MS in Applied Data Science

Spring 2020

Eric Rodgers



Contents

[Introduction 3](#_Toc41850613)

[Reflection 5](#_Toc41850614)

[Data Science 5](#_Toc41850615)

[Pre-Syracuse 5](#_Toc41850616)

[Post Syracuse 6](#_Toc41850617)

[Knowledge Gained 9](#_Toc41850618)

[Collect and organize data 9](#_Toc41850619)

[Pattern Identification (Data Mining) 13](#_Toc41850620)

[Statistical Analysis 13](#_Toc41850621)

[Visualizations 16](#_Toc41850622)

[Business Decision Implementation 20](#_Toc41850623)

[Communication 22](#_Toc41850624)

[Closing 24](#_Toc41850625)

[Appendix 25](#_Toc41850626)

# Introduction

My Background

For the last 12 years my experience and expertise has been in, tracking, forecasting and analyzing data on capital energy construction projects around the world. I have traveled to many parts of the world, speak 2 other languages other than English and I have worked in many challenging environments across the globe. At the same time of this evaluation, I can still recognize that feeling that something was missing in my life during some of my later work years. Some may have called it an interest or even happiness that comes from being fulfilled with your job, I would describe it as a drive or passion that I had lost while landing on a new tarmac, following the next project in another foreign country. A period of 12 months went by while I was abroad where I lost my family; both my mom and dad had passed away. This event marked a clear change in my life which took shape in my mind that something was amiss and that I needed to be doing something that not only challenged me but drove me to another level of success in my life, something that would fulfill me; something that would my parents would always be proud of.

Why Data Science

This new chapter of my life started in August 2018. I was picking schools to go back and get my Masters. After much deliberation and discussions with my wife, we wanted to find a route in which I could still use my background in project management but marry this with a more technical skill set, one that would scratch the itch of my more curious side but also avoid the trap of starting from scratch again. There was a lot of talk in the tech industry about analytics and there seemed to be a trend of more companies running “algorithms” and developing “models” for their data but at that time it had not meant as much to me as it does now. After running up a collection of New Tabs in the internet browser and searching DuckDuckGo for the best results of “Best Data Science Masters” I came across Syracuse University. Being originally from the East Coast, I knew of Syracuse but needed to know more about the program. After diving deep into the curriculum and developing a better understanding of how Data Science would help my career or in this case help me realize I would be taking a detour away from the Project Management life (what a relief), I applied and then was accepted into the program. From what I knew, this new program would challenge me in ways I have never thought possible but also help direct me in a new path in my life.

Twenty months later, I can see the MS in Applied Data Science at Syracuse has done just that. I have now seen the potential of so much of the data I had been using for years and how it could have been better harnessed to give better and more accurate forecast and trend analysis, or even just better visuals. The question now for me, is why not Data Science as too much time has passed for the question of, why not Data Science earlier? Data Science represents the future but not just specifically “THE” future, but my future. As systems become more complex, new algorithms are developed and companies continue to pursue AI, I can see that at the very least I have some of the tools to speak the language and the appreciation of time it takes to do this work.

MS Applied Data Science & My Career

Where does this leave me now, as I am about 1 month away from being completed in this program. I have started to garner interest within my own company, one of the largest natural gas providers in the US and Canada for a position regarding anomaly detection within the pipelines. For now as I am only a few years away from being 40, and plenty of experience in the oil and gas industry, it seems as though I will be destined to remain in the industry for some time to get some practical real-world experience under my belt. It comes to mind during my time at Syracuse what one professor mentioned that the oil and gas industry is not known for its analytics, and for the most part this is true; therefore it behooves me to ensure I utilize the skills I have learned to change that trend around and see to it that these last twenty months are put to good use. I also have an interest or we can call it a passion for natural language processing so post completion of the masters program, I would like to try and apply some of the new applications I picked up in those courses to the new potential position I may be offered in the future.

If all else fails, I would like to start taking the data that I currently work with, apply the models that I have learned to the company’ current cost and schedule data and ensure I can provide value in my current role. All of this will be to bide my time before I find the right opportunity to venture off and I achieve the data centric position I envision myself being in. Anything is better than nothing and when the company sees the value, there is bound to be a bite or two.

# Reflection

This section of the portfolio will speak to the program in its overall state and what or how I feel that I have taken from my role as a graduate student in the Applied Data Science master’s degree. I intend to cover many topics such as instructors, general class setup and topics of interest and how those topics played a role in adding strengths to my skill set. I also plan to discuss my evolution of coming from a non-technical position and transitioning to more of a programming and statistical mind-set but also discuss some of my current and ongoing weaknesses. Am I going to run the Data Science department of Amazon in the future, probably not, but will I enjoy what I do moving forward and scratch the itch of the technical career I had been seeking for so long, hopefully and most definitely yes!

## Data Science

### Pre-Syracuse

Prior to entering the MS in Applied Data Science at Syracuse, I did not know what the p-value of a x variable in a regression model anymore then I knew how to order breakfast in Japanese; this is obviously not one of the languages I have picked up over the years. One of the languages I have picked up within this program has been Python. Yes, some R was learned and yes R has a particular spot in data analytics, but it is no where near garnering the attention and support that Python has. Rather then repeating what the reader already knows though, why is Python so important to me and why am I so happy now I have some fluency in it? Prior to coming to Syracuse, I nearly completed a second BS in Cyber Security. I took this as a small achievement to better understand the changes in the data environment and it will be something I go back and complete once graduated from Syracuse. While enduring this program through Utica University, I had found myself trying to learn Python in a very ugly and non-intuitive way. Without going into detail, the program instilled a certain fear in me regarding Python. It was an intense fear that included an insurmountable mountain, aptly named Mt. Intimidation. Python to someone without a technical know-how would be difficult for anyone and then add on statistical measures and now we have a mixture that could only be described back in 2018 as impossible.

For most of my career visualizations of data has never been more then pie charts and bar graphs. I even reflect on some of my previous homework at Syracuse and see that these were non-acceptable accomplishes that I will graduate with but are not obviously my proudest of moments. But these are all the great things of the program, challenges, and reflections on what worked, what did not and what you can do to do better moving forward. Its an expensive and time-consuming way to do things but hopefully just through this tone of this portfolio, it is coming through as more then worth it.

My Pre-Syracuse time seems ages ago, its almost an effort to go back and look to see how little I actually knew about these topics. Pre-Syracuse I had an interest, an itch that had no way to be itched until the right steps were made to satiate my curiosity to the unknown. For myself, and hopefully the reader of this portfolio, hopefully it can be or has been instilled how very much I have enjoyed looking back at the past to better reflect on the present day.

