Analysis of UGA Students Advising Appointment on Student Advising and Guidance Expert

Authors: Lingyu Zhao, Zetong Jin, Jianwei Ren

Client: Dr. Julia Butler-Mayes, UGA Academic Advising Services

STAT 5020W

University of Georgia

05/10/2021

I. Introduction

Student Advising and Guidance Expert (SAGE) is a website for students to schedule their academic advisor appointments online and for advisors to track each student's success at the University of Georgia. With this feature, students can quickly go online and make an appointment with their academic advisor to explore their academic interests, identify resources for additional information and support, and develop appropriate study plans for their education goals. Thus, the UGA Academic Advising Service helps students with their academic and life problems through the SAGE system.

In this report, we aim to analyze the Covid-19 impact of the advising appointment service and the speed of each department responding to the students' questions. Those research questions are significant, which will offer our registering office a clear overview of the efficiency of each department and the impact of Covid-19 on our school. By analyzing the data, we will first use the data cleaning technique to clean the data and find the correlation between the variables. After that, we will use some statistical methods to show the results and answer the research questions.

Based on the conclusion for our analysis, we can see that each department has different efficiency on responding to the students' questions, and the advisors' working pattern has been changed. The time point of change in workload has been delayed in the year 2020.

The remainder of our report is organized as follows: In section II, we will describe the data and the data cleaning method. In section III, we will find the correlation between each variable. In section IV, we express our analysis methods, present the results and compare our findings with each plot. In section V, we state our conclusions and suggest possible topics for future research. Finally, in appendix, we put all of our programming codes and figures for reference.

II. <u>Data Description</u>

From the UGA Academic Advising department, Dr. Julia Butler-Mayes and Ashley Whitten offer the data, which has a separate Excel file containing different reports each semester from 2016 - 2020. As a complete advising process, students will schedule a meeting appointment with their advisor for questions. Then advisors will keep track of those questions by labeling them as tracking items, which all records are stored on the SAGE. Since one meeting could solve multiple questions, and one question may require several meetings to be solved, meeting records and tracking item records are stored separately. Advisors keep track of those tracking items by labeling them as "Raised" or "Cleared," which means the question has been brought up or finished. Therefore, our data is all of the meeting details and tracking item details with a start and end time, advisor's name, meeting details or tracking item details, and reasons for raising and closing with over 50,000 meetings and 60,000 tracking items in two separate Excel files per semester. Since the meeting format in 2020 is entirely different, we paid more attention to the tracking reports.

We cleaned the data and created a new Excel by counting the number of "Raised" and "Cleared" items per week and day, then re-organized by the calendar years. For example, there are usually 18 or 19 weeks in the Spring and Fall and 12 weeks in the Summer. However, since SAGE is still relatively new that just started running in 2016, we had many periods of missing values or inaccurate data, especially in 2016 and 2017.

In addition, we made an Excel about the speed of each department responding to the students' questions. In this Excel, we focus on each UGA Advisor. First, we found the number of "Total raised" and "Total cleared" of each advisor through the filtering function in Excel. From these two variables, we calculated the ratio between them. Then we focus on which departments these advisors belong to. Since the advisors' department was not given in the data, we Googled each advisor to determine which college they belonged to. In this Excel, we used the data of 2017, 2018 and 2019.

Item Name	Trigger	Raise Date	Clear Date	Days to Clear	Raiser Name	Clearer Name	Status
EC Intended Business	Manual	2017/2/6	2018/4/30	448	AS	GA	Cleared

Table 1. data dictionary table

III. Exploratory Data Analysis

From our data, we have found the details of various advising appointments in the Tracking Report. This analysis will figure out the relationship between "Raised" and "Cleared" items on assignments. We extracted data from the Fall semester of 2017 through the Spring semester of 2019. Due to the small number of summer classes and short study cycles, many students do not attend advising appointments. At the same time, there will be many online courses and students from other universities. These are going to affect our analysis to a large extent. Thus, we did not consider using data in the summer semester.

After importing collated data from our dataset, we started to figure out the correlation between the "Raised" and "Cleared" for each semester. Firstly, we use the Pearson correlation test to see if those two variables are highly correlated or not. Then, we used scatter plots to visualize our data. As shown in Figure 1, based on the definition of scatter plots, we can see that the lines are linear in most cases, but we notice that the line might change its direction at some point. Thus, we went back to the Tracking Report and found out that advisors could clear a case between zero days to three years, which means an item that cleared in one semester could be raised many years ago. However, we can still see a clear linear trend and conclude a correlation between "Raised" and "Cleared" in Fall 2017, Spring 2017, Spring 2018, and Spring 2019. Based on what we found, we decided to analyze the raised and cleared items separately.

