Portfolio Milestone

Daniel Scholnick

dscholni@syr.edu

SID: (526754961**)**

December 2020 Graduation

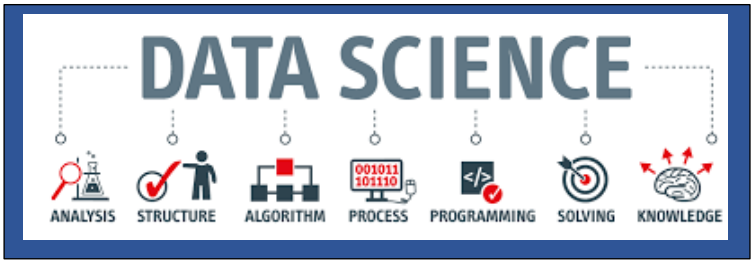
SYRACUSE UNIVERSITY

Introduction

Deciding to transition from a career as a professional chef for over twenty years to a data scientist was a leap of faith. Many people questioned why and how I would fit into the data science field. At the beginning, there were definitely more questions than answers. My first class, Intro to Data Science, allayed many of these concerns.

Professor Saltz began the class with a discussion about what tools you need to succeed as a data scientist. He emphasized the need for an enduring curiosity, the drive to find answers and the ability to communicate well with not only data scientists, but all invested parties. I had studied science extensively as an undergraduate and felt immediate comfort with this scientific process.

In addition, twenty years working as a chef was an act of both science and artistry. My methodology for cooking followed the scientific process. My clients would describe what they wanted in detail; I would research numerous recipes; and then I would create a new recipe that matched the client's goals. I would then cook the recipe, listen to the client's feedback, and make adjustments accordingly. The open communication was the critical part. I learned early on as a chef that I was not cooking what I liked, but instead what the client enjoyed. With professor Saltz's introductory lesson always in my mind, I had the confidence to dive into the data science program.



Analysis

I will present high-level descriptions and some of the lessons learned from five of the projects I completed in the Applied Data Science program. These projects each represent a key element or elements of the program's learning goals.

Dan's Sleep Study Project

The first project I will review was completed in the Data Analysis and Decision-Making class. The objective was to design and execute a case study of my sleep process and then to present it in the form of a story board. The study was to follow the Six Sigma DMAIC format which is a traditional science approach.

I believed that I was not getting the recommended 7-9 hours of sleep the National Sleep Foundation recommended and this was affecting my life in a negative way. I created a process map to delineate the inputs in the process that I believed could have a potential impact on my

sleep (Figure 1). The goal was to determine which inputs and their respective values were critical for me to achieve 7 hours of sleep. The inputs were predominantly continuous but did contain one discrete measure. This represented the **D**efine phase or D in DMAIC. This project emphasized the importance of the Define phase. This phase is the opportunity to probe the subject matter, outline the design of the project and question the value of the insight to be gained.