### Post Syracuse

Nineteen months is a long period of time for anything and many things will have changed in a person’s life. My decision to take on a brand-new challenge in my life has added new opportunities and more complex thought processes that I do now on a regular basis during my job. Do I do data science on my off time, maybe not so much, but at the very least I can now say I have something that if I wanted to bide my time and learn more about the in’s and out’s of a new topic outside of work, I would be more than happy to do so.

Data Science not only within Syracuse through the last twenty months but outside of my sphere of influence, there have been new ways of seeing data, working with data and enhancing data; this new vision of the data world is ever changing and now I am a part of it more than I could have ever imagined to be. If that is not something to be excited about, I am clearly not sure what would. With all of the positive outlooks that I see in the future for myself, it would be good to see what are some of my largest takeaway’s from the program as well as some of the lingering issues I may or may not be able to fix leaving the program.

#### Strengths

First and foremost, the Syracuse program has enabled me to be able to touch large data sets and understand them at an exceedingly high level, very quickly. This does not consider the data cleaning process, rather this refers to the loading of data quickly, run a visualization and start understanding what the data says and what it does not. Quickly being able to find out how complete a data set is or how incomplete it is, can be a time saver in itself as no further analysis can reasonably be done on an incomplete data set. Data cleaning will be spoken more at length further in this portfolio document, but it should also be mentioned here.

Data cleaning, pending on the class, has been quoted to be in the upwards of 70 to 80% of a Data Scientists or Data Analysts time. This is quite a unique metric, one that brings to light a question that has no real answer that can provide a 100% result; why don’t we ensure all of our data is uniform so instant analysis can be done? Again, this rides the line where I can now ask such questions that I had never asked before. Is it easy enough to understand that data comes in all formats, sure it is, but its an entirely different topic to evaluate the time it takes to ensure that unstructured data is cleaned and can now be worked with. It can also be said that there is some value or a lot of value to unstructured data in that we can run algorithms to better understand patterns without the use of labels. The future has been said to include AI that does not rely on labeled data rather sophisticated algorithms that understand the data and pick up patterns natively.

Prediction is of course like looking into a crystal ball, sometimes its right and sometimes its wrong. Its 2020 and many of the most sophisticated algorithms were incorrect in determining the devastation or lack thereof for the current pandemic ravaging many of the countries. What does this say for the curriculum or my understanding of Naïve Bayes, SVM or logistic regression? Only that there will always be a limitation to models or an appropriate time for these models to be deployed and model validity is always hindered by the data (the completeness of) that it is fed. What I have learned is just because your model may appear to be sound, such things like the r squared value and p-values should be reviewed before any real business decision is made. This program has also provided code that helps get the student involved in code nearly immediately and then helps them understand the output. These of course are not the same between every class but that will be discussed in the weakness section.

Understanding where data is generated or how it can be and then how it can be loaded and then transformed in a database environment is a topic a lot of data scientist may not appreciate as much as running the algorithms for predictions. I say this as data tends to be just there for many of us in some way, whether its off of a website, prepared in-house at a company or something that has been self-generated. I know I do not work with a company that processes transactional data or regular updated data that is transformed or loaded into a central repository that can be brought into different models or utilized in other ways. At the most, data is something that is stored in so many places and this makes the ETL process impossible for some companies to implement. This program has enabled me the opportunity to appreciate and even practice building out multi-leveled data with the vision of bringing some of these opportunities to my current company.

I come from a world where most visualizations have consisted of pie charts and bar graphs so obviously if there is one thing that this program did well for me, was to at the very least question how data should or can be seen. Looking at previous work outside of Syracuse, I can see most project managers would put a simplistic bar graph together or a histogram with not much detail. I can see it meant something to them at the moment but that was because all context lay within their head as they presented the findings. Imagine though you are working with hundreds of thousands of rows of data and you need to present something that speaks about the data but you need to be able to interchange variables and summarize in ways that pie charts or bar graphs cannot. When you think about it, its not just being able to change the hue of a color on a scatter plot on the fly, its being able to present the data in a meaningful way. I would not say that the Data Visualization class gave me those answers, rather it gave me code to develop such plots, but it was the other parts of the curriculum where these concepts rang home for me. Develop plots and other visuals that mean something to the reader, not just yourself; your data will appreciate it and so will the people who consume it.

Reporting & thoughtful engagement are in my mind two topics that mean the same thing and they are the final strengths I believe I have picked up from this program. When I see data now, I think about its structure, I think how it can be visualized, but I ultimately think that this data will be reviewed by someone who can make a decision about it, and that person needs to know what the data is saying without me making a formal document they won’t read. If I have learned anything through the years that working with technical minded people is one of the most difficult things to do for business professionals, i.e. engineers and programmers. I come from the other side, the non-technical side for the majority of my career and I can see that people who have the technical prowess often struggle to deliver their ideas in a more consumable format. Building out models and applying different x variables to study the movement of the dependent variable is all jargon and topics that a business professional typically does not want to find themselves drowning in. Instead they would rather be able to look at the findings in a way that makes sense to them. Whether this be a report or a dashboard with current sales forecast or predicted sales numbers, they want to know how the data is affecting the business, not how you came up with the answers. In this, I can see that the program at Syracuse has assisted in being able to speak intelligently about certain topics and I have much more confidence in doing so. This has come in the flavors of data collection, data visualization and my most appreciated accomplishment, coding.

#### Weaknesses

The statistical side of the program or what I have taken out of it has been a very mixed bag. Though this is a reflection on my takeaways, obviously this is an equally important evaluation of the program for the University to understand what one of their graduating students is saying. I will be an alumnus of Syracuse, I take the name and reputation with me, so my output I believe is of significance at to the very least to the present reader of this portfolio. With the additional jargons of above, the program did not sit on statistical notions for too long before jumping into the next model type, so my weakness still resides in some of my own statistical prowess or lack thereof. Obviously, this will vary between all students as some students do this kind of work for a living and now get more time to appreciate the math and statistics behind the algorithms while in school. As this was something that I am brand new too, this is a topic I will continue to need to learn and study post-graduation.

Additionally, the program skirts around heavy usage of Python where some courses start very easily and ramp up, other classes expect some level of intermediate expertise to get through comfortably. As many of the classes focus on R, which is much easier to get into, the Python focus classes were far and few between; only 4 to be exact. With more of an emphasis on Python I would have hoped that I may be a bit more fluent in the language but I can certainly seeing myself engaging in different courses post-Syracuse to learn additional skill sets to further the topics that have really interested me, particularly in Natural Language Processing.

# Knowledge Gained

This section will cover in detail the projects that offered the most comprehensive additions to my new skill set. I will additionally go over a few homework assignments throughout the course that had large impacts on my studies, along with a few final projects.