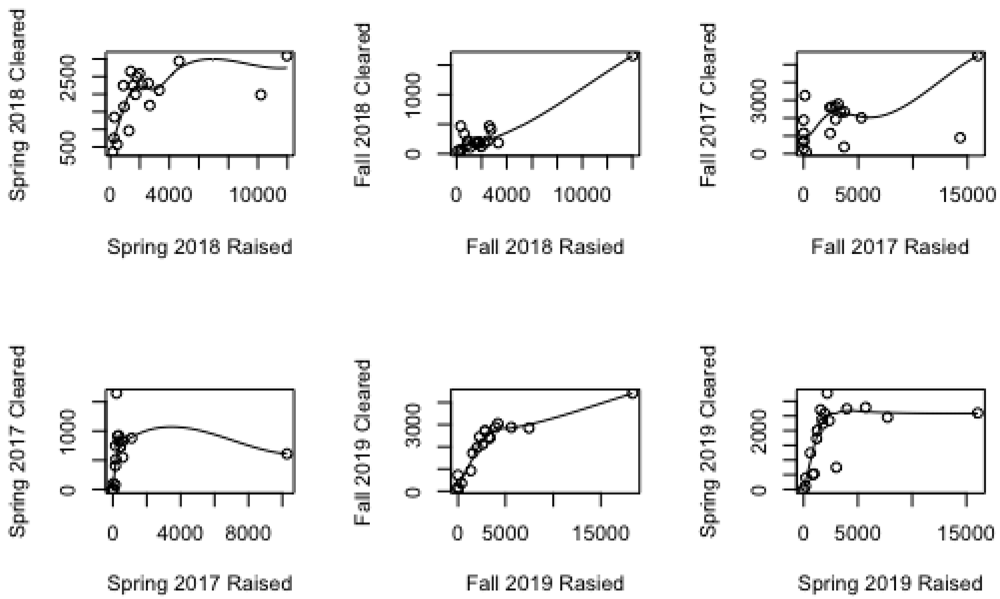

Figure 1. Scatter Plot of Each Semester

IV. <u>Methods and Result</u>

<u>Covid-19 Impact</u>

Covid-19, as a worldwide epidemic pandemic, has dramatically influenced our life at this particular time in history. The patient population increased exponentially in the United States starting January 2020. Responding to the situation, the University of Georgia began online class after the spring break in March 2020 and resumed part of the in-person class in August as the new semester started. Due to the enormous impact and it will not disappear in the short term, we decided to focus on the effect of Covid-19 on UGA's academic advising service by comparing the data from the calendar year 2019 and 2020.

From the new data file that we created, it is clear that advisors are usually very busy after the new semester starts and reach the maximum within a month, and then the workload will eventually slow down until the semester ends. Thus, we will use the Anderson-Darling test and change point determination to solve this research question. Anderson-Darling test is the test for proving if two samples are from the same distribution or the same parent distribution. Changepoint determination in time series could detect the specific time that if an event has significantly changed, which is there, we are looking for the workload shift points. There are multiple factors to determine the change point, and we used two kinds of methods to detect the change point. The first one uses the change point algorithm PELT with a nonparametric cost function based on the empirical distribution to find multiple changes in a time. The second is to use the difference of change in mean to identify one time point that the advisors' workload has decreased per semester. The R package, kSample, changepoint, and changepoint.np, are used in this research question, which corresponds to the two statistical methods used.

We used R commend dygraph for seeing the general trends of raised and cleared cases in 2019 and 2020 by sketch scatter plots shown in Figure 3 below. In both plots, the x is the number of weeks, which 0 is the start of spring semesters and 20 is the start of fall semesters. The y is the number of cases each week, which raised items is on the left, and cleared items are on the right. 2019 is in blue, and 2020 is in green. In the raised items plot, two years are pretty similar during the spring semester. But in the fall semester, advisors usually raised questions until the 38th week, but they finished earlier on the 28th week in 2020 with the number of questions doubled compared to the previous year. In the cleared items plot, advisors were as usual until the 10th week, which is the Spring Break week. After that, advisors cleared all of the questions during the week after Spring Break while working from home in 2020 and didn't resume until the 20th week, which is the fall semester. The maximum number of cleared items per week in 2019 is 4500, but in 2020 the maximum was almost 8000. The last local maximum is during the week after the withdrawal deadline of 2020. Therefore, the local maximum in the raised items during the 28th week might be because students had questions about withdrawing from their classes, and advisors cleared them after students made their decisions.

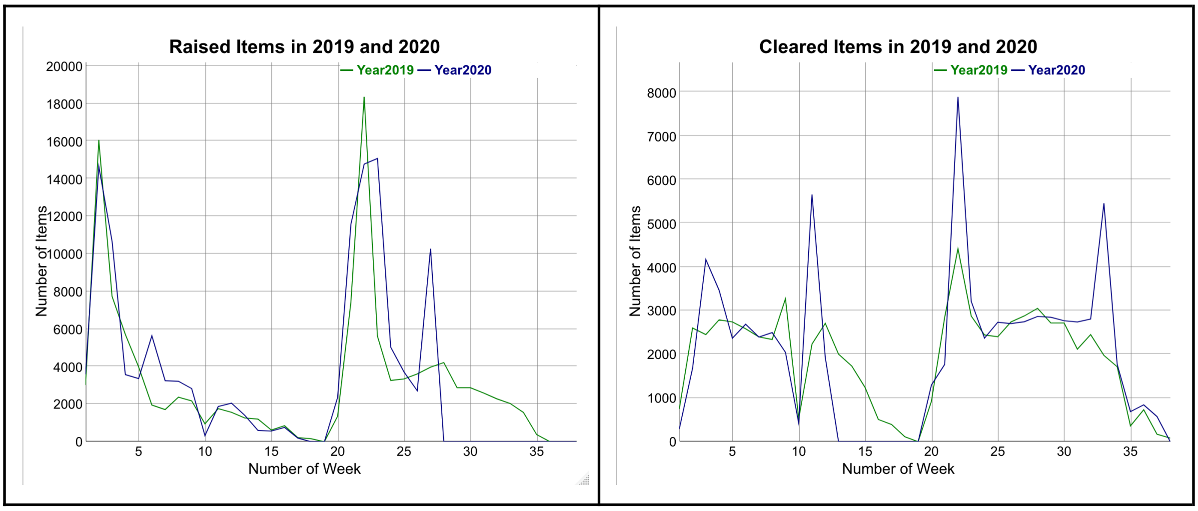

Figure 3. Comparison of "Raised" and "Cleared" between 2019 and 2020

Then, we tested the distributions of Raise and Cleared items in 2019 and 2020, conducting an Anderson-Darling test. As a result, we got p-values of 0.1982 and 0.1927 for the cleared items and 0.05400 and 0.05284 for the raised items. The Anderson-Darling test proved that the distributions in 2019 and 2020 are neither the same distribution nor from the same parent distribution for both raise and cleared items as having all p-values greater than 0.05. Advisors' workload was not as same as the previous year in 2020, which we believe is because of the Covid-19, school schedule, and class format changes.