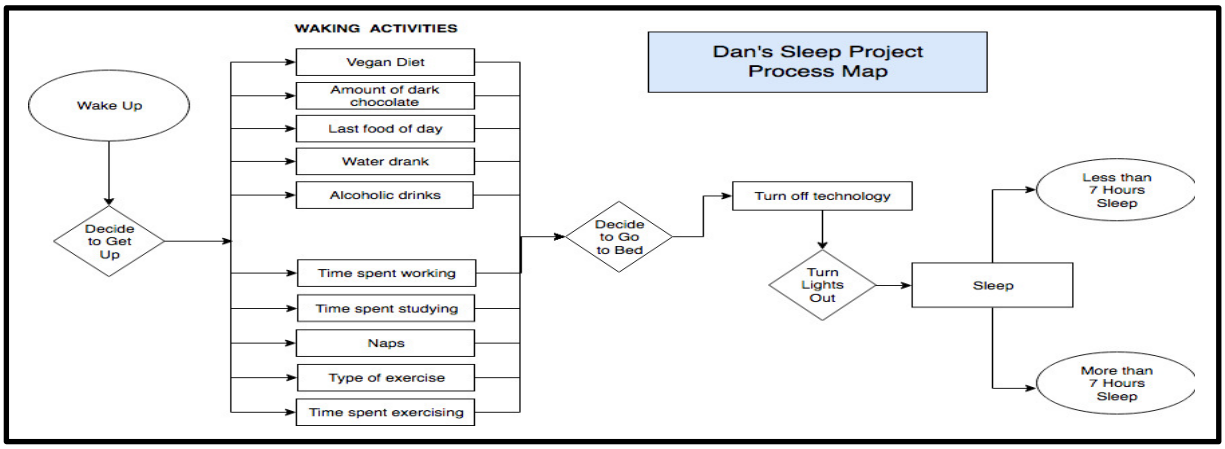


Figure 1. Process Map

In the **M**easurement phase, I began to record the inputs on a daily basis (Figure 2). I utilized the Sample Size formula to help ensure the quality of my results. Using this formula, I calculated that I would need thirty nights of data to attain 95% confidence in my results. The data from the thirty nights were then used to calculate the initial Sigma Quality Level of 1.3 where the defect was less than 7 hours of sleep.

Date	Water (ounces)	Chocolate (ounces)	Vegan diet (y/n)	Exercise (minutes)	Exercise type	Naps (minutes)	Study (minutes)	Work (minutes)	Alcoholic servings (count)	Time between last food and lights out (minutes)	Time between technology and lights out (minutes)	Lights out (time)	Sleep (minutes)
4/15/19	68	2	n	50	walk	25	330	110	1	135	15	22:45	185
4/16/19	62	1	n	60	walk	30	315	25	0	125	35	21:10	315
4/17/19	72	1	n	75	bike	15	305	45	1	145	25	22:00	405
4/18/19	80	1.5	y	55	walk	0	180	120	1	225	60	22:30	360
4/19/19	74	2	n	85	walk	35	120	80	2	95	25	21:30	427
4/20/19	72	1	y	38	walk	13	185	135	2	185	10	22:45	460
4/21/19	63	1.25	n	61	bike	0	195	165	2	175	15	21:20	300
4/22/19	68	1.5	y	52	walk	0	360	45	0	133	10	22:20	345
4/23/19	61	0.5	y	56	bike	14	360	47	2	93	15	21:45	255
4/24/19	70	0.5	y	72	walk	36	165	75	0	225	20	22:30	195
4/25/19	72	0	y	50	walk	25	195	95	2	70	17	22:45	428
4/26/19	88	0	n	69	bike	20	255	63	0	175	27	22:22	431
4/27/19	85	0	n	65	walk	0	193	0	3	55	31	22:23	435
4/28/19	62	0.5	n	48	walk	0	175	0	3	96	28	21:38	285
4/29/19	84	1	y	59	bike	14	305	65	1	105	10	22:02	330
4/30/19	61	0	n	52	walk	24	247	75	0	235	30	22:39	325
5/1/19	68	1.5	y	94	walk	0	215	65	1	185	5	22:10	345
5/2/19	62	0.5	y	66	bike	13	255	0	2	115	35	22:01	449
5/3/19	63	1	n	48	walk	0	175	185	1	148	20	21:45	435
5/4/19	68	2	n	99	walk	12	145	125	2	107	44	23:15	229
5/5/19	54	0	n	47	walk	7	196	92	1	47	26	21:37	437
5/6/19	72	0.5	y	68	walk	0	315	45	0	116	12	22:42	195
5/7/19	69	1.5	y	67	bike	27	275	67	2	55	25	22:07	468
5/8/19	87	2	n	61	walk	9	236	32	0	252	24	23:15	358
5/9/19	64	1.25	y	52	walk	0	295	55	1	67	27	21:34	335
5/10/19	77	1.5	n	47	walk	0	255	63	1	285	4	22:52	370
5/11/19	58	1	n	51	walk	17	325	76	2	22	19	23:00	455
5/12/19	52	1.5	n	74	bike	0	165	95	0	75	16	21:40	335
5/13/19	73	1	y	61	walk	9	286	96	0	57	22	21:47	456
5/14/19	48	1	n	47	walk	14	315	40	2	25	13	21:54	423

Figure 2. Sleep Study Data Collection

After the measurements were completed, I began the **A**nalyze phase. In this phase, I discerned whether I had been able to get a clear picture of my sleep. I focused on whether I had attained meaningful insight into the inputs required to attain my goal of 7 hours of sleep. Using multivariate linear regression, I generated an Adjusted R-Squared of -.249 with low F and P values that pointed to a very weak correlation between my selected inputs and sleep. None of my inputs were driving my insufficient sleep which was an unexpected outcome.

