Logo

Description automatically generatedSyracuse University

School of Information Studies

**IST 782: APPLIED DATA SCIENCE PORTFOLIO MILESTONE**

Instructor: Dr. John Stinnett

Student Name: Jeffrey Thomson

SUID: 688596641

Email: jethomso@syr.edu

Table of Contents

Introduction 3

Describe a Broad Overview of the Major Practice areas of Data Science 4

Collect and Organize Data 4

Project Connection 5

Data Collection Summary 5

Project Connection 7

Data Source 7

Reflection and Learning Outcome 7

Identify Patterns in Data via Visualization, Statistical Analysis, & Data Mining 8

Project Connection 8

Background 9

Pattern Recognition 9

Reflection and Learning Outcome 13

Develop Alternative Strategies based on the Data 13

Project Connection 13

Background 13

Alternative Strategy 16

Project Connection 17

Alternative Strategy 17

Reflection and Learning Outcome 18

Develop a Plan of Action to Implement the Business Decisions derived from the Analyses 18

Project Connection 18

Background 18

Business Decision 19

Project Connection 20

Background 21

Business Decision 21

Reflection and Learning Outcome 21

Demonstrate Communication Skills regarding Data and its Analysis 22

Project Connection 22

Project Connection 22

Reflection and Learning Outcome 23

Synthesize the Ethical Dimensions of Data Science Practice 23

Reflection and Learning Outcome 23

Introduction:

I have worked in the financial analytics department for various companies over my twenty years as a professional. These have varied in size from the United States Postal Service to currently for a local hospital in my town. Each company had very similar issues when it came to analytics. Data was very hard to collect, organize and analyze to provide meaningful results. Either datum was spread across multiple systems that were not compatible with each other or, often, data was missing or wrong due to human error. Our only tool was to use excel and manually copy data into spreadsheet. You can already see the issues. For a place like the Postal Service, there was sometimes too much data to bring over. Or with my current company, the data is hard to export and systems that hold similar metrics have incompatible systems, making merging data with excel time consuming. When we do manage to get the data into a spreadsheet, performing predictive analytics in excel is hard so we revert to simple mechanisms like averages. While this is fine, we lack any robust software that could open us to more meaningful insights and drive operational improvements.

To that, I found the Syracuse University School of Information Studies. Not having any prior programming experience, I assumed I would be ignored when I first inquired about the program. Not the case. Turns out, lack of experience is welcomed and understood amongst every professor. My goal is to bring what I have learned in the program and apply it to my current work so that we can, over time, produce more robust and meaningful analytics.

To ultimately succeed and graduate, students must show that they have acquired significant capabilities in the below competencies. I hesitate to use the word “mastered” as that implies there is no more to learn which is never the case.

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patters in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professional, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy)

The below write-up will show specific examples of each learning outcome and show that I have acquired the capabilities to succeed in this field.

1. **Describe a broad overview of the major practice areas in data science:**

The continued theme within the program is that data science merges the use of programming language like R or Python with statistical analytics. To do this you must become proficient with data collection and mining, statistics and probability, algebra, various regression techniques, classification, predictive analytics, machine learning, big data analytics and data visualization.

In my role as a financial analyst, we ask ourselves everyday what our financial future looks like. This insight allows us to predict cash flow which can then be used to buy better equipment and expand operations. So, of the proficiencies listed above, the predictive analytics has always been a keen area of interest for me.

**2. Collect and organize data:**

Data collection is an integral part of data science as it is the first step once the scope of your project has been defined. Collection has many forms, from using human entered spreadsheet files that contain individual rows and columns often used for record keeping, to pulling feeds from Facebook or a stream of data obtained in POS (point-of-sale) transaction or camera sensor or key card transactions.

Each class required some form of data collection. As most professors pointed out, and we learned from experience, collecting, and organizing the data into proper structures could take 50%-60% of the entire assignment or project timeframe.

**Project Connection: IST 652: Scripting for Data Analysis:**

**GitHub Repository:** IST 652: Scripting for Data Analysis/IST\_652\_Jeffrey\_Thomson\_Final\_Project\_final.py

Data collection summary:

The goal for this project was to analyze medal winners in the 2016 Summer Olympics games to determine if a pattern emerged of characteristics leading to most medals won.