Since they are in different distributions, we used the changepoint algorithm PELT based on the empirical distribution each year with R package changepoint.np. The result is in the table shown below, and the data was organized by week. Change points in 2019 for raised items are the 5th, 20th, 22th and 33rd week; the cleared items change points are 13th, 20th, and 32nd week. Change points in 2020 are the 9th, 20th, and 27th week for the raised items; for the cleared items are the 12th, 20th, and 34th week. Each time point means the advisors' workload has shifted during this week. If it was normal and no pandemic going on, those time points should be the same for each year. However, among all change points in the number of weeks, only week 20 is consistent, which is the first week of the fall semester. The change point before 20 is the change point during the spring semester, and after 20 are the change points during the fall semester in the corresponding years. Since we got two change points, we could not directly compare the change points for raised items during the fall semester. But except that only cleared items during the spring semester shows an earlier change point in 2019. For a closer look at the difference in day and want to have one changepoint per situation, another detection is placed with different methods.

	Change Points in 2019	Change Points in 2020
Raised Items	5, 20, 22, 33	9, 20, 27
Cleared Items	13, 20, 32	12, 20, 34

Table 2. Change points in number of weeks

After calculating the number of items per day, we used the R package changepoint to detect the changepoint by changing the mean. In this way, we can control our changepoints to be one per semester, which the plot is shown in figure 4. As a result, for the spring semesters, the change point of the raised items altered to the 17th day in 2019 and 19th day in 2020; the cleared items change point was the 92nd day in 2019 and the 87th day in 2020. For the fall semester, the raised items' change point was the 23rd day in 2019 and 47th day in 2020; the cleared items' change point was the 87th day in 2019 and the 92nd day in 2020. The change points are different since we use different detection methods, but the results are consistent, speaking of which change point is earlier in those two years. In four pairs of comparable change points, only one has an earlier change point in the year 2020, which is still the spring semester cleared items as we found in the other method. All of the other pairs indicate that year 2020's workload decreased later in 2002 compared to 2019, especially the fall semester raised items which shows the most significant delay that the workload decreasing point for advisors was almost a month late in 2020.

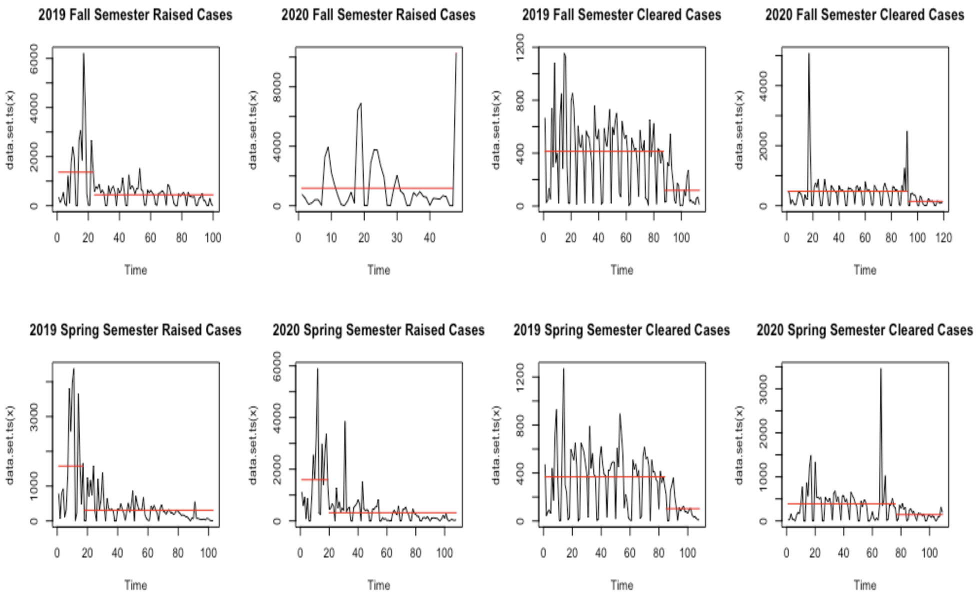

Figure 4. Change Point by Day Plots

<u>Department Efficiency</u>

For our second research question, we plan to compare the ratio of cleared and raised cases with each department to see each department's efficiency in responding to their students. We are using the number of cleared items divided by the number of raised items for each advisor. Then group the advisors by colleges and take the mean of the ratios as the department's efficiency ratio on responding to student's questions. Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more data groups to gain information about the relationship between the dependent and independent variables. Using the ANOVA table and the statistical significance, we can see the pattern and know the results. Initially, our data only consisted of each advisors' name, but we do not know which department they belong to. Thus, we plan to use Python to ask a web page to run the search for us. However, we found that some advisors might have retired or changed their job, so we cannot find their names on the UGA website. Since this technique can only search for one web page, many advisors' names cannot be found on the UGA website. Therefore, we decided to eliminate this technique and search the advisors' names manually. After seeing the advisors' department, we notice that we have one unknown category, which means we cannot find their name online. We notice that there is no way for us to figure out the title since the data imputation method does not have the accuracy on the unknown department. Thus, we decide to delete the unknown category.