## Collect and organize data

Across the course catalogue, many of the projects and homework assignments focus on collecting and organizing data, more heavily on the latter then the former. Obviously to stream-line class time and ensure it is used efficiently, the assignments were setup so that code whether or working or broken was provided at a high and detailed level but more importantly data sets were provided to use immediately to get the code running and the output reviewed. This would mirror real world scenarios if a company was setup to utilize data that could be pulled easily and it also mirrors a scenario that nine times out of ten, some data cleaning or manipulation is going to have to be done. On the flip side, very few courses outside of the 700 level courses required fetching data sets that were not either readily available or were not Kaggle approved. Some of the classes did involve bringing in outside data if available and even combining data sets based on similar variables.

The focus on cleaning was strenuous at first, but as classes moved along, and new challenges approached, I could see myself being able to identify that working in pandas or sub setting a certain part of the data for ease of use came more natural. I can still see myself struggling through regular expressions since it is just a vast world of possibilities but utilizing these for cleaning became an essential part of streamlining some of my data sets. Organizing data was also an essential processing after collecting and cleaning. Organizing the data whether your dictionary needed to be changed to a matrix or data frame was something that came with the requirements of the project or assignment. If no literal summarization were necessary, then maybe using the most basic data structure would be ok, as long as it was cleaned! Coding to manipulate these various structures was challenge and required a lot of research but while browsing through many code forums, you realized quickly, this problem was not only not unique to the community it has been answered 100 different ways before you even thought to ask it.

Where does this leave us, as we now have collected, imported, cleaned and organized the project; no where else other then to discuss the projects that made the most use out of the collect, import and clean process mentioned above. One key assignment and one key project will be discussed in the next section, and these will be mentioned throughout this document regularly through the other sections along with another sample or two from the program.

#### Key Projects and Skills Gained

The first assignment comes from Big Data Analytics and it requires the student to look over a small data set but also add to it. It is essentially a harmless data set that requires only a small bit of cleaning, for example, there were several rows with no data in some of the columns, the columns with money would need to be converted once they came into python as integers. This all came into play as the assignment asked that we predict the salary of a coach at Syracuse. As the data set was limited in what we would call predictor variables, the class was asked to obtain additional data. Of course, we would have to merge these data sets and we would do so based off of the school name. The most difficult and time-consuming part of the assignment was data gathering, but this was also the point of the assignment. Collect the data, import in the data and then organize it in such a way that the data frames could then be used as a whole even though they were not originally together.

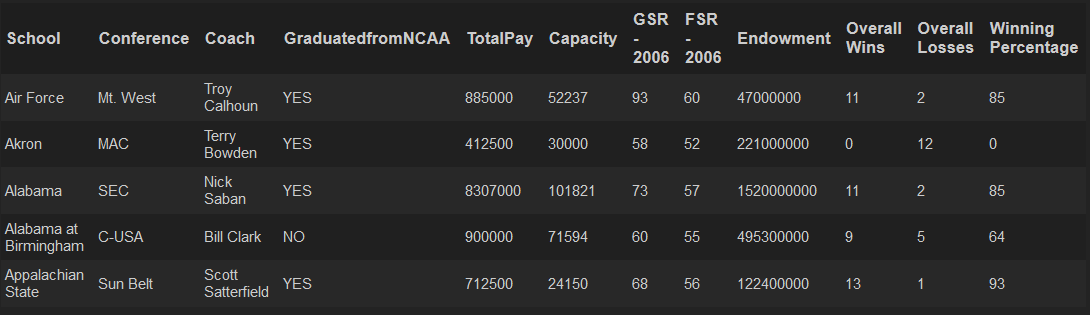
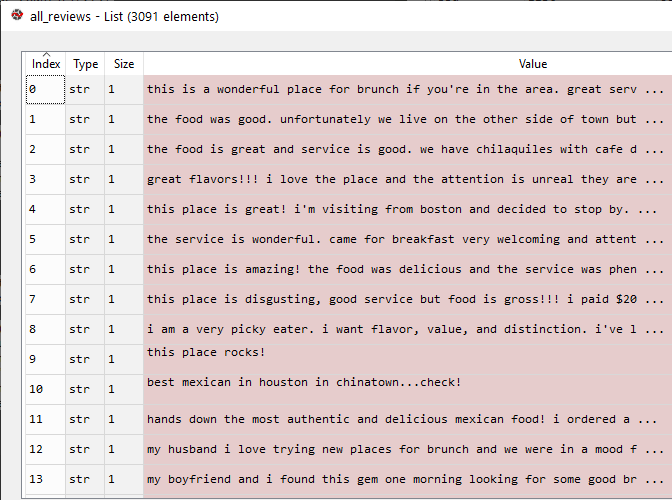
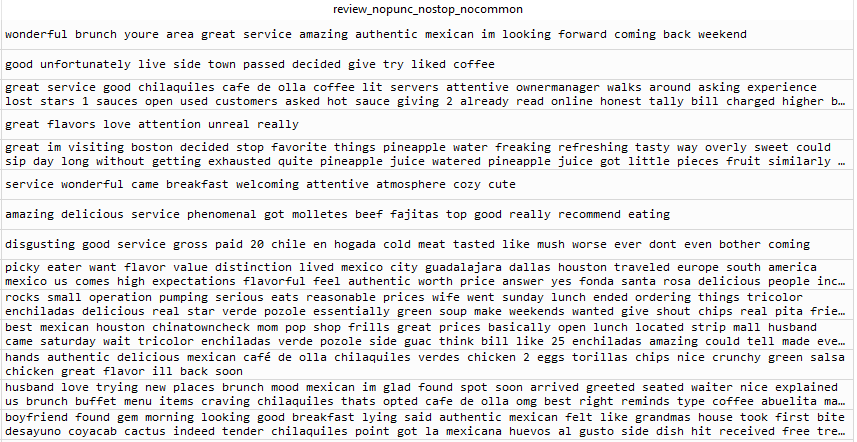
The skills gained here came in a multitude of flavors. First, Big Data should really be taken last in the program, and the reason for this is that it brings many of the concepts together from previous classes, yes even the business analytics courses. I know that there were some key concepts I had to go back over to round out my understanding of the concept and also make sure it was fresh enough that I still remembered the concept well enough to apply it. Secondly, this was the first class that I had to bring a multitude of data sets together that there was not necessarily a link to and that was exciting. It was challenging but it also forced me to think about the usual places that would have said data and then find out the data did not exist in the forms I needed. This also challenged me to code and ensure I could merge successfully and then look after a large data set with more variables. It also promoted me, not demanded of me to think about the assignment, i.e. what would be some defining factors in predicting the worth of a coach to a University. This for me was obviously not a data set I had any interest in, but it gave me the opportunity to think out of the box and then apply it to the model at hand. In Figure 1, there is a data frame that contains additional variables from outside of the range of the original data set. These variables include, stadium size, whether or not the coach graduated from a NCAA college and several others. These variables had to be cleaned and then brought into to then be merged with the original data set.

Figure 1: Original Coaches data set plus other variables added at the end.