Two data sets were obtained. First, a “athletes” dataset from Kaggle that contained attributes of each athlete organized into columns with each athlete in rows. This set consisted of twelve columns and 11,539 rows. The attributes provided were ID, name, nationality, sex, date of birth, height (in meters), weight (in kilograms), sport, gold, silver, bronze, and age. The second file, countries.csv, is a list of each country, their country abbreviation, its population, and GDP per Capita as of 2014.

Each data set needed to be organized into a data frame, cleaned, and prepped. This meant reviewing the attribute types for function compatibility and changing as necessary.

Athletes File:

Table

Description automatically generated Text

Description automatically generated

The data sets were searched for null/blanks and either deleted or updated using averages. The “isnull” function was useful to check the entire dataset for NAs and showed this data contained quite a few. Each column was checked for NAs and then filled with that column’s average metric. Again, the “isnull” function was used but on each column. From the output show above, the “dob” NAs were simply deleted since it only contained one NA. For the “height” and “weight” columns deleting these would results in shrinking the dataset by almost 10%. Instead, we used the “fillna” function and replaced the NAs with the respective mean height and weight.

Measurement attributed were converted from metric to imperial system.

Text

Description automatically generated

The final step was to merge the athletes and countries data sets. I needed a unique identifier and the countries file used ‘code’; which is the countries name abbreviation. The ‘nationality’ column in the athletes file is the country code abbreviation, so I change its name to code to link the two.

I utilized the left merge function on the code column to merge athletes with countries. The left function will merge the two while keeping any NA values within the dataset. I checked the new columns from the countries file for NAs in the population and GDP per Capita. There were 390 NAs in the population and 816 in the GDP which was more than expected. Instead of removing the NAs I decided to use the mean function again. Upon review of the now 18 column data frames, the ‘id’, ‘height’, ‘weight’, and ‘country’ columns were redundant so those were deleted. In the end, we had a clean dataset with fifteen columns that were used to answer the project questions.

A screenshot of a computer

Description automatically generated with medium confidence

**Project Connection: IST 652: Scripting for Data Analysis:**

**GitHub Repository:** IST 652: Scripting for Data Analysis/IST\_652\_Jeffrey\_Thomson\_Final\_Project\_final.py

On a smaller scale, but a good example of data collection, was to collect data from a website. We wanted to determine how many episodes of the tv show aired and compare to the IMBD website. Each episode will be displayed along with the airdate and airtime. The fields to be includes are “name”, “airdate” and “airtime”.

Data Source:

The data was gathered from a json database website called Awesome Ranked. Within there was data for the TV show Mr. Robot. (<https://awesomerank.github.io/lists/jdorfman/awesome-json-datasets.html#movies>). The initial processing was straightforward and all I needed to do was open the url, decode using utf-8 and parse the data using loads.

A function was written opens the url where the json data is stored and used the loads function to bring it into a dictionary type. It then prints the keys used in that dataset. Using those keys as well as visual printouts, it digs deeper to find the episode data and detail. The episode detail is then saved into the MongoDB as a collection called TV shows. From here we wrote a function that loops through the data to print out each episode’s name, the airdate and airtime. The output table shows the episode names, the date and airtime.

Reflection and learning outcome:

Data collection and organization is a first, and very important, step in the analytics process. The overarching question you need to ask is, can I trust my data source and its integrity? Then, how do I turn it into a data set that is compatible with either R-studio or Python functions. I learned quickly that preparation is vital. To that, I mean you need to understand your end goal, plan out your complete analysis and understand the functions you want to use for your analytics so that you know what your variable data types need to be. You must be diligent in your initial review and study the data using summary and str() functions to view your dataset types. Changing them upfront saved valuable time, instead of spending hours only to find you have a dataset that doesn’t work with your functions.

This translates to my current work environment as well. We must pull data from multiple databases that were not designed to operate together. To that I mean, columns signifying the same metric are named differently or the metric is scaled differently. Missing values are scattered everywhere. We must be very diligent to understand these differences, make the necessary corrections and clean it up before performing the analysis.