We are using the ANOVA table to test the differences between the two groups of experiments. Based on the Anova table, we can see that the p-value is 0.0076, which is less than the critical value of 0.05. Thus, we can conclude that the efficiency of each department is different. Then, we are trying to compare the mean of each department and determine which department has the fastest efficiency. Using the ggplot shows that the mean ratio of "Cleared" and "Raised" of each college and the bars with different heights show different mean ratios, and the Warnell School of Forestry and Natural Resources has the highest proportion.

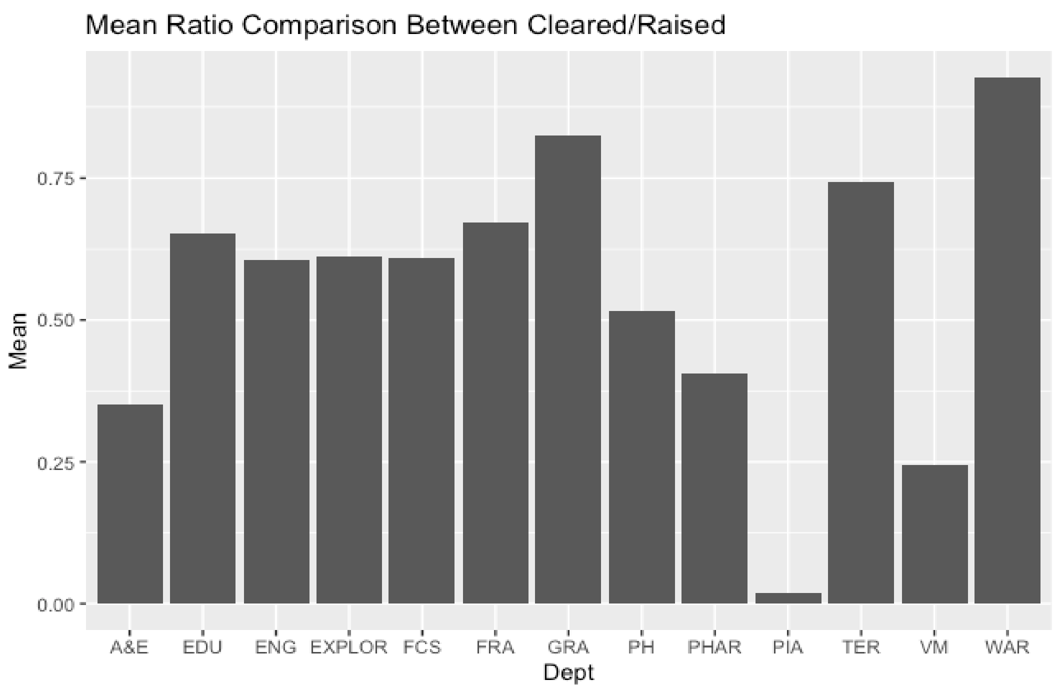

Figure 5. Mean Ratio for Each College at UGA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Department	13	7.3	0.5612	2.289	0.00618 **
Residuals	460	112.8	0.2452		

Table 3. ANOVA Table

Now, we are using the Tukey Method to compare the significance of two departments and see if they are significantly different or not. Tukey method is a single-step multiple comparison procedure and statistical test. It is helpful to find means that are substantially different from each other. Since we have lots of value to compare, we cannot see the pattern in the figure. However, by looking at the Tukey method plot, we can see that the Grady and Agricultural & Environmental Sciences departments and the Franklin and Agricultural & Environmental Sciences are significantly different, showing that those two groups of colleges have different efficiency in solving the questions.

