College Student Pandemic Prediction

Emil Matti

Wagner College

Abstract

Data science is the process with programming while using mathematical and statistical methods to pull meaningful insight from data. From recommended music playlists to sports statistics, data science/econometric modeling are playing a significant role in human society today. Linear regression is one of the simplest and most commonly used statistical techniques and algorithms applied in the real-world. Using data science and machine learning algorithms help the world make better decisions based on the computer's findings. I have connected my econometric and machine learning background to implement a linear regression and exploratory data analysis on the decision of a student during the current pandemic. Data was collected through an electronic survey on google form and was sent out through the Wagner College student mailing list. SPSS was used to run the regression and collect statistical data on the variables used. Python, which is a high level programming language, was used to clean the data, provide visualizations and apply machine learning through a series of libraries within python. The regression run on the dataset and testing dataset achieved scores ranging from 55-59.9%. This end-to-end experiment provides the ease of understanding of modern analytics on a real-world problem.

I. Introduction

Today's world is driven by data. Data is vital and necessary for us as we use it to operate in our everyday lives. An immense amount of data is being recorded now than it has ever been, therefore it is the only way to make decisions. For example, sports organizations use their recorded data to optimize their team's performance, companies show their growth in revenue and social media platforms collect users' data to learn more about their interests. In relation to current events, it is apparent that we use data in the crisis of the pandemic in 2020 to present day. Throughout the course of this pandemic, states and governors are basing their decisions on the information they receive on positive cases. Hospitals and testing sites are constantly recording the amount of deaths and positive cases caused by COVID-19. These numbers are very important to us as a country because it tells us whether to make restrictions stricter to stop the spread or show us that there is a decline in deaths and cases which will allow us to start returning back to normalcy. Because of the current pandemic, it affected everyone's work life forcing people to work from their homes. A controversial topic surrounds schools having to learn remotely while kids lose their sense of youth and social skills. More specifically, students are frustrated with the idea of learning remotely as some find going to campus with a large amount of people quite daunting. In this project, we will explore a college student when faced with a decision to come back to campus in the Fall of 2020 or stay home and attend courses online.

This econometric project will help predict whether or not a student will attend courses on campus or fully remote. An issue that arose on college campuses during the process of bringing their students to campus was spacing. In a time where people must be socially distanced to stop the spread of the disease, it was imperative that colleges had space on their campus. Universities had to limit class seating, organize living situations, limit indoor dining accesses and many more accommodations to make the environment safe for all students and faculty members. In New York City, laws and regulations were constantly changing during the summer of 2020 leading up to move-in day for most colleges. Universities had to adjust the procedure of getting students back to campus accordingly to the law of the state they are located in. With all the troubles universities must deal with in order to get it up and running, it would be ideal to get a sense of who will come back to campus. This number will better prepare the university to set up living, class and dining accommodations for students to be safely socially distanced. The objective of this project is to add value to Wagner College and come up with a model that will accurately predict a student's decision based on certain variables. As this being a start, hopefully this model will be valuable and applicable to other universities in time of a pandemic.

Since this is an end to end data science/econometric project, collecting the data is a key step in this process. This project is to solely add value to Wagner College, therefore, the data is collected within the University. To collect the data, a survey on google form was sent out to the student mailing list and also sent out personally. The survey was enclosed to Wagner email addresses only which means no other email addresses that do not end in 'wagner.edu', will not be able to access the google form. The survey consisted of thirteen questions including an informed consent form which verifies that the participant is eighteen years old, verifies that they consent and informs them that their responses are anonymous. The length of the survey was taken into consideration due to college students preferring to spend less time on it and convenience of an online survey. The thirteen questions consist of 'yes' or 'no' answers or dropdown checkboxes. It was necessary to limit the possibility of various responses because it would be an easier dataset to clean in order to run a statistical analysis. After the survey was sent to the student population, responses were recorded from google forms and linked to google sheets. From google sheets, the dataset was converted into a CSV file and transferred over to python for cleaning and a statistical analysis on IBM's SPSS. Lastly, the dataset will make its way back to python where there will be visualizations and applied machine learning.

II. Literature review

The pandemic brought many emotions out of people and were forced to make tough decisions for their livelihood. This deadly disease dates back to December of 2019 and roots to Wuhan, China. Studies show that hospitals in Wuhan reported cases of severe pneumonia as it was unknown to what had caused it (Singhal, 2020). According to Singhal (2020) , coronaviruses are RNA viruses that have spike-like crowns and range from 60 nm to 140 nm in diameter. The four different corona viruses named HKU1, NL63, 229E and OC43 generally cause disease in the respiratory system and have been circulating in humans. The initial cases of these unknown pneumonia, show evidence that it comes from the seafood markets where they traded live animals. The World Health Organization was notified on December 31, 2019 by China that there was an outbreak from Wuhan. Singhal (2020) established in late January of 2020, cases were rising at a rapid rate in countries outside of China where the cases suggested that it was human to human transmission. From there cases kept rising as many countries started to shut down and ordered people to stay at home.

Supporting and helping people's physical health against Covid-19 was only half the battle. The disease caused many countries to shut down which meant businesses went bankrupt, workers had to work from home, people lost their jobs and many more. With this economic crisis at the hands of Covid-19, people's mental health was at risk. It is apparent to see that it is human nature to struggle mentally when being isolated with minimum social interaction. A study was performed and showed the immediate stress outcomes of patients who were infected MERS and put into quarantine (Shah, Kamrai, & Mekala, 2020). The results show that families who were affected by the disease claimed that the general public avoided them which made them feel isolated. When a human is being pushed away and looked down upon because of being sick, it can make them feel quite lonely. In relation to Covid-19, having the virus was scary not only for your physical health but also for the way that people will look at you. As the media outlets and social media show people the risks of receiving the virus, the more scared and anxious people feel toward the virus. In result, people have turned their back against their own family and friends because of them getting infected. Overall, people's thoughts have changed due to this disease and have given tough situations and decisions.