The second example is a final project for the Text Mining class. Here the ask from the student was to conduct some type of research across any of the social media platforms or other data collection warehouses, i.e. Yelp, Zillow, etc. My project was to collect some data of restaurants here in Houston, Texas and find out what were some of the key drivers for a customer to be visit that particular restaurant. It was also to understand if there were key drivers that differentiate types of restaurant, i.e. Vietnamese and Mexican. Lastly it was to analyze if people preferred more expensive food with more “quality” or would a cheaper more accessible establishment garner more positive sentiment. The easiest part was writing code (taken from a coding forum) and scrapping Yelp of over 5,000 reviews. The hardest part and least satisfying was cleaning the data and then finding out the vocabulary of most patrons were the same across all restaurant types.

The skills gained in this class came from investigating code that I was not familiar with but then merging it with code that I was. Scrapping data is something that is common practice, but I had not had that much opportunity to do so. Merging it with the code that I had worked on before though was fresh and exciting and made the experience of finding out the sentiment of the restaurants around me that much more enjoyable. The skills learned here allow me to think in such a way of viewing data as either a bag of words or a corpus and understanding that I need to clean more then I typically would on other sets of data. There is no need to capitalize or take out the pre-fix of numbers and words present a challenge that other sets of data do not. As well, when working with words, one must be mindful that utilizing the built in dictionaries of both Python and R may not be adequate or reasonable for your analysis, therefore taking a look at a dirty set of words and looking at the frequency first may give you that first insight. Finally, this project presented the opportunity to work with real world, real-time data that could or would be important to someone in marketing or even the business owner of some of the restaurants. This was an opportunity to see the fruits of our labor in the context that it may even help someone make a decision that could impact the success of their business.

Figure 2 and 3 represent the scrapped review data from Yelp. The data was needed to be normalized to ensure words were not duplicated and the sentiment was not buried within the noise. In this, all words were lowered, numbers had to be removed as did punctuation and custom stop words. Stop words were necessary per restaurant type as we did not want words like taco, or beef to overtake the percentage of words in a particular restaurant. The intent was to focus on what made the restaurant special or not so special in some cases. Words like service, delicious, friendly, and return were our objectives as they represented what the customer was saying.

Figure 2: Uncleaned python list of Yelp Reviews. Figure 3: Cleaned python list of Yelp reviews.

## Pattern Identification (Data Mining)

Knowing how to spot patterns and how to interpret them are key features in the data science skill set. Patterns can be found in both structured and unstructured data, and although the delivery of the models will be different for each type of data, what we are seeking is essentially the same, what message is the data telling us. Patterns can be picked up in different ways, which include the three topics below, statistical analysis, visualizations, and data mining. We will cover some of the projects that covered these types of pattern recognition and how they were used in their respective assignments and projects.

This section will be broken down into three sections, covering, statistical analysis, visualizations, and data mining. As before, the Syracuse Big Data Assignment regarding College Football Salaries will be discussed as will various visualization examples that were apart of some of the assignments. Finally, the same assignments will be referenced for data mining examples.

### Statistical Analysis

Breaking down a data frame whether in R or Python helps to look at the basics of what the data set is telling you. For example, the df.describe() function in Python gives you a quick look at the data frame you have pulled in and what are the most prominent features; such that the mean, maximum, median and minimum values can be seen so that you can get a very brief but summarized view of your data. This may help in being able to make a quick determination to see whether the data is skewed, is it evenly distributed or does it even tell you what you thought originally going into the data set. This function of course should be followed up with is.NA to ensure that your dataset is at the very least complete, and if it is not, it is time to dive deeper into the data before trusting the full summary statistics.

Statistical analysis can also involve modeling, although this is farther down the process or chain of events. The p-values and r-squared values are extremely important in understanding how relevant the independent variable is to your dependent variable while the r-squared or adjusted r-squared value will assist in telling you how accurate your model is. Of course, the idea to get the highest model score with regards to r-squared or adjusted r-squared is great, one must be mindful that the data is not overfitted or that (if they exist) junk dependent variables are not driving up the r-squared score. One must also check the p-values as the independent variables may have rather high p-values, meaning they are not statistically significant which then would result in those variables being replaced or removed all together. If the choice is to remove, then other independent variables will be re-weighted, so different combinations of independent variables should be tried before coming to a full conclusion. This is where the original step of looking through a complete data set and understanding where some of the summary statistics stand, so that better decisions about the model can be made. What if the data is skewed, would running a model even be worth the effort?

The coach’s salary prediction will be discussed below to demonstrate the statistical summarization and model output for this section. Additionally, the Yelp review project will be briefly discussed on the model accuracy and why the findings are statistically significant or not.

#### Key Projects and Skills Gained

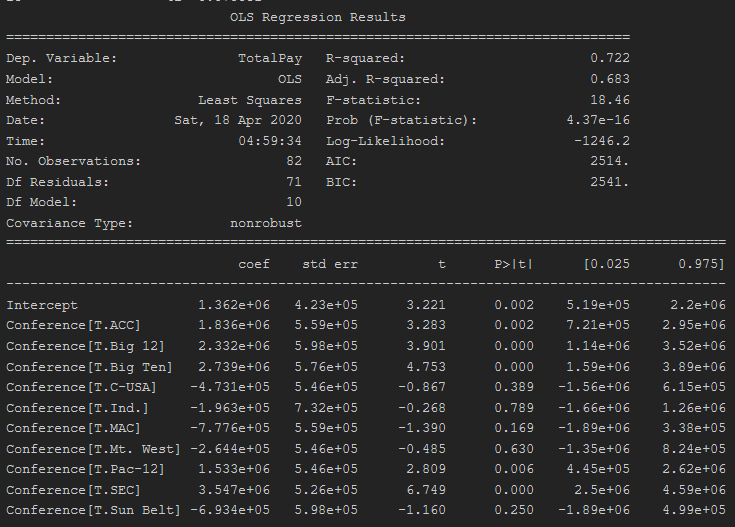
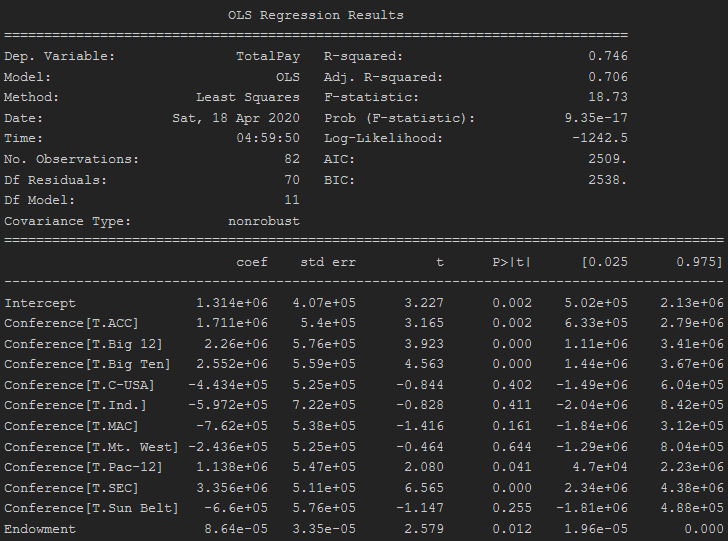
The Coach’s data set after being cleaned and added to was used in a liner regression model. After adding more variables for prediction, we wanted to see what was important or what was not in predicting the coach’s salary.