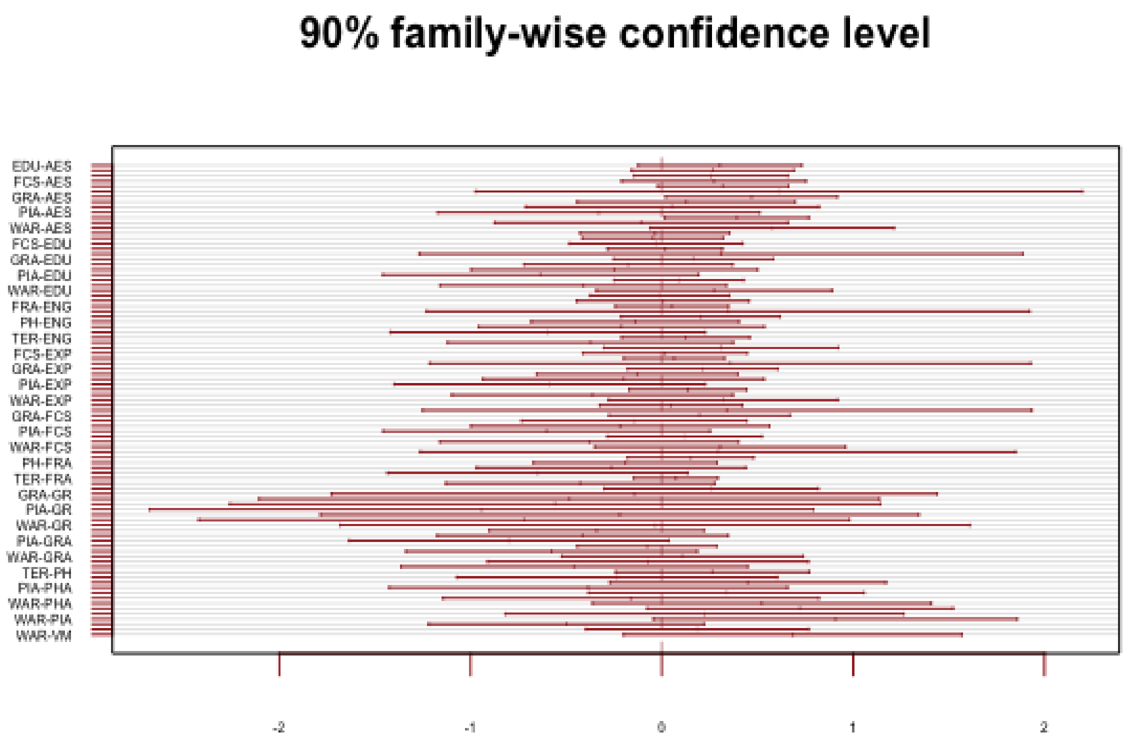

Figure 6. Tukey Method Plot

V. **<u>Conclusion and Recommendation</u>**

From the observation from previous years, advisors always have a significant workload as the semester starts and reaches the maximum within a month, eventually decreasing until the semester ends. It is reasonable that students might have more questions because they are not familiar with the class or school, and the number of questions will decrease as they fit into the new schedule. Therefore, this working pattern will continue in the future unless we change the semester format entirely.

Compared to 2019 and 2020, the number of raised items has increased by 4.23% and broke the highest record in the past years. Even though the number of raised items has been rising in recent years, we still believe that Covid-19 and remote learning has increased the number of questions from students as they raise and solve a lot more questions during the 28th week and the 33rd week in 2020, which is a month before and a week after the withdrawal deadline. In addition, students ask more questions to their advisors around the unique schedules, such as during the add and drop period and before the withdrawal deadline, but more minor before the summer, winter, and spring breaks.

From this analysis, we believe that Covid-19 has delayed the advising process as having most change points later in 2020. According to the change point by day plots, all the 2020 plots have a bimodal shape where 2019 plots are all unimodal. This indicated advisors had another big wave of workload in 2020, which does not show in 2019. If we continue online classes in the following semester, the changepoints should be more similar to 2020, and advisors could use those changepoints as a reference when scheduling their works.

As for the second research question, our results show that each college handles cases with varying efficiency, which is reasonable. It is not because the advisors from different colleges have different working speeds, but many factors influence days to clear for questions from different colleges. For example, the student population, number of advisors in each department, steps to solve a question will all cause the efficiency to be different. Franklin does not require additional applications, but both Terry and Grady have another round of admission. Another critical influencer is sometimes departments collaborate. One item could be raised by an advisor but cleared by another advisor from another department, which we did not include those cases in the analysis. By seeing the mean ratio for each college at UGA, the Warnell School of Forestry and Natural Resources has the highest percentage, which means it is the college with the highest efficiency and solves questions in the shortest time. Based on the result for the Tukey Method, we can conclude that the two groups of colleges have different efficiency on solving the questions, which are Grady compares with Agricultural & Environmental Sciences and Franklin compares with Agricultural & Environmental Sciences.

In the beginning, we thought we could use this method to find the number of students who changed their major because of the Exploratory Center by checking if the question was raised by an Exploratory Center advisor but solved by another advisor. But in the actual process, we were still not able to identify the number of major changes. For example, we can not split the students who want to apply to business school initially or because of Exploratory Center events. Therefore, a direct major change from the Exploratory Center is needed to test the Exploratory Center's efficiency. A survey after each Exploratory Center's events could also help solve this problem.

In addition to our reports, we did a general analysis of our advising meeting from 2017-2019, when we explored the data sets. They contain important information but irrelevant to our research questions, therefore we put this part in the discussion. Most meetings are scheduled one-to-one meetings hosted in advisors’ offices in the morning that last 32.8 minutes on average, in which the supporting plots and tables are included in the appendix. Friday has significantly fewer meeting appointments compared to the other weekdays. We do have meetings during the weekend, but it is rare and less than ten meetings per semester. Days to clear for reasons vary in each semester, but all applications to college or program usually take a longer time. Those reasons all have an extensive range which could be 0 to over 1000 days and the reason cause the gap is unknown. The plot of average days to clear for 2020 fall is included in the appendix as an example. A close reason or a situation description would be beneficial for the future analysis if needed and help us find the factor that causes a longer time to clear.

From the three year's data, most students attended their meeting as scheduled, which is shown in figure A2. Overall, less than 4.5% of the students will miss their advising meeting each semester. Even though the percentage is small, an SMS texting feature for reminding students for their advising meeting is still needed due to the enormous number of meetings. For testing the efficiency of the SMS text message system, the current record is not enough. We don't know neither how many students saw the message nor if the student who received the message attended their advising meeting. One of a plan to test the efficiency could be to conduct a chi-square homogenous test. We will need to collect if the student wants to receive the texting message with the other meeting details. Then we will have four situations:

- The student received the message and attended the meeting.
- The student received but did not attend the meeting.
- The student attended the meeting but did not receive the message.
- The student did not attend the meeting or received a message.

We will be able to identify the efficiency of the reminding message by comparing the percentages in different situations. In addition, if we can let each advisor ask if the student thinks the feature is functional, that would also be good information to have in the analysis process.