**3. Identify patters in data via visualization, statistical analysis, and data mining.**

**Project Connection: MAR – 653 – Marketing Analytics:**

**GitHub Repository:** MAR 653: Marketing Analytics/MAR\_653\_Final Project\_python\_code.ipynb

For our marketing project, we had a business problem to fix. We represented an ecommerce website and needed to identify which customers to market to by predicting their propensity to make a purchase. This ecommerce company is currently marketing to all website visitors equally, though only a small percentage of visitors are likely to make a purchase

**Project Goal:**

* What variables/activity make a customer more likely to place an order?​ (Pattern)
* What customers should the ecommerce company market to?​ (Business Decision)
* At what stage of the customer journey should the ecommerce company implement a marketing campaign? (Business Decision)

Background:

We collected a dataset from Kaggle that contained one day’s worth of website traffic transactions from 455 thousand customers with 25 variables measuring behavior on the website. For example, did the customer use a mobile device, did the visitor add to their shopping cart or was it a returning customer.

An ordered column showed whether each transaction ended up with an order being places (1) or not (0). All the data was recorded in a binary so only a “1” or “0” was listed based on where the customer performed the action represented in the variable columns. In this case the “1” represents, yes, the customer placed and order and “0” if they did not. The initial data was slip and 70% was used for the training and 30% for testing.

Pattern Recognition using Statistical Inference:

Using the data, we want to predict whether a customer will place an order based on past behavior. Given the binary nature, logistic regression was used. In logistics regression, the model will aim to predict the probability of an ordered being placed so here “ordered” is the dependent variable.

The initial analytics step was to run a correlation matrix on all the variables to look for the ones with the highest correlation to our dependent variable. Applying a heat map type view, gave an extra level that color coded the results so those with the highest or lowers values would pop out. Upon initial review we identified areas of multicollinearity where the independent variables had high correlations with each other. This resulted in removing two variables, “sign\_in” and “device\_mobile”.

Chart, histogram

Description automatically generated

Using the results of the correlation matrix, we developed three logistic regressions to determine p-values, coefficient rates and R-squared results. A hit rate was calculated using a confusion matrix to show the observed to predicted matches and mismatches.

**Model 1:**

The initial model used a large number of variables in the hope we could narrow the results based on the regression outputs. The R-squared was high and there was a mix of positive to negative coefficients. The variable “check delivery detail” had the highest coefficient showing this attribute added the most positive value to the probability score. The p-values were the most telling because four had values above 0.05 which was above our threshold used those were thrown out.

 

Variables “sort by” , “image picker”, “Promo banner click” and “saw sizechart” had p-values below 0.05, so we concluded that these factors are not associated with changes to “ordered” and should be removed.

You see in the confusion matrix the sheer number of transactions that resulted in no order being placed. This made us realize we had an imbalance classification issue. This occurs where a dataset is highly skewed in the class distribution. We went back and saw that of the 455,401 transactions, 436,308 did not place an order. This bias in the training dataset can influence the accuracy calculation, leading some to ignore the minority class entirely. We found that using the “Recall” calculation for accuracy score which uses the total number of true positive predictions divided by the sum of the true positives and the false negatives, needed to be used instead of the hit rate. Model 1 gave a recall score of 87.35%.

**Model 2:**

Using the statistical results from Model 1, we removed the p-values over 0.05 and reran the model. Interestingly, the R-squared dropped to 82.64% from 87.26%. Three variables had p-values over 0.05. Again, “check delivery detail” had the highest coeficient value.

**Table

Description automatically generated** **Graphical user interface, application

Description automatically generated**

The confusion matrix resulted in a recall rate of 88.60% slightly over the 87.35% from Model 1. It looks that removing those variables helped increase the accuracy of the model.

**Model 3:**

Here we narrowed the variables down to the five that had the absolute highest correlation to ordered. Each variable gave positive coefficient direction, p-values below 0.05 and a decent R-squared of 86.34%.

**Table

Description automatically generatedGraphical user interface, application

Description automatically generated**

Unexpectedly, the recall rate fell to 85.60% from 88.60% in Model 2. We decided that the ecommerce company should focus their marketing on visitors that sign into their account, visit the delivery FAW section and click on the shopping basket icon, as these predict the higher probability that the customer will place an order.

