).median()
h=x.groupby(by=['gender','Undergraduate school level','Is a
→ freshman','Whether to start the question on time']).median()
```

Appendix B Data

Table 2: Data

	gender	Is a freshman	Whether to start the question on time	Undergraduate school level
0	male	no	yes	985
1	female	no	yes	211
2	female	yes	no	general
3	female	yes	no	211
4	male	no	yes	general
5	female	no	no	211
6	female	no	yes	211
7	male	no	no	211
8	female	no	no	985
9	male	no	no	general
10	male	no	yes	general
11	female	no	no	general
12	male	yes	yes	211
13	female	no	no	211
14	female	no	no	211
15	male	no	yes	985
16	female	no	no	211
17	male	yes	no	211
18	male	no	yes	general
19	male	no	no	general
20	female	no	yes	985
21	female	yes	no	general
22	female	yes	yes	211
23	male	no	no	general
24	male	no	no	211
25	female	no	no	985
26	female	no	yes	985
27	female	yes	yes	985
28	female	yes	no	211
29	male	no	no	general
30	female	yes	yes	general
31	male	no	no	985
32	female	yes	yes	general
33	male	no	no	general
34	female	no	yes	211
35	female	no	yes	211
36	female	yes	no	general
37	male	no	no	general
38	female	yes	no	985
39	female	no	yes	211
40	male	yes	yes	211
41	female	yes	no	general
42	female	yes	no	985
43	male	yes	yes	211
44	male	yes	no	general
45	male	no	yes	985
46	male	yes	yes	211
47	female	yes	no	985
48	female	no	yes	211