VI. Appendix

Figure and Tables

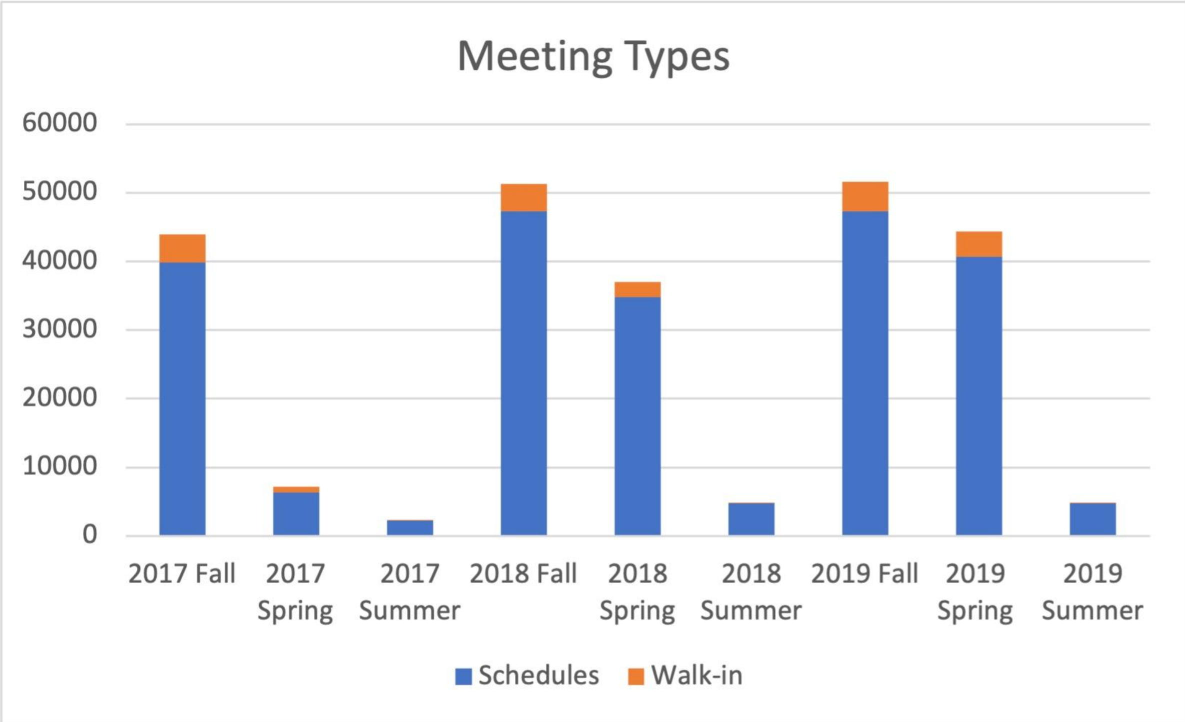

Figure A.1. Meeting Types Summary 2016-2019

Figure A.2. Attended Session Summary 2016-2019

Location	Percentage
ELSEWHERE	0.37%
OFFICE	99.04%
ONLINE	0.26%
PHONE	0.33%

Table A.1. Meeting Location Summary 2016-2019

Schedule Block Type	Percentage
One-on-One	98.74%
Group	1.26%

Table A.2. Schedule Block Type Summary 2016-2019

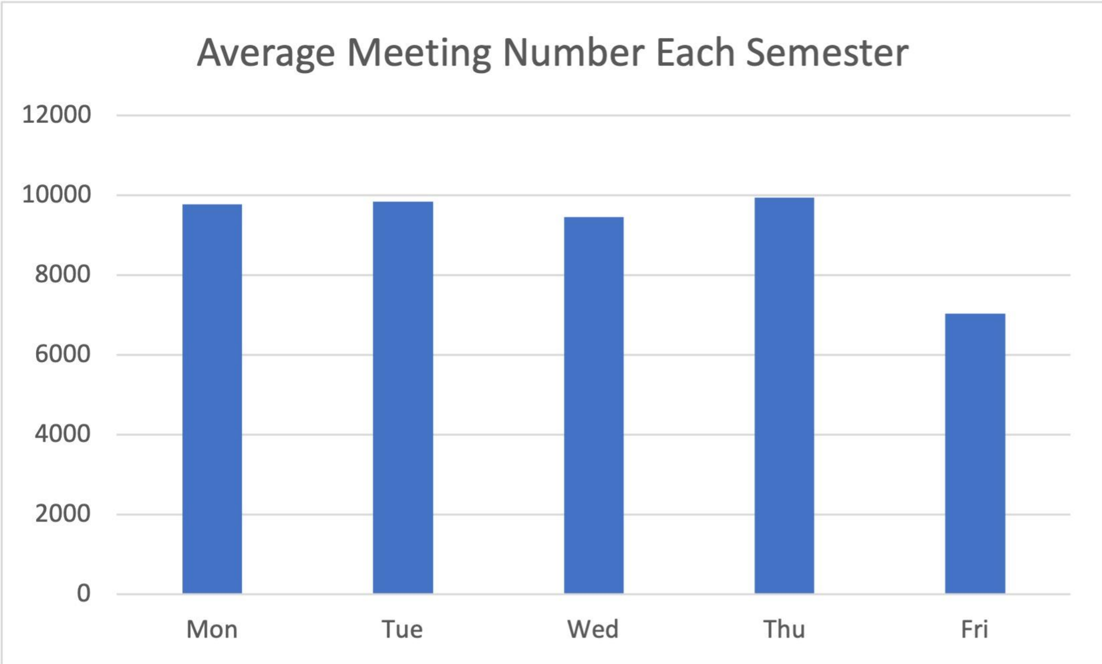

Figure A.5. Average Meeting Numbers for All Semesters

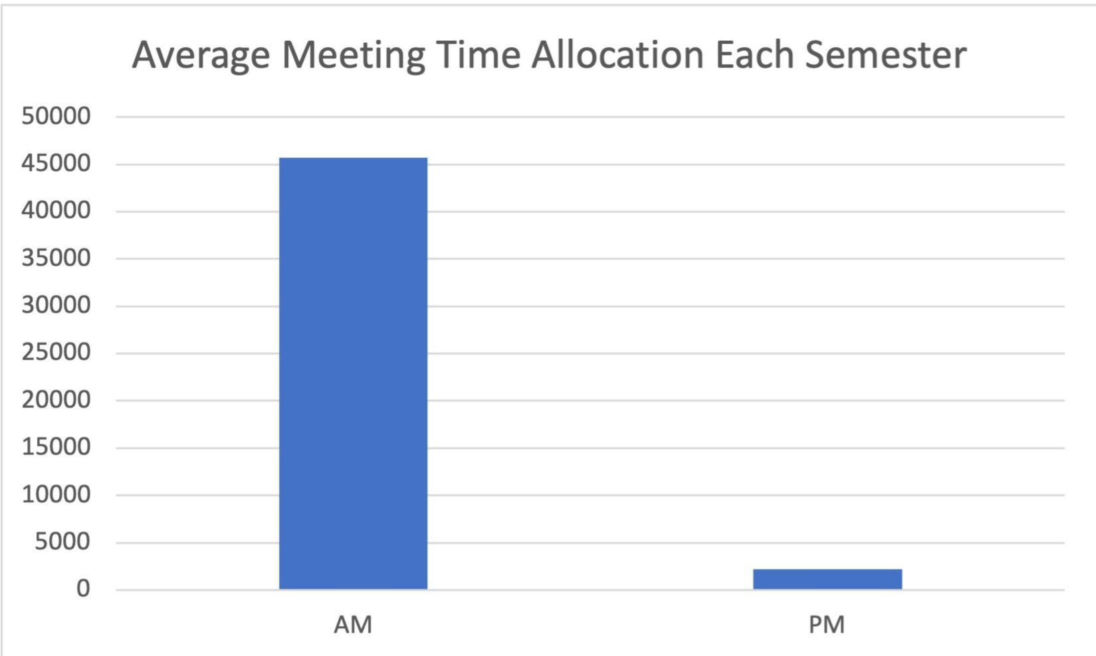

Figure A.6. Average Meeting Time Allocation for All Semesters

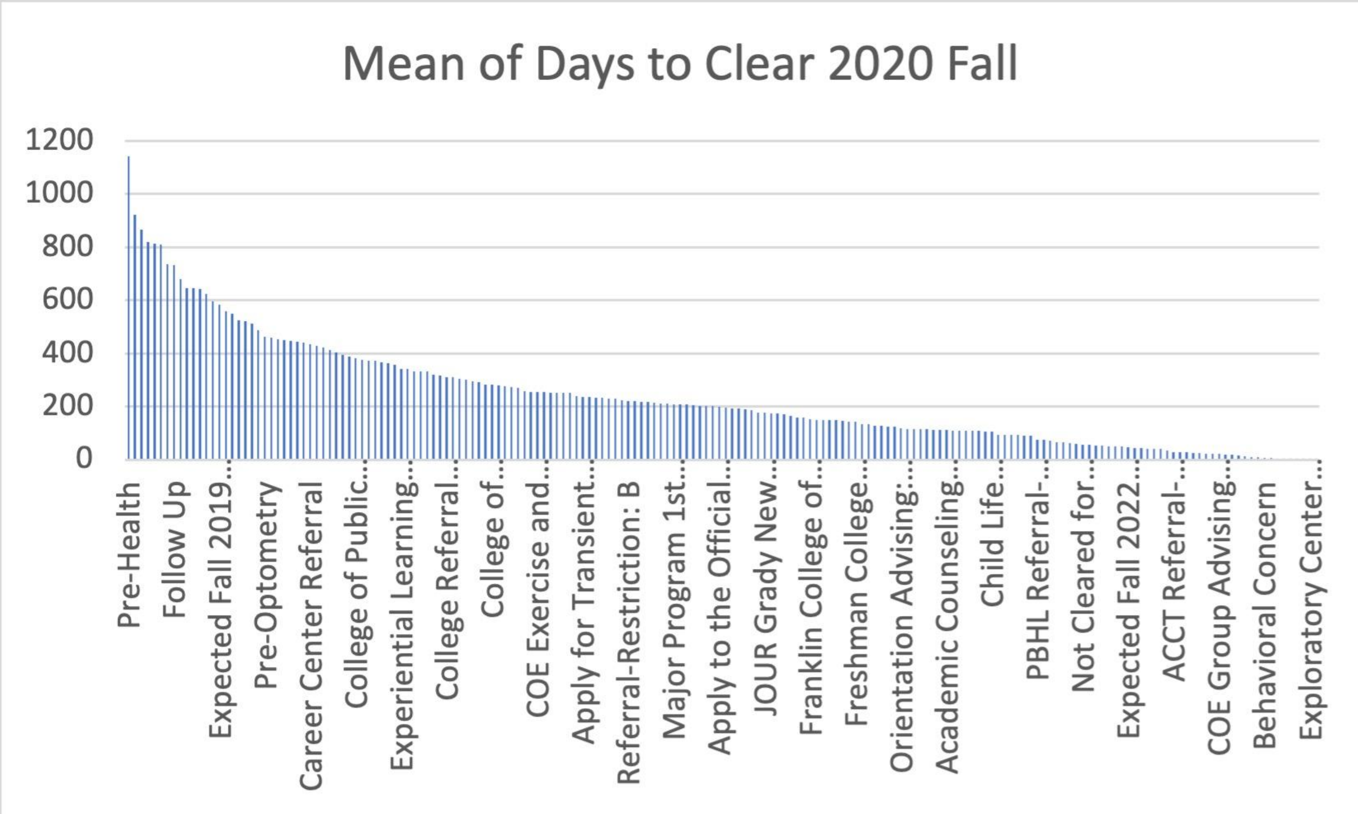

Figure A.7. Mean of Days to Clear in 2020 Fall Semester

R code

```
install.packages("xlsx")

library(xlsx)

data = read.xlsx("/Users/Thomas/Desktop/Projectdata.xlsx",header = T, 1)

head(data,20)

cor.test(data$Spring.2017.Raised,data$Spring.2017.Cleared,method = c("pearson"))

cor.test(data$Fall.2017.Raised,data$Fall.2017.Cleared,method = c("pearson"))

cor.test(data$Fall.2018.Raised,data$Fall.2018.Cleared,method = c("pearson") )

cor.test(data$Spring.2018.Rasied,data$Spring.2018.Cleared,method = c("pearson"))

cor.test(data$Fall.2019.Rasied,data$Fall.2019.Cleared,method = c("pearson"))

cor.test(data$Spring.2019.Raised,data$Spring.2019.Cleared,method = c("pearson"))
```

```
par(mfrow=c(2,3))
scatter.smooth(x=data$Spring.2018.Rasied, y=data$Spring.2018.Cleared)
scatter.smooth(x=data$Fall.2018.Raised, y=data$Fall.2018.Cleared)