Reflection and learning outcome:

This class and project were fun in that this was not an IST course, so my fellow students were not learning about data science. I was able to bring this skill set and perform the analytics in python which significantly improved the speed over using excel. We learned quickly that the excel program being used could not handle more than eight independent variables, so being able to pivot to python was crucial. We also learned that surprises come up that you were not expecting. For example, we believed that our last model would give the highest recall rate because we followed the statistical results and based variable inclusion decision based on those, but our predicted rate was an only second best. This is an important lesson because it requires us to pivot our beliefs and go where the data takes us.

**4. Develop alternative strategies based on the data.**

**Project Connection: MBC 651: Business Analytics**

**GitHub Repository:** MBC 651: Business Analytics/MBC\_651\_SU\_Recreuitment\_Campaign\_Google\_Analytics.xlsx

Background:

This project revolved around choosing a recruiting advertising campaign strategy for the Whitman School of Management. The marketing department from the Whitman School of Management launched a series of campaigns designed to increase enrollment into the MBA program. Google Analytics supplied data to help measure the effectiveness of those campaigns. By effectively analyzing past campaigns, the marketing department can determine the best way to optimize future campaigns and how to best spend the limited budget of $100,000.

A summary of the past campaigns and metrics used are below in Figure 1.

Table

Description automatically generated

Based on the data, we see that the number of clicks has been decreasing over time while the cost per click increased significantly. We assume this is due to the popularity of using Google allowing them to increase their prices. Assuming a linear relationship, we forecast the cost per click and determined that by 2021 it would be around $69.12.

**Figure 2 Illustrates the Forecast of Students Cost Per Click (CPC)**



Next, we wanted to understand the audience better. What key words were effective? Where were they searching from? When were they searching?

Table

Description automatically generated

Most searches for each campaign originated from New York, specifically areas close to the university including Syracuse and the greater Syracuse area including East Syracuse, Rochester, and Liverpool. New York city also accounted for many clicks likely due to its size and proximity to Syracuse. Significantly fewer returns came from larger California cities including Los Angeles, San Francisco, and San Diego as well as other large cities including Boston, Washington, Atlanta, and Dallas.

Table

Description automatically generated

The above chart illustrates the top keywords used in searches across the different campaigns. These top keywords included “MBA,” “online MBA,” “AACSB MBA” and “MBA without GMAT.” Almost all keywords contain a variation of “MBA,” “online,” “AACSB,” and “without GMAT.” As a result, these represent the key reasons why Syracuse MBA are obtaining web traffic.

The below charts show the activity by days of the week and time of the day for the historic campaigns.

Table

Description automatically generated

Graphical user interface

Description automatically generated

The best days of the week to advertise are relatively consistent, with a slight uptick on weekends. The time to advertise is once many workers finish their day jobs, around 4pm and stretching until 11pm.

Alternative Strategy:

Based on the data, we decided that the campaigns should focus advertising around the greater New York area, near campus and south through New York City because that population is interested in the school. We also expect that population will be most familiar with the school, its history, brand and likely have met alumni, all of which would help increase their awareness and likelihood to attend. The campaigns should utilize the top keywords words and the different combination of such words. Last, advertising should occur Saturdays through Monday from 4pm until 9pm to capture the highest likelihood of prospective searchers.

**Project Connection: IST 652: Scripting for Data Analysis:**

**GitHub Repository:** IST 652: Scripting for Data Analysis/IST\_652\_Jeffrey\_Thomson\_Final\_Project\_final.py

Alternative Strategy:

Another approach to this question is that an alternative analysis is sometimes needed when results do not meet in initial expectations, so you want to dive deeper into the data to confirm. In the Olympic Medals project described in the Organize Data section above, the initial hypothesis was that countries with higher GDPs per Capita would have more medals given the higher access to quality resources. However, the initial correlation of Total Medals to GDP per capita was only 0.14 (Figure 1). Using the “groupby” function we sorted the data from highest to lowest GDP per capita. The results (Figure 2) showed that some of the richest per capita countries have smaller populations which could mean their pool over Olympic quality athletes is also small.

Figure 1: Figure 2:

Chart

Description automatically generatedTable

Description automatically generated

Reflection and learning outcome

The marketing campaign project did not use programming in R or Python, but rather relied only on google analytics. It was beneficial to see how other analytic platforms function and represent data. I found the google platform initiative and easy to navigate. It also shows that there are multiple ways to approach a problem.