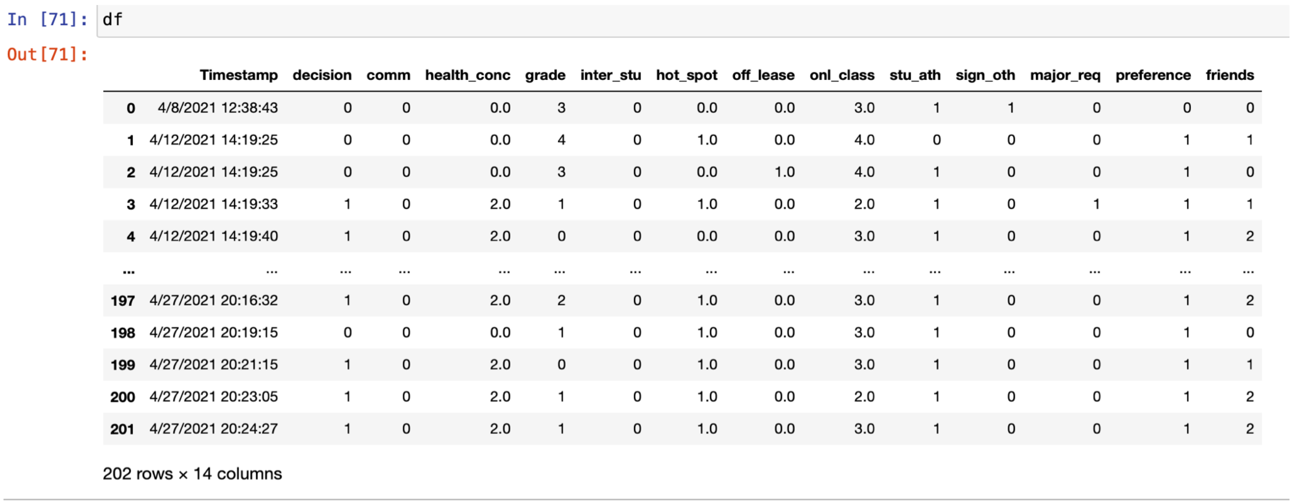
More specifically, the country's shutdown has affected businesses as some couldn't produce enough revenue and stay afloat. One business that may not strike you as a business are universities. Schools at all levels were affected by the pandemic as classes were pushed from onsite to online. However, universities all over the country were fighting the struggle as they were losing money by not having students on campus. In the summer of 2020, universities had to brainstorm and strategically come up with ways to bring their students back to campus in a safe manner. Bradley, An and Fox's study concluded that to open colleges up safely, colleges must test frequently to exceed guidelines from the US Centers for Disease Control and Prevention (Bradley, An & Fox, 2020). At Wagner College, students must take a PCR test once a week and report any new symptoms daily via a mobile app. These mandatory testing for students will ensure parents and students safety and awareness of what is going on within the college.

On the other hand, students are put in a situation where they must decide whether or not they will attend class on campus or stay home and take classes online. Due to the current pandemic, universities gave students the option to stay at home for their safety. The decisions of the students are weighed heavily by many factors. In a study conducted by Tumen Akyildiz, he picked 12 undergraduate students randomly to analyze their views on COVID-19 (Akyildiz, 2020). The study shows that 10 of the 12 students feared getting infected while 9 students claimed that they feared losing relatives due to the virus. Alyildiz (2020) also showed that 7 students reported being hopeless on economic issues. However, 9 students agreed that they were lacking social activity during the pandemic. It is evident in this study that there are many factors that play into a young college student's view on COVID-19. These factors translate into their decision when it comes to coming back to campus in the fall of 2020 where cases were still rising.

One of the biggest obstacles that students and Wagner College faced was the state travel restrictions that were placed by the governor of New York, Governor Cuomo. Starting on June 24th of 2020, governors of Connecticut, New Jersey and New York announced that people travelling to their state must quarantine for 14 days. However, these restrictions only applied to 8 states that had an average number of infections of 100,000 residents. Throughout the whole summer, Governor Cuomo began to add states to the restriction list while also removing states. Wagner declared that the first day of classes will begin on August 24, 2020 and required a mandatory quarantine in New York, New Jersey or Connecticut 14 days prior to the first day of classes. Along with the 14 day quarantine, the student must show a negative PCR test upon arrival to quarantine. With this mandatory quarantine rule for students, students planned to arrive in the tristate two or more weeks prior to the first day of classes. Therefore, most students had to start their quarantine at the latest, August 10th. On August 10th, Governor Cuomo (New York), Governor Murphy (New Jersey), and Governor Lamont (Connecticut) had a total of 35 states on the joint travel advisory that required travellers to quarantine for 14 days. The states on the out-of-state quarantine list included: Alabama, Arkansas, Arizona, California, Florida, Georgia, Iowa, Idaho, Kansas, Louisiana, Mississippi, North Carolina, Nevada, Oklahoma, South Carolina, Tennessee, Texas, Utah, New Mexico, Ohio, Wisconsin, Alaska, Rhode Island, Indiana, Maryland, Missouri, Montana, North Dakota, Nebraska, Virginia, Washington, Illinois, Kentucky, Minnesota, and Puerto Rico (Ballotpedia, 2021). This played a big role in a student's decision to come back to campus because there was an immense amount of planning and expenses in order to quarantine if in one of those hotspot states.

III. Data and Exploratory Data Analysis

Figure 1



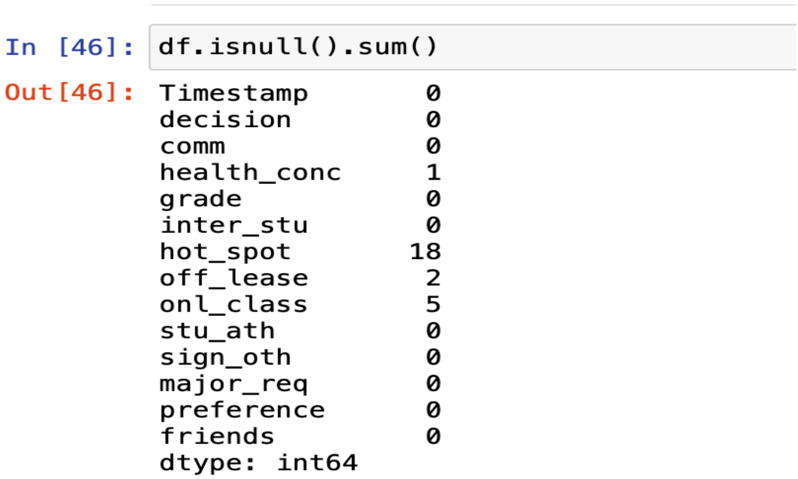
```
In [71]: df
Out[71]:
```

	Timestamp	decision	comm	health_conc	grade	inter_stu	hot_spot	off_lease	onl_class	stu_ath	sign_oth	major_req	preference	friends
0	4/8/2021 12:38:43	0	0	0.0	3	0	0.0	0.0	3.0	1	1	0	0	0
1	4/12/2021 14:19:25	0	0	0.0	4	0	1.0	0.0	4.0	0	0	0	1	1
2	4/12/2021 14:19:25	0	0	0.0	3	0	0.0	1.0	4.0	1	0	0	1	0
3	4/12/2021 14:19:33	1	0	2.0	1	0	1.0	0.0	2.0	1	0	1	1	1
4	4/12/2021 14:19:40	1	0	2.0	0	0	0.0	0.0	3.0	1	0	0	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
197	4/27/2021 20:16:32	1	0	2.0	2	0	1.0	0.0	3.0	1	0	0	1	2
198	4/27/2021 20:19:15	0	0	0.0	1	0	1.0	0.0	3.0	1	0	0	1	0
199	4/27/2021 20:21:15	1	0	2.0	0	0	1.0	0.0	3.0	1	0	0	1	1
200	4/27/2021 20:23:05	1	0	2.0	1	0	1.0	0.0	2.0	1	0	0	1	2
201	4/27/2021 20:24:27	1	0	2.0	1	0	1.0	0.0	3.0	1	0	0	1	2

202 rows × 14 columns

The data was collected through an online survey on google form and was sent to the student mailing list of Wagner College. After two weeks of mass emails to the university and personally reaching out to students to complete the survey, there were a total of 202 responses. The survey consisted of 13 questions with very simple and easy answers. The google form was linked to a google sheet as the sheet updated every time someone answered. It needed to be linked to a google sheet because it made the response into a structured dataset. This then made it easier to convert into a csv file and ready for data cleaning and data manipulation. In Figure 1, it shows the complete dataset after the responses have been encoded to numeric values and missing values were filled in.