scatter.smooth(x = data$Fall.2017.Raised, y=data$Fall.2017.Cleared)
scatter.smooth(x = data$Spring.2017.Raised, y = data$Spring.2017.Cleared)
scatter.smooth(x= data$Fall.2019.Rasied, y = data$Fall.2019.Cleared)
scatter.smooth(x = data$Spring.2019.Raised, y = data$Spring.2019.Cleared)
ccf(x = ts(data$Spring.2018.Rasied) , y = ts(data$Spring.2018.Cleared),type =
"correlation",main = "Spring 2018")
ccf(x = ts(data$Fall.2018.Raised) , y = ts(data$Fall.2018.Cleared),type = "correlation",main
= "Fall 2018")
ccf(x = ts(data$Fall.2017.Raised) , y = ts(data$Fall.2017.Cleared),type = "correlation",main
= "Fall 2017")
ccf(x = ts(data$Spring.2017.Raised) , y = ts(data$Spring.2017.Cleared),type = "correlation",
main = "Spring 2017")
ccf(x = ts(data$Fall.2019.Rasied) , y = ts(data$Fall.2019.Cleared),type = "correlation", main
= "Fall 2019")
ccf(x = ts(data$Spring.2019.Raised) , y = ts(data$Spring.2019.Cleared),type = "correlation",
main = "Spring 2019")



head(cleared)
library(dygraphs)
data <- data.frame(
  time=c(1:38),
```

```
  Year2019=cleard$r19,
  Year2020=cleard$r20
)
p <- dygraph(data,main="Compareson of Raised Cases in 2019 and 2020")
p

data2 <- data.frame(
  time=c(1:38),
  Year19=cleard$c19,
  Year20=cleard$c20
)
p2 <- dygraph(data2,main="Compareson of Cleard Cases in 2019 and 2020")
p2

library(changepoint.np)
changepoint.np::cpt.np(cleard$r18)@cpts
changepoint.np::cpt.np(cleard$r19)@cpts
changepoint.np::cpt.np(cleard$r20)@cpts
changepoint.np::cpt.np(cleard$c18)@cpts
changepoint.np::cpt.np(cleard$c19)@cpts
changepoint.np::cpt.np(cleard$c20)@cpts