After a bit of disappointment and reflection, I started working on the **I**mprove phase of DMAIC. As none of the inputs were significant, I could not attempt to adjust any of them to measure if things improved. Instead, I focused on how I might improve my study by adding new inputs or possibly focusing on only one or a few inputs at a time to decrease dispersion. Another issue that I became acutely aware of in the process was the difficulty in reliably measuring my amount of sleep per night. I used a sleep tracking watch that I knew after a few days was not accurate. It would mistakenly count the time where I laid still in bed but was awake as sleep. A possible solution could be to enter a sleep study.

The final phase is **C**ontrol where the process is monitored to make sure the results are reliable. This is often accomplished using control charts. As none of my inputs were found to be of significance, I did not proceed to the Control phase. Instead, it was time to go back to the drawing board. The resulting Story Board is shown below (Figure 3).

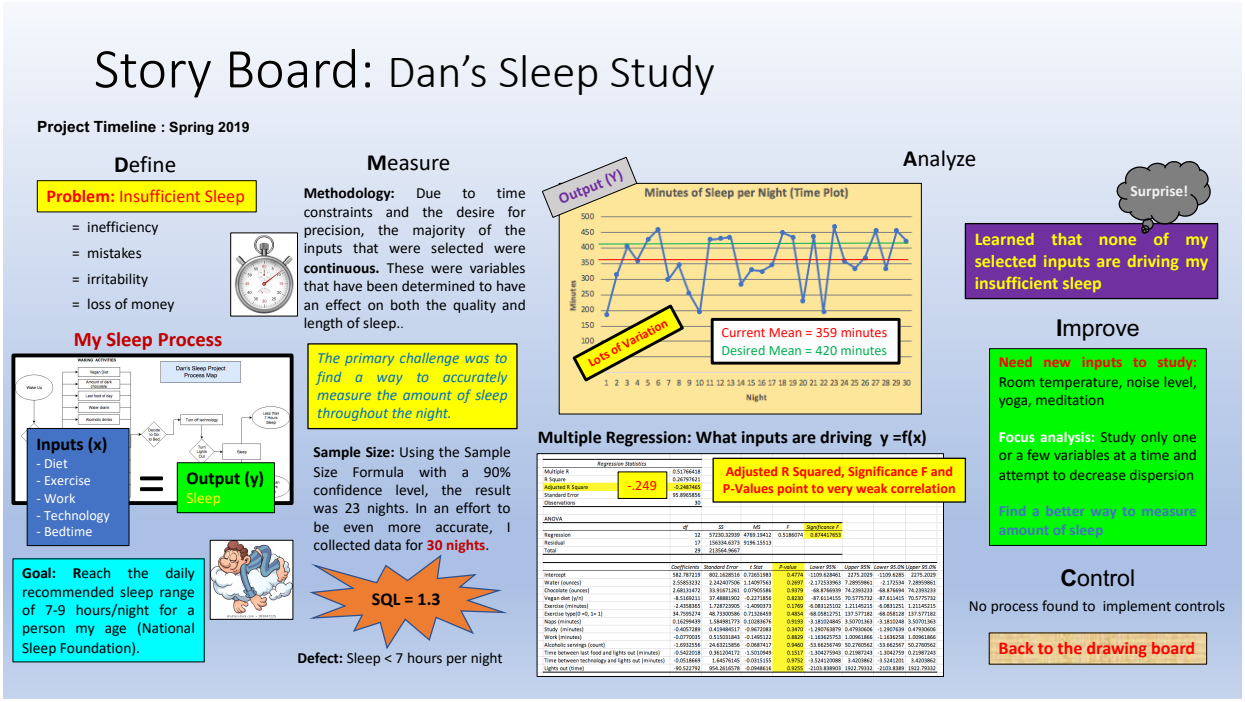


Figure 3. Dan's Sleep Study Story Board

The sleep study project helped define the data science process. I discovered that many projects/studies lead to more new questions than answers. And that frequently, iterations through the process will lead to better intelligence gained. I learned the importance of utilizing high quality data. In my sleep study, I determined that the inputs I thought were important and could measure were not driving my lack of sleep. New inputs and possibly a better measuring system would be a good next step. This study also introduced me to the need to visualize the results in an easy to interpret manner. If others cannot easily interpret the results, the project has little or no value.

Global Literacy Project

The second project I will describe was from the Information Visualization class. The purpose of this class was to learn how to present information in an attractive, easy to comprehend visual. I selected Global Literacy as the topic for my poster project.