Figure 2

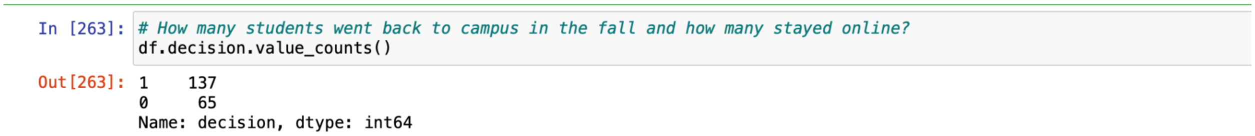


```
In [46]: df.isnull().sum()

Out[46]: Timestamp        0
         decision         0
         comm             0
         health_conc      1
         grade            0
         inter_stu        0
         hot_spot        18
         off_lease        2
         onl_class        5
         stu_ath          0
         sign_oth         0
         major_req        0
         preference       0
         friends          0
         dtype: int64
```

An important part in preparing the data for analysis is handling missing values within the dataset. Missing values can mess up the regression and present you false results or missing results. As shown in Figure 2, python shows the sum of all missing or null values in each column. In the health_conc column there was 1 missing value, 18 missing values in the hot_spot column, 2 missing values in the off_lease column and 5 missing values in the onl_class column. It was taken in consideration that each missing value must be filled in correctly or it can have a misinterpretation in the analysis. Therefore, the missing value in the health_conc column was filled in with the mode or the majority of the column which was that the student attended class in person. The missing values in the hot_spot was determined that 17 of the 18 missing values were international students which forced them to skip the question since they do not reside in any state in the United States. In result, those missing values were filled with 0 or they did not reside in a restricted state. The off_lease variable was filled in with the mode as the students skipped the question assuming that the student did not have an off campus lease. Finally, the onl_class missing values were filled in with the median which was the value of 3 online classes.

With the given data that was collected and cleaned, it is important to get insight on the data to get a better understanding of the sample. An exploratory data analysis is performed on python and is presented with

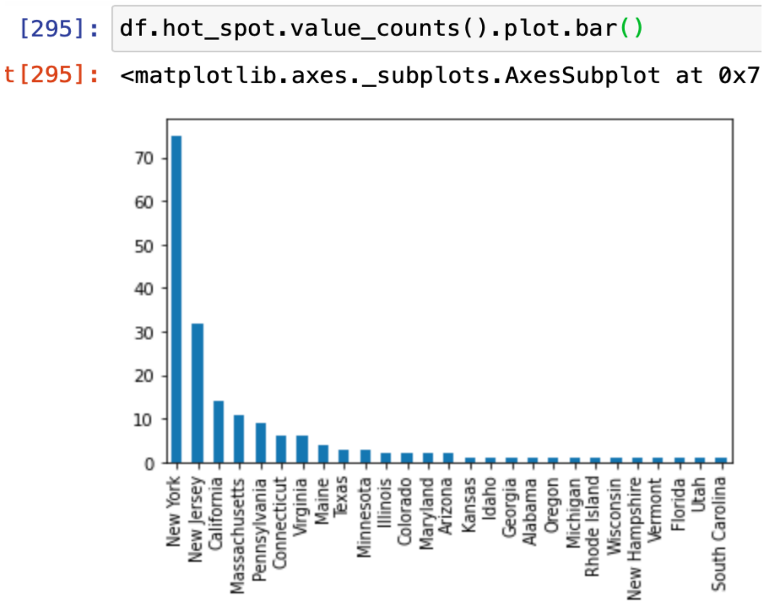
Figure 3



```
In [263]: # How many students went back to campus in the fall and how many stayed online?
          df.decision.value_counts()

Out[263]: 1    137
          0     65
          Name: decision, dtype: int64
```

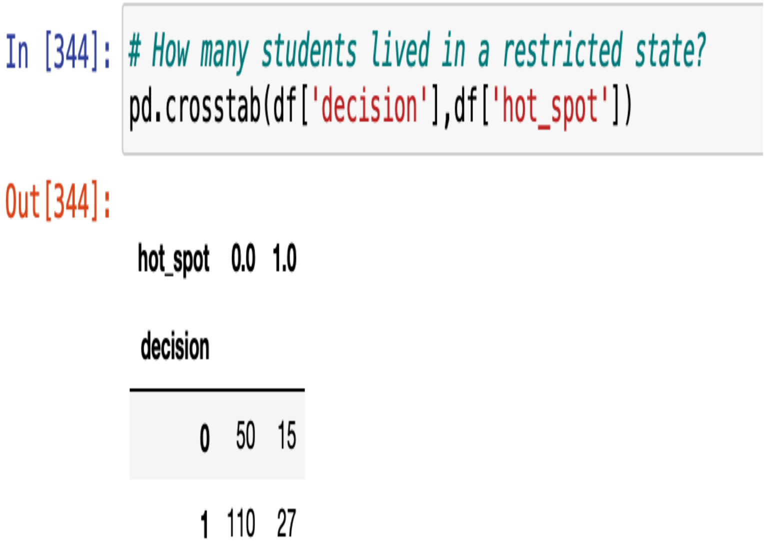
It is best to start with the decisions that students made in the fall semester of 2020. As shown above, there were 137 students that came back to campus while 65 students stayed home and attended courses online.

Figure 4



```
[295]: df.hot_spot.value_counts().plot.bar()
t[295]: <matplotlib.axes._subplots.AxesSubplot at 0x7
```

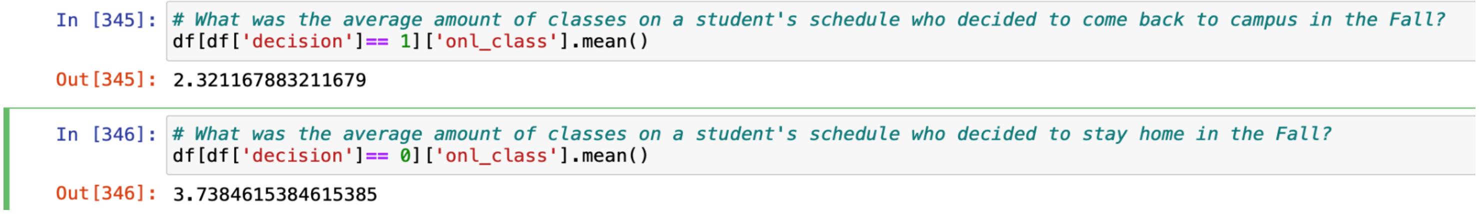
Figure 5



```
In [344]: # How many students lived in a restricted state?
          pd.crosstab(df['decision'],df['hot_spot'])

Out[344]:
          hot_spot  0.0  1.0
          decision
                 0   50   15
                 1  110   27
```