library(zoo)
library(changepoint)
springcp18=as.vector(table((cpbyday$`18 spring`)))
```

```
cps18 = cpt.mean(springcp18)
c(ints = param.est(cps18)$mean,
  cp = cpts(cps18))
plot(cps18)
springcp20=as.vector(table((cpbyday$`20 spring`)))
cps20 = cpt.mean(springcp20)
c(ints = param.est(cps20)$mean,
  cp = cpts(cps20))
plot(cps20)
library(kSamples)
ad.test(cleard$r19,cleard$r20)
ad.test(cleard$c19,cleard$c20)


library(readxl)
Sage = read_excel("/Users/Thomas/Desktop/Overall.xlsx")
attach(Sage)
Sage = data.frame(Sage$Ratio,Sage$Department)
Terry = mean(Sage[Department=="Terry",1])
Franklin = mean(Sage[Department=="Franklin",1])
AES = mean(Sage[Department=="Agricultural & Environmental Sciences",1])
Education = mean(Sage[Department=="Education",1])
Engineering = mean(Sage[Department=="Engineering",1])
Exploratory = mean(Sage[Department=="Exploratory",1])
FCS = mean(Sage[Department=="Family and Consumer Sciences",1])
Grady = mean(Sage[Department=="Grady",1])
```

```
Pharmacy = mean(Sage[Department=="Pharmacy",1])

PIA = mean(Sage[Department=="Public & International Affairs",1])

PH = mean(Sage[Department=="Public Health",1])

VM = mean(Sage[Department=="Veterinary Medicine",1])

Warnell = mean(Sage[Department=="Warnell",1])

g = aov(Sage$Sage.Ratio~Sage$Sage.Department)

summary(g)

#Efficiency different

TUKEY<-TukeyHSD(g,conf.level = 0.90)

plot(TUKEY , las=1 , col="brown")

ggplot(Mean,aes(x = Dept, y =Mean))+geom_col()+ggtitle("Mean Ratio Comparison
Between Cleared/Raised")
```