The ease at which the google platform was to use, shows how important it is to set up a dashboard so that the end customer can us it with ease. A product can have mounds of data available, but if is difficult to understand, it will not be used.

**5. Develop a plan of action to implement the business decisions derived from the analyses**.

**Project Connection - IST 718 – Big Data Analytics:**

**GitHub Repository:** IST 718: Big Data Analytics/IST\_718\_Big\_Data\_Analytics\_REIT\_Project.ipynb

This case study asked the question on can we predict which ZIP codes provide the best investment for a Real Estate Investment Trust (REIT).

Background:

Data was collected from the housing app Zillow showing single family residence by ZIP code. This required reading in a csv file that contained over 30,000 rows and 300 columns of housing data from January 1996 to March 2020. The rows represented unique ZIPs and median house prices by month.

The REIT project had two goals. First, we looked at house price increases in four metro areas within Arkansas from 1996 to 2020. To visualize this, we created a separate data frame for just Arkansas and calculated the mean value increase for each metro area and plotted this on a line graph. Second, we looked at the entire dataset and calculated the compound annual growth rate (CAGR) to show which ZIPs had the highest growth. This would help narrow where to invest.

A separate data frame was created for just the Arkansas ZIPs. Next, a chart was created for each metro area and then combined into one.

Chart

Description automatically generated

Business Decision:

The above graph, while simple, tells a lot. Deciding where to live depends on the goals of our REIT. If we want price appreciation, Hot Springs, Searcy and Little Rock have appreciated around 65% since 1996 while Fayetteville is closer to 95% price appreciation. If we want consistency, Searcy has steadily appreciated and did not experience the large price drops the other three did during the 2008 financial crisis.

Next, we went back to the full data set and included the Compound Annual Growth Rate as a new column. Adding CAGR adjusts for seasonality to show the smooth price appreciation or depreciation percent over the time. These results can them be sorted to show highest to lowest. This analysis showed that the ZIP codes with the highest CAGR are in:

02108 – Boston, MA

02116 – Boston, MA

02118 – Boston, MA

Based on these results, the REIT should invest a higher percentage of its fund in the Boston, MA metropolitan area.

**Project Connection - IST 623 – Information Security:**

**GitHub Repository:** IST 623: Information Security/IST 623\_Information\_Security\_SolarWindsHack\_Presentation.pdf

For this project we researched the recent 2020 SolarWind’s cyber security hacker infiltration to determine the root causes and what businesses can do to prevent future attacks.

Background:

In September 2019, cyber hacker group Nobelium, installed the malicious code, “Sunburst”, into a batch of software distributed by SolarWinds as an update. More than 30,000 public and private organizations -- including local, state and federal agencies -- use the SolarWinds network management system to manage their IT resources. More than 18,000 SolarWinds customers installed the malicious updates. Through this code, hackers accessed these customers’ information technology systems, which they could then use to install even more malware to spy on other companies and organizations.

This was considered a supply chain attack, since SolarWinds is a vendor to these companies. Because of this, customers never fathomed an infiltration would originate from them. The SolarWind’s supply chain attack is a global hack, as threat actors turned the Orion software into a weapon gaining access to several government systems and thousands of private systems around the world. The malware had access to entire networks; many government and enterprise networks and systems face the risk of significant breaches.

The hackers also found their way into the Cybersecurity and Infrastructure Security Agency, or CISA, the office at the Department of Homeland Security whose job it is to protect federal computer networks from cyberattacks.

As more was learned about the attack and how it was implemented, it became clearer that, while this was an incredibly sophisticated attack, multiple prevention techniques could have been done to prevent the spread. This includes zero trust architecture, limiting access rights, implementing a cyber kill chain program and continued education to employees of the danger of cyber-attacks.

*Zero Trust Architecture*- Only known and authenticated communication would have been allowed.

*Limiting Access Rights* - Many organizations give full access rights to suppliers because it’s easy. This is part of the reason the malware was able to cause such widespread issues, but the update software did not need full access rights. Had privileges been limited, damages would have been less

*Cyber Kill Chain* - A series of steps that trace stages of a cyberattack from the early reconnaissance stages to the exfiltration of data. It has defined how organizations map out their security controls but also determines how they measure their cyber resilience.