The survey asks students to verify the state that they reside in. This was a really important factor in determining their decision to make the trip and quarantine in New York before move-in day. As seen in *Figure 4*, students from 27 different states filled out the survey as a student at Wagner College. The majority of the students state that they reside in New York and New Jersey since Wagner is in New York City with New Jersey being a neighboring state. With Wagner being a very diverse university, students being out of state makes it difficult to come to New York with the travel restrictions it had placed as of August 2020. *Figure 5* shows a crosstab of the amount of students who live in a restricted state in terms of New York's state law, who don't live in a restricted state and both of their decisions. It shows a total of 42 students who live in a restricted state and a total of 160 students who do not live in a restricted state. However, there were 18 students that were added to the total of students who did not live in a restricted state because they resided outside the country. Out of the 42 students who live in a restricted state, 15 decided to not go through the quarantine process in New York while 27 students decided to come to New York and attended in person courses. Of the students who do not live in a restricted state, 110 students came back to campus while 50 decided to stay home.

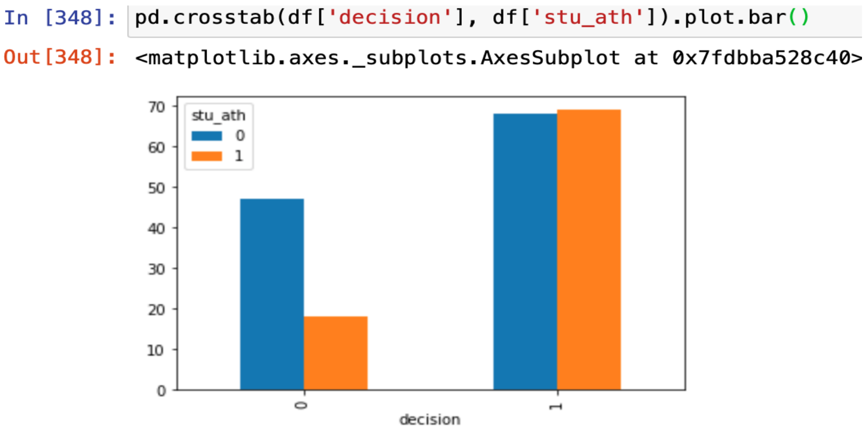
Figure 6



```
In [345]: # What was the average amount of classes on a student's schedule who decided to come back to campus in the Fall?
          df[df['decision']== 1]['onl_class'].mean()

Out[345]: 2.321167883211679

In [346]: # What was the average amount of classes on a student's schedule who decided to stay home in the Fall?
          df[df['decision']== 0]['onl_class'].mean()

Out[346]: 3.7384615384615385
```

Professors also had the option to teach in person or construct their course(s) online through Zoom. This factored in a student's decision to come back to campus or stay online based on what their schedule looked like. As seen in *Figure 6*, for the students who decided to come back to campus, there was an average of 2.3 classes that was made online by their professors. However, for the people who decided to stay home, there was an average of 3.7 online classes on their schedule. As expected, the more online classes a student had on their schedule gave them more of a reason to not pay for room and board. The more classes a student had in person, the more the student would want to take advantage of the resources available provided by the school.

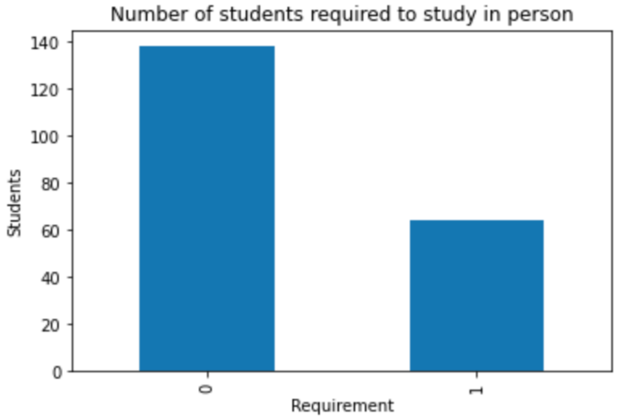
Figure 7



```
In [348]: pd.crosstab(df['decision'], df['stu_ath']).plot.bar()
Out[348]: <matplotlib.axes._subplots.AxesSubplot at 0x7fdbba528c40>
```

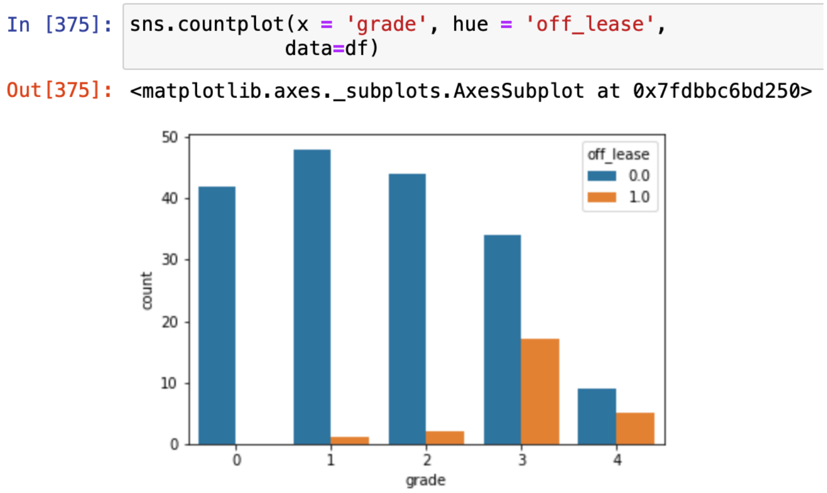
The student-athlete population at Wagner College makes up almost a quarter of the whole student body. According to the Equity in Athletics Disclosure Act (Fahey, 2020), there are a total of 562 participating student-athletes in 25 different sports as of August 31, 2020. In this sample, *Figure 7* highlights student-athletes in orange, non student-athletes in blue and their decision. The bar chart shows that 69 of the 87 student-athletes decided to come back to their respective teams on campus and participate in practice. In other words, 79.3% of student athletes decided to return to their sport. There are a total of 115 non student-athletes while only 59% of those students decided to come back. There are multiple reasons as to why a student-athlete wanted to come back to campus. One being that the athlete misses their sport since sports had been shut down 6 months prior to the fall semester. Another reason is that because since the fall semester was an off season for all teams at Wagner, athletes felt the need to show up and earn a spot on the team or a spot as a starter. The data clearly shows that student-athletes wanted to get back to their sport and their fellow teammates.