Figure 4: Output of Liner Model

In Figure 4, the r-squared shows a decently high score as the only independent variable to predict a coach’s salary is the conference the school is in. As this expected to be the strongest indicator for the prediction, the score is rather good and generally speaking the p-values are significant for at least five out of the ten conferences. In Figure 5, we run an additional an additional independent variable called Endowment (donations received by a University), to estimate the impact on a coach’s salary. As expected, adding an additional independent variable drove up the r-squared value, as well as the adjusted r-squared which we will ultimately use since we are using multiple independent variables. The p-value of the endowment is also low indicating endowment it is statistically significant. Finally, what is not shown here, during the analysis the coefficient is weighted and then figured into the salary of a coach and what was found was there was more then a five hundred thousand difference between schools with sizable endowments and ones that did not. Based on two different models, we can at least surmise, that having more money at a school will definitely impact a coach’s salary and now we know by how much.

Figure 5: 2nd output of a liner model with additional independent variables

For the second project to demonstrate statistical analysis, the topic of vectorization of a corpus will be discussed. As the Yelp reviews were taken from the website, and cleaned, the terms had to be vectorized for processing and prediction. This process takes sentences within the corpus and breaks them up in bite size pieces. This process is extremely important in understanding sentiment but the proceeding step of cleaning the data is equally important. The reason being, too much chopping of data can result in inaccurate results, so custom stop words are a good idea, but being mindful of what you take out and addressing this to your reader will be important as this could effect your model. Now vectorization implies that a word structure like, “I want to eat spicy chicken today” could be broken up into chunks like, “eat spicy chicken” which is a trigram, or “spicy chicken” which is a bigram, or even just the word “spicy” which is a unigram. Once this vectorization was done on the data set, the model was then to try and predict what kind of restaurant these words belong to. The idea was to teach a model what to look for so that when a new review came in, it would understand where it belonged to. The statistical importance here is the relevance of each model. As the prediction was ran against models that looked at vectorization of a unigram, bigram as well as a trigram, the different groupings would assist in identifying what words belong to what restaurants. In Figure 6, we can see our scores were different depending on whether our test was based on sentiment, restaurant expense or restaurant type.

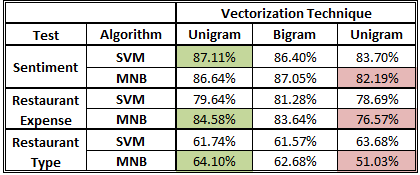


Figure 6: Prediction success based on vectorization.

As we can see the difference between a 61.74% success rate compared to an 87.11% score is relatively large. This means our model was much better at understanding how to read sentiment then it was to identify the restaurant type. This is important as if we wanted to put in additional data for let us say a new restaurant, we would have to be mindful that there would be at least a delta of nearly 25% when prediction the type of restaurant and sentiment. The clear issue was some of the overlapping words that the restaurant types shared which confused the model. Better tagging of the data would have to be done to get better results it was concluded.

### Visualizations

Visualizations serve two purposes in data science. First, they can really make a report or dashboard standout, i.e. make it more appealing to its viewer. There is nothing worse than have collected a lot of data for it only to hide behind ugly graphs, texts and other visualizations that do not make sense or are not at least appealing to look at. The second purpose is to tell a story, one that gives the impression that the data has been thoroughly cleaned and summarized and what you want to tell the reader is easily communicated, there is no ambiguity.

#### Key Projects and Skills Gained

In Figure 6, we can clearly see the average pay among the college football conferences, but we can also see the maximum and lowest values of the data set. Also, the dots on the side represent the schools so we can understand some of the clustering of those data points and where they align within the distribution.

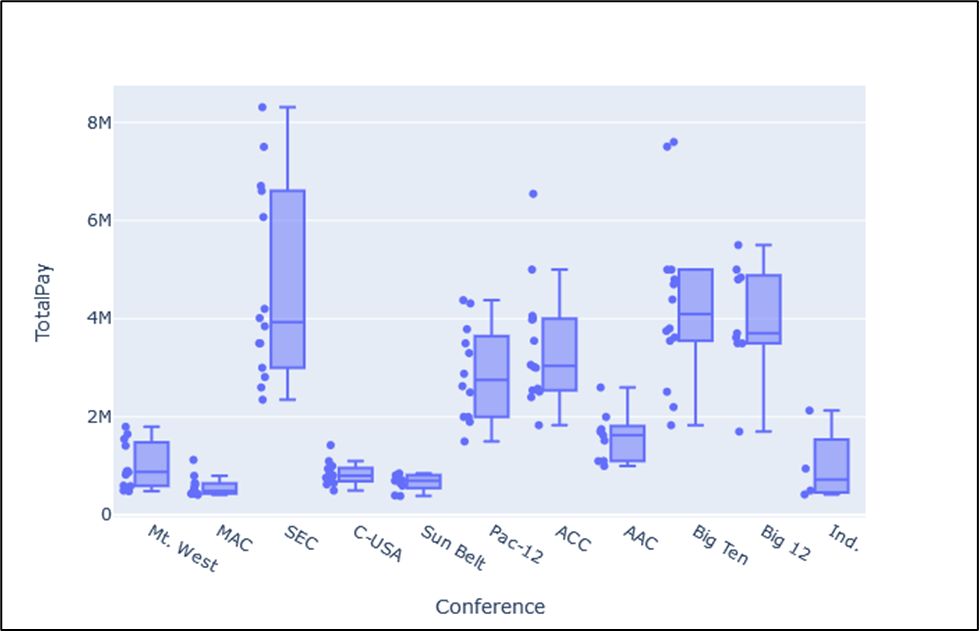


Figure 6: Total Pay of a Football Coach across Conferences.