Business Decision:

Businesses should implement the above three measures to significantly increase their cyber security. Understanding the cyber kill chain allows them to see what steps hackers must follow to gain access to systems. Use this knowledge to see what weaknesses your organization has and plug them.

The hack could also be the catalyst for rapid, broad [change in the cybersecurity industry](https://searchsecurity.techtarget.com/podcast/Risk-Repeat-SolarWinds-backdoor-shakes-infosec-industry). Many companies and government agencies are now in the process of devising new methods to react to these types of attacks before they happen. Governments and organizations are learning they must actively seek out vulnerabilities in their systems, and either shore them up or turn them into traps against these types of attacks.

Reflection and learning outcome

The action plan is the culmination of the analysis. At this point we have explained the data, the analysis goals, and the results. The idea here is that everything learned can be used to improve business operations by finding the best ZIP codes to invest money or solidifying your company defenses again cyber-attacks. While this is exciting it is also the scariest. To implement the changes, companies generally spend capital and are expecting a return on this investment.

In my current role as a financial analyst for a hospital, my job is to create pro-forma profit and loss statements the either prove or disprove the need to new equipment or programs. You are always putting a lot of trust in data sources and insight from subject matter experts that can carry a lot of weight in the results.

**6. Demonstrate communication skills regarding data and its analysis for managers, IT professional, programmers, statisticians, and other relevant professionals in their organization.**

Every course required some sort of communication, whether that be a presentation or a project write-up dictating your analysis and results. The importance of this can sometimes be overlooked. If you are not able to clearly articulate what you did and how you obtained your results, even the best analytics can fail if this falls short of expectations.

**Project Connection: IST 652: Scripting for Data Analysis:**

**GitHub Repository:** IST 652: Scripting for Data Analysis/Final\_Project\_Presentation.pdf

For the Olympic games project described above, both a project write-up and presentation were required. The presentation included a slide deck summarizing the key points of the analysis, results, and conclusion.

**Project Connection - IST 623 – Information Security:**

**GitHub Repository:** IST 623: Information Security/IST 623\_Information\_Security\_SolarWindsHack\_Presentation.pdf

The SolarWinds project required a group presentation that was no more than 20 minutes long. Here, no write-up was required which was honestly harder. This is because you needed to articulate to the professor and class solely by verbal communication. This requires you, as a group, to find the most meaningful aspects of your research so you can clearly give the audience a clear picture of the project goals, analysis, results, and conclusion.

Reflection and learning outcome

This is a day-to-day occurrence at my work. I am trained in finance but work with colleagues in other departments that need your analysis and results, but do not understand the conspectus like margin or growth rates. Over the years, and through courses here, I have been able to find ways to explain results so that my audience can understand what they need to. I believe communication is a practice that people should always be honing as this is where you can captivate your audience or lose them quickly.

**7. Synthesize the ethical dimensions of data science practice (e.g., privacy)**

Two theme that come up in this context are bias. For bias, the question often comes from machine learning and is it can produce bias results because of the assumptions that go into it. In IST 687, Applied Data Science, I vividly remember a class discussion related to this. Here were looked at a machine learning tool that was used to approve mortgage application. If the model used geographical data, it might take a person’s application and show that they house in located in a poor neighborhood, resulting in the application being turned down.

Reflection and learning outcome

This is a hard concept. One on side, you have raw data that provides factual transaction. So as a data scientist writing an algorithm, your mindset is probably that you are not biases because you are simply providing results based on facts. On the other side, these facts being imputed are still written with certain assumptions that can easily be biases. To me, one solution is to not rely only on the algorithm to determine action. A responsible company should still review to determine if the results are bias and may not represent the case before them.

Working for a hospital, this is a constant dilemma. Patients often come in that need immediate care but do not have insurance. Knowing it is possible you will not receive compensation; do you proceed to care for the patient, or do you turn them away? I am happy that my organization has written into its mission statement that we must strive to care for all people without prejudice to their financial situation. In my role as a financial analyst, I see the financial strain this puts on the hospital. Luckily, we can absorb this cost, but we are seeing it grow especially over the last two years. I hope we do not have to one day decide whether to cut this program.