Figure 8



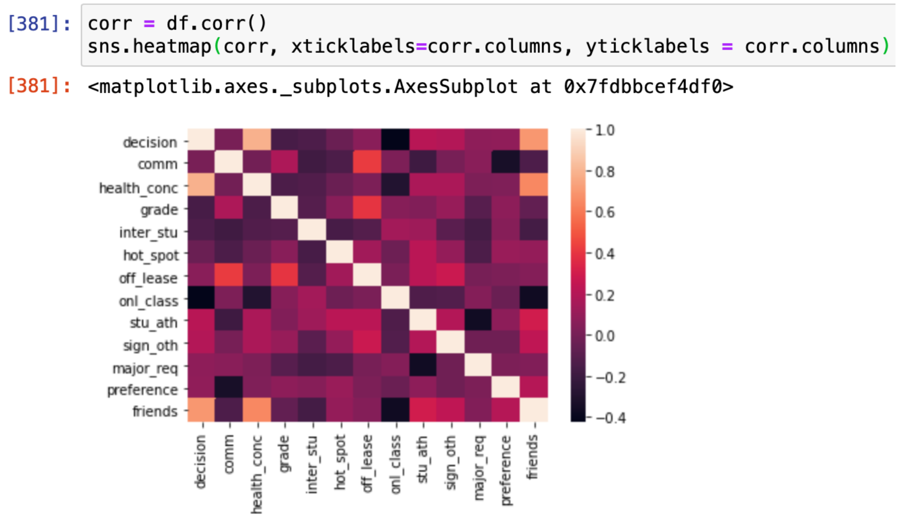
Data on the students' major was also vital insight to their decision in the fall. Certain majors at Wagner were required to study on campus. The nursing and physician assistant programs are quite large at Wagner College. The nursing students were required to be in person because of clinicals and simlab. Clinicals and simlabs that the program runs are a shadow of another nurse at a hospital and cannot be conducted online. In the physician assistant program, 4th and 5th year students within the program were required to show up to campus. To earn a certification, the physician assistant students were not allowed to do online clinicals in place of in person work. In order to graduate, the students needed to physically go to the hospitals in New York City. Out of the 202 students in the sample, 64 students were required to come back to campus for their field of study.

Figure 9


```
In [375]: sns.countplot(x = 'grade', hue = 'off_lease',
                        data=df)

Out[375]: <matplotlib.axes._subplots.AxesSubplot at 0x7fdbbc6bd250>
```

An interesting finding shows the students that have an off campus lease. Students usually sign a 12 month lease before the fall semester and cannot break the lease. Therefore, once a student signs a lease for an apartment or house, they are obligated to pay each month's rent for 12 months. At Wagner College, all students sign a contract to live on campus until senior year but if a student moves off campus before senior year, they are charged with $1,000. The data in *Figure 9* shows that the majority of students that have an off campus lease are seniors (3) and graduate students (4).

Figure 10



```
[381]: corr = df.corr()
       sns.heatmap(corr, xticklabels=corr.columns, yticklabels = corr.columns)

[381]: <matplotlib.axes._subplots.AxesSubplot at 0x7fdbbcef4df0>
```

Before the economic modeling, it is important to look at the correlation between the variables. In the package *Seaborn* in python, it is best to visualize the covariance with a correlation matrix. A correlation can best describe the strength of a linear relationship. The correlation matrix in *Figure 10* shows the relationship between each variable. The lighter the color is, the more correlated it is with its corresponding variable. As seen in the heatmap, the decision of the student is highly correlated with the variable 'friends' where their friends are a primary factor in their decision to come back to campus.

IV. Economic Specification

A regression analysis is to quantify estimates of an economic relationship. It is a statistical technique that tries to explain the relationship between one variable and the dependent variable. It's important to note that in order to find movements in a single variable, all other explanatory variables must be held constant while changing the variable. To predict the direction of the change in a single variable, you must have a general sense of knowledge on the characteristics of the product in the question. For example, if you wanted to predict the direction of change in a variable of the amount of hours spent studying on the effect of a student's grade, you must use intuition and have knowledge that the more hours a student spends studying, the increase in the students performance and grade. The regression model may be written as:

$Yi = B0 + B1Xi + ei \quad (i = 1, 2, \ldots, N).$

In this equation, you have 'Yi' which is the *i*th observation of the dependent variable. B0 is the constant and 'B1' is a regression coefficient. 'Xi' represents the *i*th observation of the independent variable. 'N' is equal to the number of observations in the sample or data. Lastly, 'ei' is the *i*th observation of the stochastic error term. A stochastic error term is a variable that is added to a regression equation to show that there are more variations in the dependent variable that cannot be explained by the included independent variables (Studenmund & Studenmund, 2015).

The model in this project has a predictor variable (attending on campus **or** staying home online) of only two possible outcomes. Therefore, the model calls for a logistic regression as the variable will encode 1 being on campus while 0 being the student stays home and takes courses online. With all variables considered, the model has been cut down to 6 explanatory variables. All variables were coded to a simpler name to make it easier when analyzing the data. Note: a dummy variable is a variable that only has two possible outcomes where it is usually encoded as 1 or 0.

The 10 explanatory variables are:

comm: dummy variable where 1 means they are a commuter student and 0 being they are not a commuter student.

grade: range from 0-4 where 0 represents freshman, 1 represents sophomore, 2 represents junior, 3 represents senior and 4 represents a graduate student.

inter_stu: dummy variable where 1 represents that the student resides outside the United States and 0 represents that the student lives in the United States.

hot_spot: dummy variable where 1 represents that the student lives in a hotspot state at the time of August 2020 and 0 represents that the student did not live in a hotspot state.

off_lease: dummy variable where 1 represents a student having an off campus lease and 0 being the student does NOT have an off campus lease.

onl_class: scale from 0-7 as it represents the number of classes the professor made their class online.

stu_athl: dummy variable where 1 represents a student being a student-athlete and 0 represents a student that does not play a sport.

sig_oth: dummy variable where 1 represents the student having a significant other attending Wagner College and 0 represents the student not having a significant other attending Wagner College.

preference: dummy variable where 1 represents the student favoring to learn in a classroom setting and 0 represents the student favoring to learn online.

friends: range from 0-2 where 0 represents that the student was online for the fall semester, 1 represents that friends weren't a primary factor for their reason to come back to campus and 2 represents that friends were a primary factor for their reason to come back to campus.

The dependent variable is:

decision: dummy variable where 1 represents a student attending class on campus and 0 represents a student staying home and taking classes online.