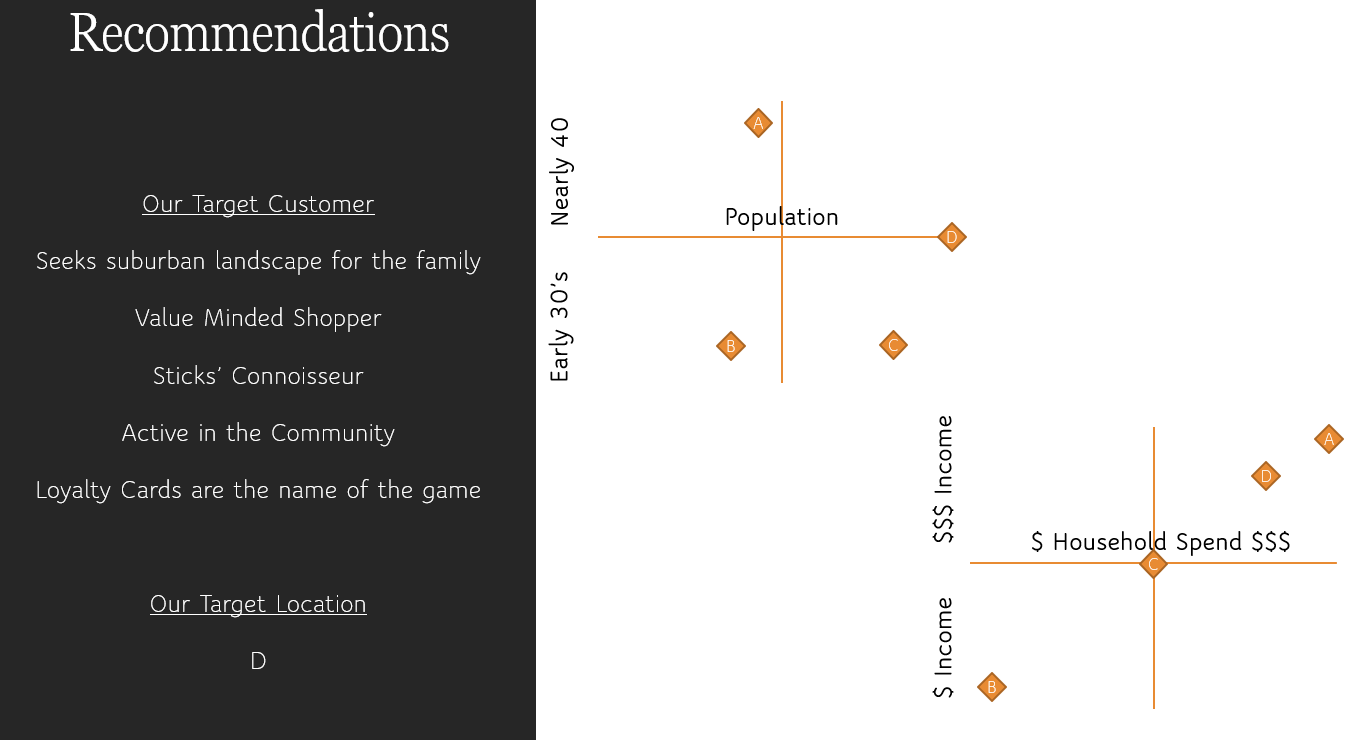
A screenshot of a cell phone

Description automatically generatedSticking with the same dataset but summarized on different variables, in Figure 7, we are now looking at Graduation Rates across the conferences. Both visuals can clearly tell the story and we can see a clear trend. We can easily see which markets pay the most in schools and we can see the upward trend for the scores across the same conferences.

Figure 7: Graduation rates across conferences.

Changing gears and projects, we will now look and some different visuals that give the reader a clear view on what we want to say about our data. Taken from a project assignment for the marketing analytics course, we want to give a quick idea on who out customer is and how did we identify the next restaurant location for our franchise.

In Figure 8, we can see what makes up a customer of our restaurant bit then we also plot out where do the locations line (represented by letters) among the types of customers. This was then used to look at what customer type currently eats at our restaurant to then see where we can continue to grow the brand. Why open up in a new area that does not fit your demographic?

 Figure 8: Recommended location to specific customer of restaurant.

Our next example takes a look at word clouds. Two projects will be looked at to show some of the themes of the word clouds and then we can discuss what they mean.

In the first word cloud, Figure 9, we can see the major themes that are within the Kickstart campaigns. This data includes the most mentioned words within each campaign without considering sentiment. What is represented here was there were a lot of campaigns that had to do with different types of media. It can actually be seen that the majority of words in the center have something to do with a campaign that required someone’s creativity, which in essence is what Kickstart is all about! We will discuss a bit later more about the actual data found in the Kickstarter campaign project so that we can elaborate a little bit more on what does the word cloud actually mean for the data set.

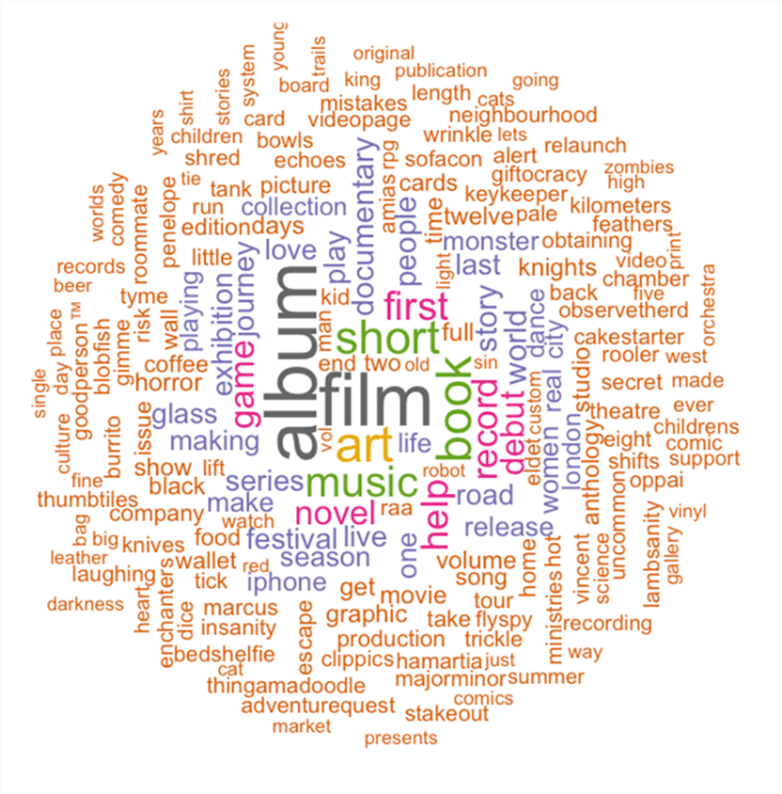


Figure 9: Word cloud of Kickstarter campaign data

The second word cloud in Figure 10, comes from a cleaned corpus, of Yelp reviews for the project that was looking at the sentiment of cheap and expensive priced restaurants. This is the same project as mentioned in the collect and organize data section. Take the written review and converting them to a corpus was the first step after cleaning which then led to being able to identify the most frequent words and then converting this into a word cloud.

Figure 10: Word cloud on cheap restaurants

From this word cloud regarding customer reviews on Yelp for the cheapest restaurants in Houston, we can see very quickly that the food is good, if not also great, people still believe the food is fresh, it is delicious and even many people love the places that were chosen in the analysis.