This project used a more data science driven approach. The first step was again to design the process. I developed two questions that I would attempt to answer. These were: 1) How do literacy rates compare at the country, global, and regional level? and 2) How is literacy rate correlated with other key developmental metrics? Then, I searched for data that I believed would answer these questions. Data was collected from the CIA World Factbook prior to 2018 and imported into R. In R, the data was cleaned, organized and merged with map data. The resulting data set was then used to create numerous visualizations using ggplot2. This was an iterative approach where different types of graphs and charts were generated in an effort to thoroughly answer each question. These included a choropleth map, boxplot, pie chart, bar plot, tree map and scatterplots with trend lines. The objective was to tell a story of global literacy using the visualizations as the primary sources.

The creation of the poster required multiple techniques. These included aggregations and numerous advanced ggplot2 tools as well as hours of editing in Adobe Illustrator to make the visuals more illuminating and attractive. The layout was also of critical importance. The design I crafted is informative, eye-catching, thematic, and has a clear flow. In addition, I made sure it was not over-crowded and had room for the eyes to rest (Figure 4).

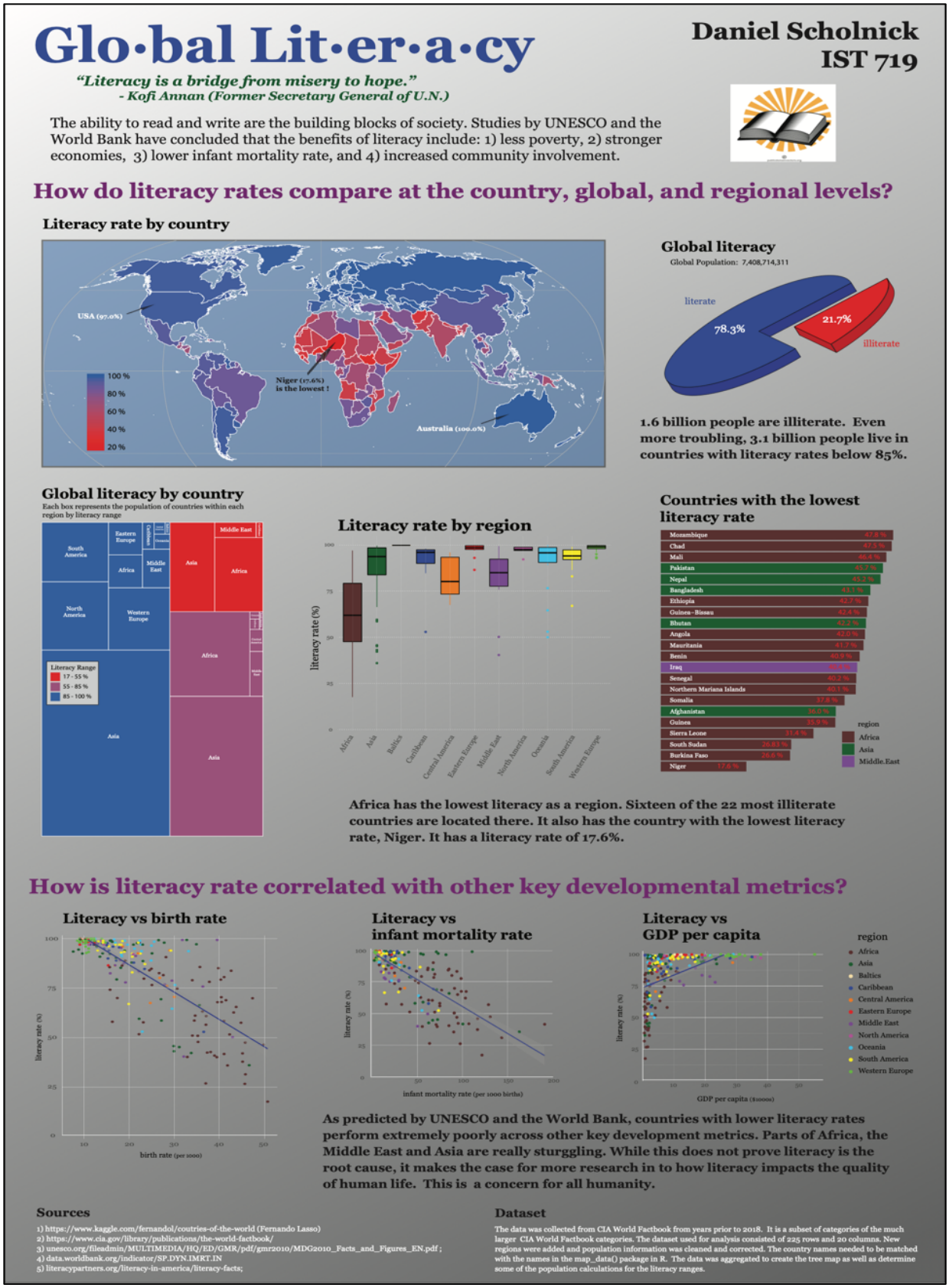


Figure 4. Global Literacy Poster

This project emphasized the power of visualization combined with data science. Visualizations can reveal patterns, trends and connections in data that are difficult to find any other way. And humans process visuals "60,000X faster in the brain than text" (Visual Teaching Alliance). According to Dr. Lynell Burmark, an education consultant: *"…unless our words, concepts, ideas are hooked onto an image, they will go in one ear, sail through the brain, and go out the other ear."* This idea is the essence of data science and was emphasized throughout the program. If you are not able to clearly communicate your data and results, you will not be able to accomplish your goals.

Fudgemart Inc. Data Warehouse, Sales Data Mart and Business Intelligence Project

The next project I will review was from the Data Warehouse class. The project was to build a data warehouse from the merging of two fictitious databases. After the completion of the data warehouse, we were required to create a data mart to produce analytics with PowerBI and Excel that would describe a business process.