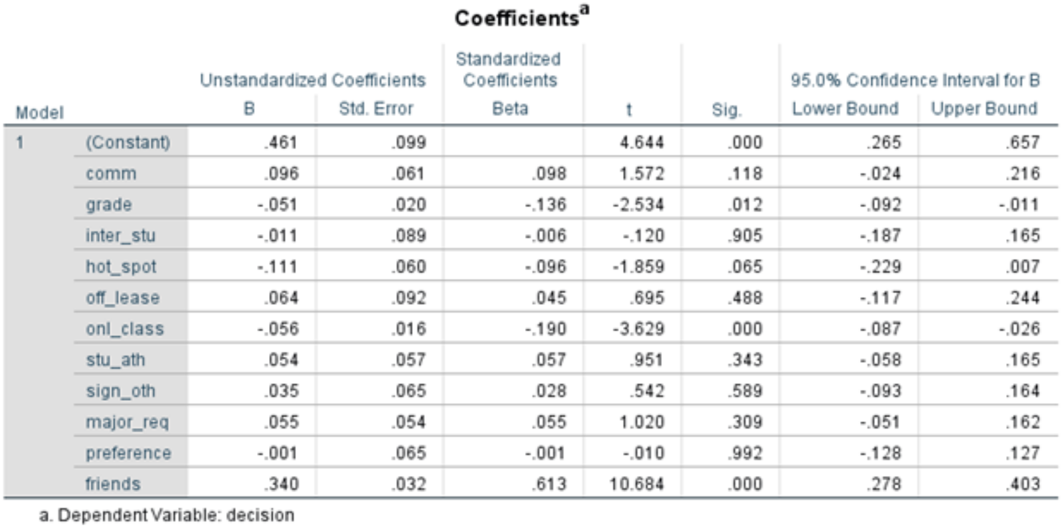
After choosing all possible variables that can affect change in the dependent variable, the estimated regression equation must be formed. The estimated regression equation for this project is:

decision(i) = B0 + **B comm - B grade - B inter_stu - B hot_spot + B off_lease - B onl_class + B stu_ath + B sign_oth + B major_req + B preference + B friends + e**

This equation is estimated with assumptions on their signs. In other words, with knowledge on the topic and intuition, each variable is assumed to have either a positive or negative effect on the dependent variable. The commuter variable is estimated to increase the probability of the student's decision to come back to campus. This is because most commuters wanted to get out of the house since there is no cost in living in their own home. The grade variable has an expected negative effect on the dependent variable because older students aren't too worried about the college experience but more focused on saving money. The inter_stu variable has an expected negative effect on the student's likelihood to come back to campus because at the time of move in day for Wagner, it was extremely difficult to fly into the United States from other countries given the restrictions. The hot_spot variable is negative because hotspot states were restricted from coming to New York or they would make you quarantine for 14 days. The off_lease variable is positive because students are usually paying for a 12-month lease no matter if they study on campus or not. Therefore, students do not want to waste twelve months worth of rent while not using their apartment or house. The online class variable decreases the likelihood of a student's decision to come back to campus because the more online classes that the professor makes on their schedule, gives them more of a reason to stay home and not waste money on living situations. The student athlete (stu_ath) variable has a positive effect on a student's decision to come back to campus because most athletes love and miss their sport and do not want to waste practice time when trying to earn a spot on the team or starting position. The significant other (sign_oth) variable is expected to increase the likelihood of the student coming back to campus because they are more likely to want to spend time with their boyfriend/girlfriend with the convenience of being on the same campus or general area. The required major variable has a positive effect on a student's decision because fields such as nursing and the physician assistant program require students to come back to campus and learn in person. The preference variable is expected to increase the probability of a student to study in person because a marginal effect of 1 represents that the student prefers to learn in a classroom setting. Lastly, the friends variable has a positive relationship with the dependent variable because the marginal effect represents that the student's primary factor in coming back to school was to be around their friends again.

V. Results

Table 1



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.461	.099		4.644	.000	.265	.657
	comm	.096	.061	.098	1.572	.118	-.024	.216
	grade	-.051	.020	-.136	-2.534	.012	-.092	-.011
	inter_stu	-.011	.089	-.006	-.120	.905	-.187	.165
	hot_spot	-.111	.060	-.096	-1.859	.065	-.229	.007
	off_lease	.064	.092	.045	.695	.488	-.117	.244
	onl_class	-.056	.016	-.190	-3.629	.000	-.087	-.026
	stu_ath	.054	.057	.057	.951	.343	-.058	.165
	sign_oth	.035	.065	.028	.542	.589	-.093	.164
	major_req	.055	.054	.055	1.020	.309	-.051	.162
	preference	-.001	.065	-.001	-.010	.992	-.128	.127
	friends	.340	.032	.613	10.684	.000	.278	.403

a. Dependent Variable: decision

The regression was run through SPSS and showed each variable's coefficient and its effect on the dependent variable. As shown above, the equation now reads as:

decision = 0.461 + 0.96comm - 0.051grade - 0.011inter_stu- 0.111hot_spot + 0.064off_lease - 0.056onl_class + 0.054stu_ath + 0.035sign_oth + 0.055major_req - 0.001preference + 0.340friends

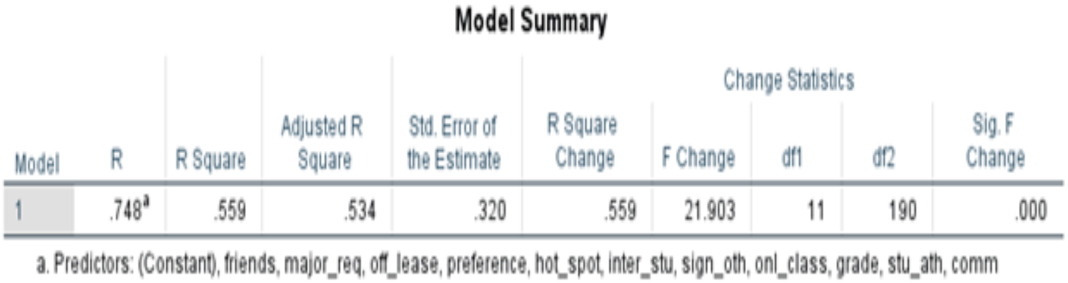
It is important to note that the expected signs in the estimated equation for each variable read true after running the regression except for the preference variable. The constant for this equation reads as .461. To interpret the equation, it states that there is an increase of probability that the student will come back to campus if the student is a commuter by .096. It is more probable that the student will stay online by .051 for every increase in grade. If the student is an international student, it is .011 more likely that the student will stay online. There is a decrease in probability by .111 that the student will come back to campus if the student resides in a restricted state based on where the student resides. If the student has an off campus lease in either an apartment or a house, it increases the probability of the student coming back to campus by .064. Every increment of the amount of online classes that the student has on their schedule, decreases the probability of the student coming back to campus by .056. If the student plays a sport at Wagner College, it is .054 more likely that the student will come back to campus. If the student has a significant other that attends Wagner College, it is .035 more likely that that the student will come back to campus. It is .055 more likely that the student will come back to campus if the student's major requires them to come back to campus and study. As stated before, the preference variable was expected to have a positive relationship with the likelihood of the student returning to campus. However, after running the regression, the coefficient of the preference variable reads that it has a negative marginal effect on the dependent variable. Although there is a very minimal negative impact on the dependent variable with the preference variable's coefficient with .001, it is important to check why this occurred. After reviewing the possible reasons for the unexpected sign, it is expected that the sample size is too small. High variances could result from a small sample size in which the sample size of the data presented here, only accounts for 202 students of the university's population. To fix this unexpected sign, the sample size would need to be increased. Lastly, the friends variable has a marginal increase of .340 on the dependent variable since friends are the student's main reason for studying on campus.