## Business Decision Implementation

Boiling data science down is a matter of looking at the functions of the science and looking at the reasons why we care so much about the data. What are the decisions we make when we have put together an algorithm or show a liner regression line, someone has to do something with it? This section will discuss some concepts that were learned in the program that gives more purpose in a business setting rather then just a technical sense.

In the real world, most people will be working for a firm that has hired on new staff or has an established team that runs analytics. These analytics will help drive new sales, new business initiatives or project decisions, i.e. my world in the oil and gas industry. As my industry becomes more focused on analytics, there has been a steady approach of understanding the reason for delays of projects, the predictability of a particular machine going out of service or even how much churn the company expects. In the examples below, I will go over similar concepts even though they are not a part of my daily duties. We will review, the salary prediction assignment as mentioned before, the assessment of yelp reviews for the restaurant project (also mentioned before), as well as a Kickstart project that was done for an analytics class.

#### Key Projects and Skills Gained

In the first assignment to be discussed (mentioned in collect and organize section as), the goal was to predict the salary of a Syracuse football coach’s salary. This prediction took the form of a liner regression model and took in several factors to predict the salary. Now where or why would this be beneficial to the management of a University? Well of course, the school wants to be competitive but what we found is that those wins come at a price, since wins are typically attached to talent; and talent normally stays in the conferences with the biggest market shares. Then how does a school lure a big-name coach to a smaller conference, and the easiest answer was, they do not! This is a business decision that the school has to realize their market is not big enough to house the amount of fans they want to draw in, so there is easily a break even point that they can no longer or ever generate enough money to bring in to pay the coach; so then who will pay? Again, the easiest answer is no one! Big-name talent does not belong in smaller conferences because the business sense is not there. Now if a coach does not mind taking a pay cut to move to another conference (generally never done), then it may make sense at the very least for the school to pick that coach up. The great thing about is, with the visualizations that were generated before the model was even kicked out told the story. Syracuse in the example does not have a big enough stadium, nor enough fan fair to take in a coach that demands the nearly double digit million-dollar price tag. In fact, the current head coach is ranked 65 out of the 100 coaches in an NCAA website report. The business decision, do not bring in talent you cannot afford, rather spend money on updating and progressing curriculum.

The second example is a final project that was covered above, and this was looking at the sentiment differences between an expensive restaurant and a cheap one as well as if there were any identifying characteristics in the yelp reviews between different cuisines. Of course, the findings were that the vocabulary was a limitation in the yelp reviews, as many patrons utilized the same speech, so that many of the restaurants illustrated main themes such as tasty, cheap, service and other non-descriptive words. What was the take-away then if all the speech was similar, well oddly enough, there were a few clever findings in the dataset that are worth mentioning First, within the five thousand reviews scrapped, there was evidence that most people were equally satisfied with cheap food compared to “expensive” food.

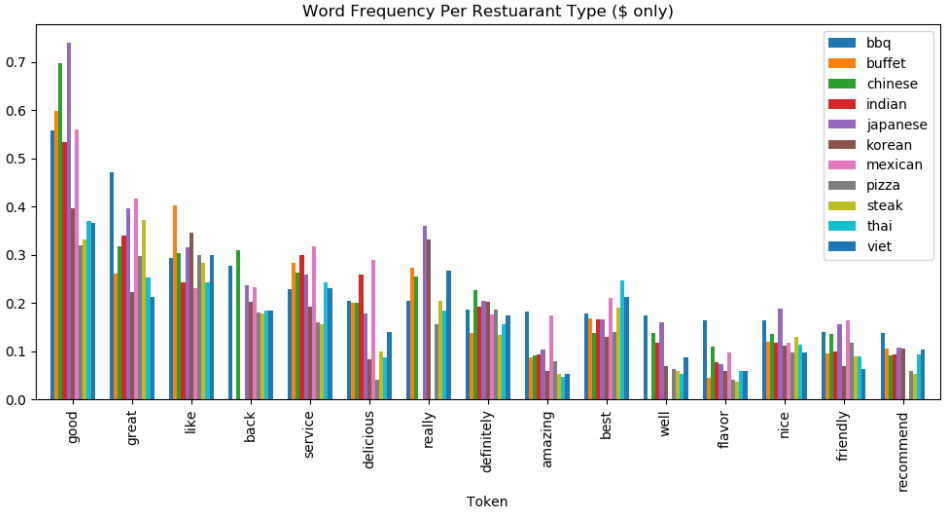


Figure 11: Word frequency against restaurant type

We also found that in order to possibly get better results, we would have to scrape many more restaurants and bring in more then the 10 different restaurant types. This would include maybe even looking at a different city altogether to see if the same thing could be found. At the end of all of this though, for a Houston business owner, they can easily focus on keeping the food at a reasonable price, the service good and the customers will come in and give a good rating. Food quality can be affected by the service and the price or at least the perceived effect on food quality so its good to be mindful of both.

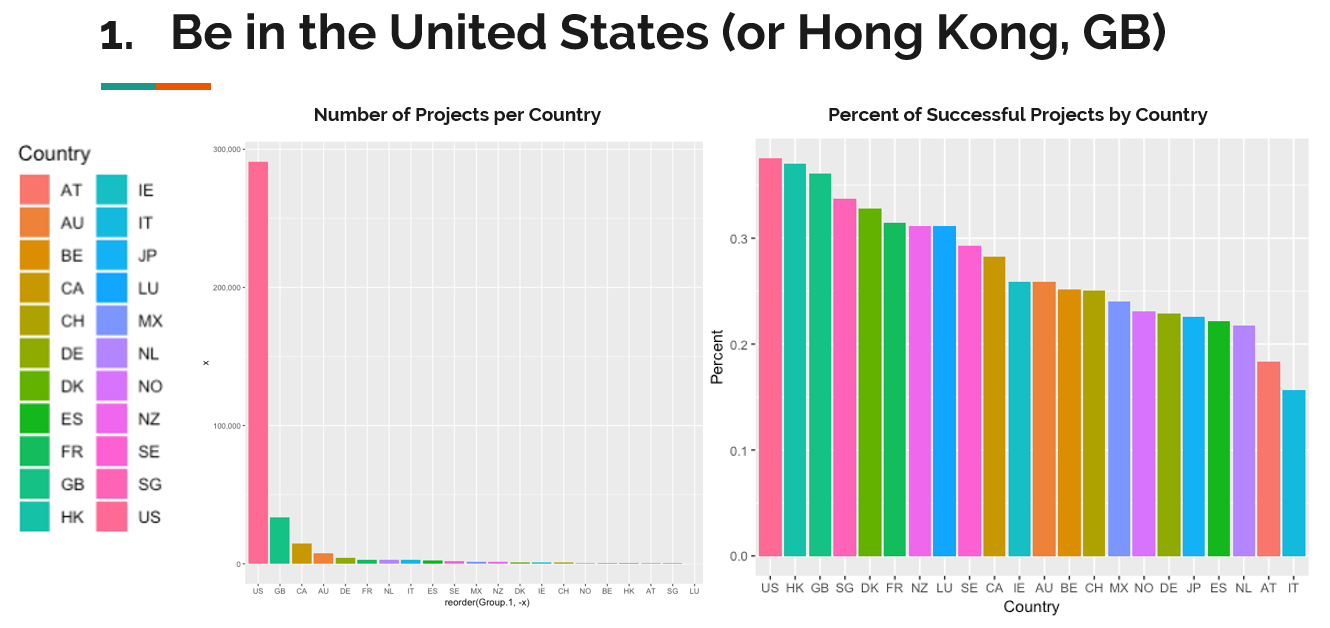
Now that we are all hungry at this point, lets discuss the Kickstarter business model. Kickstarter is a mixed bag of dream visions and disastrous projects that went down the drain. It is a risky proposition if you want to start a new business, so the Kickstarter program put most of the risk on the consumer. The project wanted to take a look at what kind of projects were a hit as well as what countries lead the way in Kickstarter projects. We also wanted to take a look at, what was the average campaign length of a successful Kickstarter compared to one that was not. What we found was that of course Kickstarter was more successful in open business environments with open idea exchanging. Markets like Canada, US and the UK led the way, while countries like China, the Middle East and others did not have such successes. This of course led us down the path of wanting to know, if you are a small business or wanting to start one, or even develop a new product, what did you need to sell, how did you need to sell it and how much money were you going to need.