This project brought data science to life in a real-world manner. The process started by exploring the data. What products or services did these fictitious companies sell? One represented a streaming movie division and the other an online retailer. Which columns did the databases have in common to share? Customer email addresses were utilized. Were the data types the same in each database? Most of it was very similar. How many customers were there in common and in total? There were fifteen customers in common with a total of forty-six customers.

After getting a basic understanding of the databases, it was time to focus on the data warehouse and data mart. Our team began by building out a High-Level Dimensional Model (Figure 5). What business processes did we want to examine more closely using a data mart? In our project, we focused on Sales as the business process. We determined that the Customers, Products and Date tables would be utilized. We selected one row per product order as our level of granularity. This would enable us to explore the data from an individual product, a single department or at a company-wide level. We would then be able to compare services of the streaming division with the products of the online retail sales division by treating each service package as a unique product.

Business Process Name	Fact Table	Fact Grain Type	Granulairty	Facts	Customer	Product	Movie	Date	Employee
Customer Analysis	CustomerFacts	Periodic Snapshot	One row per customer	Quantity movies ordered, Average movie review, Quantity fm orders, Quantity products ordered, Total Sales, Average product review, Customer Fudgemart, Customer Fudgeflix, Customer Both	X	X	X	X	
FF Inventory Analysis	FFInventoryAnalysisFacts	Periodic Snapshot	One row per genre type	Quantity of titles, Percentage of titles, Number Orders, Average review	X		X	X	
FF Order Analysis	FFOrderFacts	Accumulating Snapshot	One row per order	Days to Ship, Days to Return	X		X	X	
FM Product Analsyis	FMProductFacts	Periodic Snapshot	One row per product	Quantity sold, Revenue, Profit, Average review	X	X			
FM Employee	FMEmployeeFacts	Transaction	per employee (employee + date)	Hours worked, Total pay				X	X
Sales	Sales	Periodic Snapshot	one row per product order	Order_Quanity, Product_Revenue, Product_Cost, Product_Margin	X	X		X	

Figure 5. High-Level Dimensional Model

With all of this defined, we produced a Detailed Dimensional Model that painstakingly described the data warehouse we were going to build (Figure 6). It included the column names, descriptions, ETL rules, data types, keys, and source information as well as many more details.

This was then used to generate the SQL code we would use for the design of the data warehouse and data mart.

Table Name	FactSales
Table Type	Fact
Display Name	Sales
Database Schema	fudge
Table Description	Sales based on product by order
Comment	
Biz Filter Logic	
Size	
Generate Script?	Y