In table 1, it shows the t-statistic for each variable. The t-statistic given can be interpreted as a test to see whether the corresponding coefficient is not equal to zero. Therefore, the null hypothesis would be Ho = 0 and would try to reject this. In other words, each variable would be set equal to 0 and to prove that it is not equal to 0, it is necessary to reject this hypothesis. Rule of thumb says that at a 1% significance level, I can reject the null hypothesis (Ho=0) if the absolute value of the t-statistic is greater than 2.5. With the t-statistics given in table 1, I can confidently conclude that the grade, onl_class, and friends variables are statistically significant from 0 at a 1% significance level. For the other variables, I cannot confidently reject the null hypothesis at a 1% significance level because the absolute value for their t-statistic is less than 2.5.

With the t-statistic explained, economists tend to look directly at the variable's p-value. The variable's p-value in table 1 is labeled under "Sig.". The p-value essentially explains the same thing as a t-statistic, however, it shows exactly what significance level it is at. In table 1, it shows that the p-value for the commuter variable(comm) is at 0.118. This means that the null hypothesis of its coefficient can only be rejected at approximately 11%. The grade and coefficient has a p-value of .012 where the null hypothesis can be rejected at 1%. The null hypothesis of the international student coefficient can be rejected at 90%. The null hypothesis of the hot_spot variable can be rejected at 6% and the off_lease variable can be rejected at 48%. The onl_class and friends variables are equal or very close to equal to 0 which means that their coefficients can be rejected at a high degree of confidence. The preference's coefficient can be rejected at 99% while the stu_ath and sign_other coefficients have a significance level of 60%. Lastly, the major _req coefficient can only be rejected at a 30% significance level.

Table 1 also displays the 95% confidence interval columns on the far right. The 95% confidence interval columns say that with 95% confidence, the true beta falls between the two values. The lower you get in degrees of confidence, the smaller the gap is between the two values. For example, with 95% confidence, the true beta of the commuter variable (comm) falls between -0.101 and 0.185. Intuitively, if the confidence interval lowers it is less confident to tell that the true beta will fall in between the two values. Therefore, at a 95% confidence level the true beta for the grade coefficient falls between -0.092 and -0.011; -0.187 and .165 for the inter_stu coefficient; -0.229 and 0.007 for the hot_spot variable; -0.117 and 0.224 for the off_lease coefficient; -0.087 and -0.026 for the onl_class coefficient; -0.058 and 0.165 for the stu_ath coefficient; -0.093 and 0.164 for the sign_oth coefficient; -0.051 and 0.162 for the major_req coefficient; -0.128 and 0.127 for the preference coefficient; 0.278 and 0.403 for the friends variable.

Table 2



Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.748[a]	.559	.534	.320	.559	21.903	11	190	.000

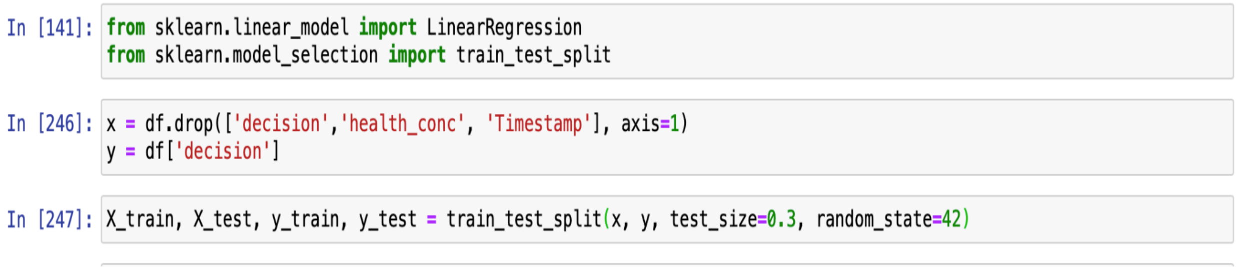
a. Predictors: (Constant), friends, major_req, off_lease, preference, hot_spot, inter_stu, sign_oth, onl_class, grade, stu_ath, comm

To describe the overall fit of the estimated equation, R-squared is commonly used. In other words, this can be measured as the “goodness of fit” for the model. According to Studenmund (Studenmund & Studenmund, 2015), the higher the R-squared is after running the regression, the closer the equation fits the sample data. On the other hand, the lower the R-squared means that the independent variable(s) are not related to the dependent variable. Therefore, if R-squared is equal to 0, then the independent variable(s) and the dependent variable are not related and there is no explanation for the variation in the Y/dependent variable. After running the regression, it is shown in Table 2 that the R-squared has a value of .559. This has a very standard R-squared value for cross-sectional data. Cross-sectional data are collected to assess and observe a random sample at a certain point in time, disregarding differences in time. An R-squared value would be higher in a time series regression model because it is the result of significant trends over time within the predictor variable and explanatory variables. The adjusted R-squared value in Table 2 reads as .534. The adjusted R-squared value can be used to compare the fits of equations with the same dependent variable and different number of independent variables (Studenmund & Studenmund, 2015).

VI. Machine learning, Python

Data science is a rapidly growing industry as the world is changing due to the immense amount of data. According to Provost and Fawcett, the ultimate goal of data science is to improve decision making based on the statistics and findings through the data collected (Provost & Fawcett, 2013). An aspect that allows data scientists to make decisions for the future is machine learning. Machine learning in data science allows for the computer to identify trends within a dataset and make predictions based on the patterns it finds. The data was transferred to python for the implementation of a simple linear regression algorithm. This type of machine learning is supervised learning because the data is already structured with columns as categories. Unsupervised learning is when the machine must learn on its own where the data does not have prior categories to classify.

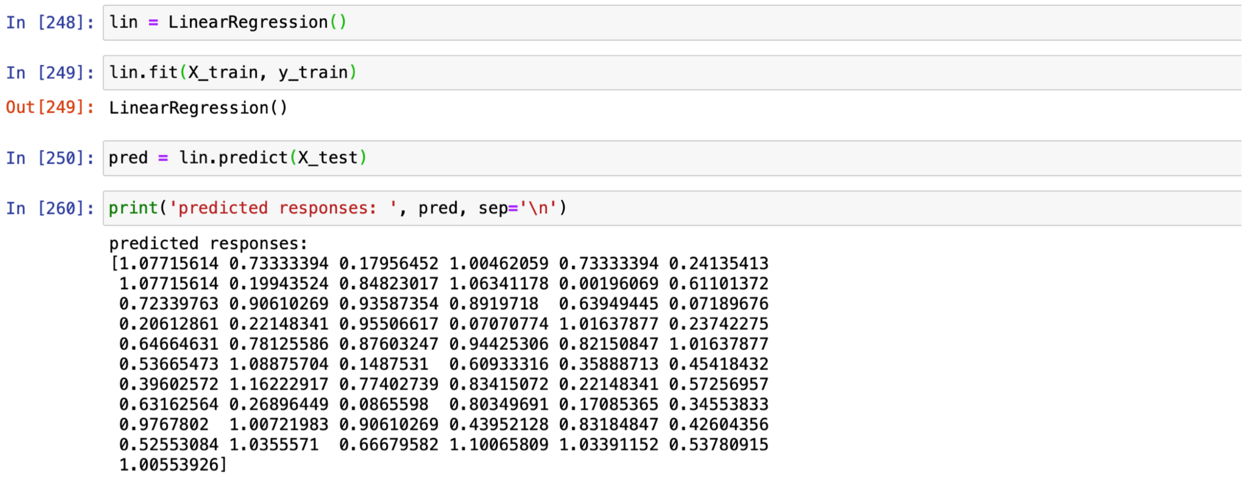
Figure 11



```
In [141]: from sklearn.linear_model import LinearRegression
          from sklearn.model_selection import train_test_split

In [246]: x = df.drop(['decision','health_conc', 'Timestamp'], axis=1)
          y = df['decision']

In [247]: X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
```