Figure 12: Countries with most success with Kickstarter

What we found was as illustrated above, keep your goal in sight, do not get to far ahead of yourself or make goals to lofty and stick with categories people want to back. Most technology campaigns set out to do too much and failed most of the time, so this was a category that failed often, but dancing and theatre, these were easier to sustain therefore more successful. As well, we found that starting your Kickstarted campaign in December was a more successful month whereas in July was an awful month do start your campaign. Many things were learned from this project and obviously we (my team) grew a bit wiser if we wanted to suggest to someone a Kickstarter campaign, pretty cool stuff!

## Communication

Delivering data findings is the last essential element to be reviewed in this document. Communicating to people who can make decisions based on the code you have ran and output you have created (in whatever medium that is) is necessary for a data scientists or analyst to be able to do in a natural way that not only gets the point across but leaves nothing ambiguous in the mind of the reader. In this, it is essential for the visualizations to capture what is being said within the data without words, it should clearly communicate your ideas. This is obviously the best way for a data centric person to communicate their findings. When all else fails, a clear and concise power point or dashboard that represents the data clearly will work to bridge that gap.

When communicating with other data centric professionals, it is also essential to be able to discuss, collaborate and learn from one another. Group work similar to what was performed in many of the classes is a key aspect of the real world and something that at times we all don’t enjoy but it is a necessity, especially when work load and deadlines are present. Changing gears but staying within the same realm of team-work, sharing code and communicating clearly what your code does, where it was derived from and what you are planning to do with it is also key to be successful when working and communicating within a team. Also setting expectations within a group project, clearly defining goals, desired results from each team member and setting a schedule to accomplish these goals are all important factors.

#### Key Projects and Skills Gained

The classes and projects within this program all required clear communication, not one class requiring more then others. Working throughout the assignments we generally a single person affair, although there were exceptions. These were not treated differently then a project though. Within the Zoom environment, many groups I was apart of, set up regular meetings to discuss the data, how progress was going and even made some sessions into working sessions. This was needed to ensure the deliverable met everyone’s expectations as well as answer or come to an agreement on any outstanding questions.

Communicating code was also something that took many different forms throughout the program. For example, early on in some classes we would share Jupyter or Python code natively but when it came to other classes, there were different avenues in sharing code, such as Google Colab. In fact, I found out about Google Colab in my last semester and would have appreciated the knowledge beforehand as it would have helped with not only running heavy code but being able to work in a shared environment with other team members. This is something I plan to pass along to my company in the case my company requires a shared environment. The other side of coding is just ensuring your code is clear enough for others to read through and keeping your comments clean and concise. I took pride in how clean my code was the longer I was in the class to the point I ensured that was something I kept in the front of my mind as even in older code that I had to review in later classes, it was difficult for me to follow what I had previously done. These are most likely rookie coding mistakes and this I understand and now when I review others data or my current, I want to be able to read through code, understand it without too many questions and be able to run it without too many issues. No one wants to find out how to run your code if it is supposed to be just a review and run of the code type exercise.

Lastly, communicating within the classroom setting is something I still work on. Being introverted is something that cannot be controlled but it can be overcome in certain situations. In many of the classes, I typically utilized my talk time in a question format as I never wanted to seem to sure about the topic at hand but wanted to clearly articulate that I had some understanding of the topic. Communicating also within the classroom setting is similar to real world work environments where people may have different viewpoints then yourself as well as different ideas that may conflict with your own. The classroom setting illustrated the idea that you should always keep an eye out for all the different points of views and be ready to discuss them respectively and intelligently.

# Closing

In closing, this portfolio document takes on many different meanings for me. It’s a culmination at a high level on what I have learned and specifically been asked to highlight but its also a symbol that represents a completion of a new beginning for my career and ultimately something that I know my parents would certainly be proud of.

My intentions as mentioned above is to allow the last twenty months to carry me forward into a new career as well as unlock a new skill set that would enable me to follow a passion of mine that had never been fulfilled until now. The next twenty years of working either for a company or for myself will represent the same passion as I had today, as I believe I owe it to not only to my school but to myself to follow-up on my goals of the future.

I want to thank the Professors at Syracuse University, both in the Asynchronous and Synchronous mediums. It is much appreciated to have taken me into the Program and I will continue to use the lessons learned from the program moving forward, as well as build on the foundational blocks that Syracuse University helped me lay.

# Appendix

All aforementioned projects and assignments have associated code on the Google Drive link below.

Link:

<https://drive.google.com/drive/folders/1iI5hjNkCPdLpmhHEDd6v9pc6CdRzPQKX?usp=sharing>

Note: Repository has been made Public

Project List

Yelp Restaurant Review Project

Class: IST 736 Text Mining

File Name: Restaurant Scrapping Project

Required Software: Python

Kickstarter Project

Class: IST 707 Data Analytics

File Name: Final\_Project\_Code.R

Required Software: R

Assignment List

College Football Prediction

Class: IST 718 Big Data Analytics

File Name: Rodgers\_\_Lab1

Required Software: Jupyter Notebook (Python)

Marketing Restaurant Analysis

Class: MAR 653 Marketing Analytics

File Name: None

Required Software: None (Graph was generated in PowerPoint)