In the code above completed in python's jupyter notebook, is the beginning of implementing a simple linear regression algorithm. The package used in python is called scikit-learn and is one of the most useful libraries in python for machine learning. It contains efficient tools for statistical modeling such as classification, clustering and in this project, regression. The second line in the code that states "from sklearn.model_selection import train_test_split" is a package that allows you to split the dataset into a training and testing set. The reason to split the dataset into a training and testing set is because the computer needs to learn patterns and make predictions. The computer will take the training dataset and use machine learning to improve the algorithm. If the computer learns the whole dataset, the model will be very familiar with the data and make near perfect predictions. Test sets are necessary after training the model because the new data that the computer has not seen yet will simulate real-world predictions. In the third of the code in *Figure 11*, the independent variables (x) are defined as the dependent variable and unnecessary variables not used in the model are dropped. In the fourth line of the code, the dependent variable (y) is defined as the column 'decision' is labeled where it contains all of the student's decision from the dataset. In the last line of the code, the dataset is then randomly split into the training and testing set where 70% of the data is assigned to the training set and 30% of the data is assigned to the testing set.

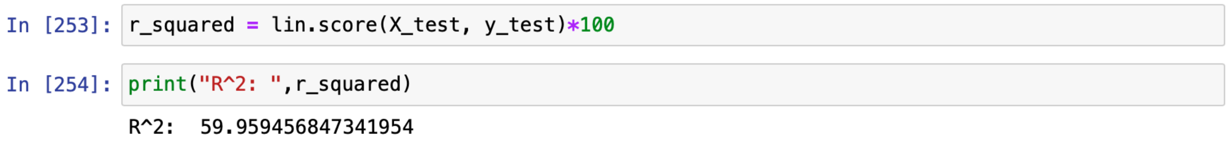
Figure 12



```
In [248]: lin = LinearRegression()

In [249]: lin.fit(X_train, y_train)
Out[249]: LinearRegression()

In [250]: pred = lin.predict(X_test)

In [260]: print('predicted responses: ', pred, sep='\n')
          predicted responses:
          [1.07715614 0.73333394 0.17956452 1.00462059 0.73333394 0.24135413
           1.07715614 0.19943524 0.84823017 1.06341178 0.00196069 0.61101372
           0.72339763 0.90610269 0.93587354 0.8919718  0.63949445 0.07189676
           0.20612861 0.22148341 0.95506617 0.07070774 1.01637877 0.23742275
           0.64664631 0.78125586 0.87603247 0.94425306 0.82150847 1.01637877
           0.53665473 1.08875704 0.1487531  0.60933316 0.35888713 0.45418432
           0.39602572 1.16222917 0.77402739 0.83415072 0.22148341 0.57256957
           0.63162564 0.26896449 0.0865598  0.80349691 0.17085365 0.34553833
           0.9767802  1.00721983 0.90610269 0.43952128 0.83184847 0.42604356
           0.52553084 1.0355571  0.66679582 1.10065809 1.03391152 0.53780915
           1.00553926]
```

The code above in *Figure 12* shows the linear regression algorithm performed on the training dataset. After the computer identifies its patterns and learns from the training set, the model was used to predict outcomes based on the independent variables in the testing set. All of the predicted outcomes are listed below the code where it says “print(pred)”. The predictions are probabilities that a student will come to school based on the independent variables.

Figure 13



```
In [253]: r_squared = lin.score(X_test, y_test)*100

In [254]: print("R^2: ",r_squared)
          R^2:  59.959456847341954
```

The last step is to check the efficiency of the model or the value of the R-squared using the syntax .score(). In this model, it received an R-squared score of 59.96% which indicates that the model explains 59.96% of the variance of the student’s decision is explained by the variance of the independent variables.

VII. Conclusion

The paper concludes with understanding the multiple components that go through a student's decision during the pandemic. The model created scored an accuracy of 55-59.9% which is acceptable for a social science experiment. The exploratory variables were narrowed down with personal experience and immense research. Although not every student's situation in the 2020 Fall semester was the same, the variables were carefully considered and were decided based on strong impact. Issues throughout the experiment come from the data collection segment of the process. The data would be more accurate if there was a bigger sample size. The small sample size caused an unexpected sign within the estimated equation. Overall, in order to make this model statistically significant, a larger sample size must be obtained.

The pandemic changed many lives in the past year, specifically universities across the country. This experiment's aim was to add value to Wagner College with a model and analytical insight. In the future, I hope to expand research at Wagner and branch out to more universities in the United States.

References

Bradley EH, An M, Fox E. Reopening Colleges During the Coronavirus Disease 2019 (COVID-19) Pandemic—One Size Does Not Fit All. *JAMA Netw Open.* 2020;3(7):e2017838. doi:10.1001/jamanetworkopen.2020.17838

Fahey, Brendan (2020). Equity in Athletics Disclosure Act. *Wagner College*, 1-10. https://wagnerathletics.com/documents/2021/1/7/EADA_2020.pdf

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Shah K, Kamrai D, Mekala H, et al. (March 25, 2020) Focus on Mental Health During the Coronavirus (COVID-19) Pandemic: Applying Learnings from the Past Outbreaks. Cureus 12(3): e7405. DOI 10.7759/cureus.7405

Singhal, T. A Review of Coronavirus Disease-2019 (COVID-19). *Indian J Pediatr* 87, 281-286 (2020). https://doi.org/10.1007/s12098-020-03263-6

Studenmund, A.H., & Studenmund, A.H. (2015). *Using econometrics: A practical guide* (Sixth ed.). Harlow, Essex: Pearson Education Limited.

Travel restrictions issued by states in response to the coronavirus (COVID-19) PANDEMIC, 2020-2021. (n.d.). Retrieved April 27, 2021, from https://ballotpedia.org/Travel_restrictions_issued_by_states_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020-2021#Active_restrictions_and_recent_news

Tümen Akyıldız, S. (2020). College students' views on the pandemic distance education: A focus group discussion. International Journal of Technology in Education and Science (IJTES), 4(4), 322-334.