Home Page

					Target				Source	
Column Name	Display Name	Description	ETL Rules	Comments	Datatype	Size	Key?	FK To	Source System	Source Table
customer_key	customer_key	Key to Customer	Key lookup from Customer.customer_key		int		FK	Customer.customer_key		
product_key	product_key	Key to Product	Key lookup from Product.product_key		int		FK, PK	Product.product_key		
order_id	order_id	Order ID to delineate unique orders	Add fm_ or ff_ to the front of the Order ID to		varchar	50	PK		fudgeflix_v3, fudgemart_v3	ff_accounts_billing, fm_order_details
order_date_key	order_date_key	Key to Date	Key lookup from DimDate.DateKey		int		FK	Date.DateKey	fudgeflix_v3, fudgemart_v3	ff_account_billing, fm_orders
order_qty	order_qty	Amount of product ordered			varchar	50	PK		derived	
product_revenue	product_revenue	Revenue of product sold			money				derived	
product_cost	product_cost	Cost of product sold			money				derived	
product_margin	product_margin	Margin of product sold			money				derived	
RowIsCurrent	Row Is Current	Is this the current row for this member	Standard SCD Type 2 Metadata		nchar	1			Derived	
RowStartDate	Row Start Date	When did this row become valid for this	Standard SCD Type 2 Metadata		datetime				Derived	
RowEndDate	Row End Date	When did this row become invalid?	Standard SCD Type 2 Metadata		datetime				Derived	
RowChangeReason	Row Change Reason	Why did the row change last?	Standard SCD Type 2 Metadata		nvarchar	200			Derived	

Figure 6. Detailed Dimensional Model

After generating the shell of the warehouse, the focus moved to staging and populating the warehouse. This encompassed a great deal of time sorting out the details of project. The data was extracted from the databases, transformed into the necessary format and then loaded into the data warehouse. Numerous decisions were made in this process to determine how to merge, sort and derive the data. The schema of the Data Warehouse can be seen in Figure 7.

A data mart or multidimensional cube was also built using Visual Studio integration services for analyzing the Business Intelligence (Figure 8). The aim was to understand the historical sales analytics and trends for Fudgemart Inc. The cube created a better option for analyzing the Fudgemart Inc. sales data than directly accessing the data warehouse. This reduced load is of great benefit to companies that are constantly updating, querying, and otherwise putting a heavy load on the warehouse. In e-commerce, as an example, this could mean faster transactions.

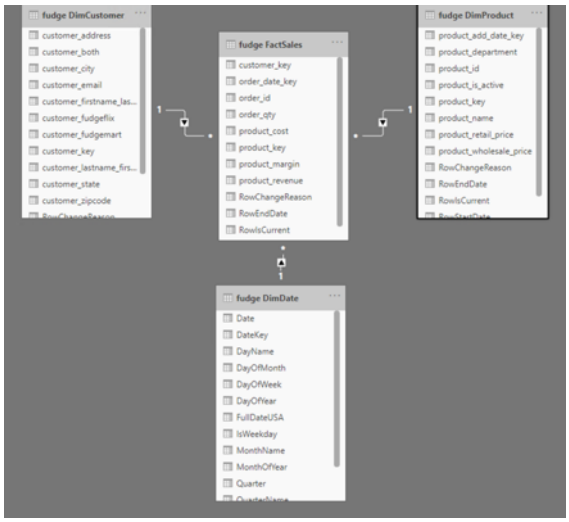


Figure 7. Data Warehouse Schema (ROLAP)

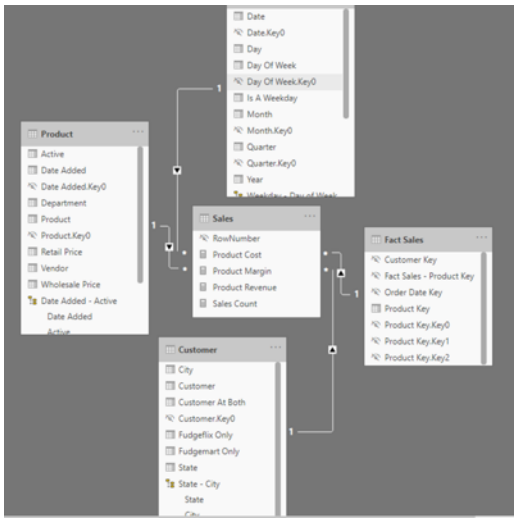


Figure 8. MOLAP Schema (Multidimensional Cube)

This project emphasized creating something of value to the stakeholders throughout the build. Every time a decision was made on the design, we asked if it was related to promoting the value for the stakeholder. The following business intelligence visuals demonstrate some of the valuable insights that can be drawn from our project (Figure 9). The design enables the user to analyze the data from the view of the order, product, department, and/or company level. This can be done based on a date or a location of the sale. Essentially, we created the desired drill down effect that enables users to get to the desired depth of information to make quality decisions.

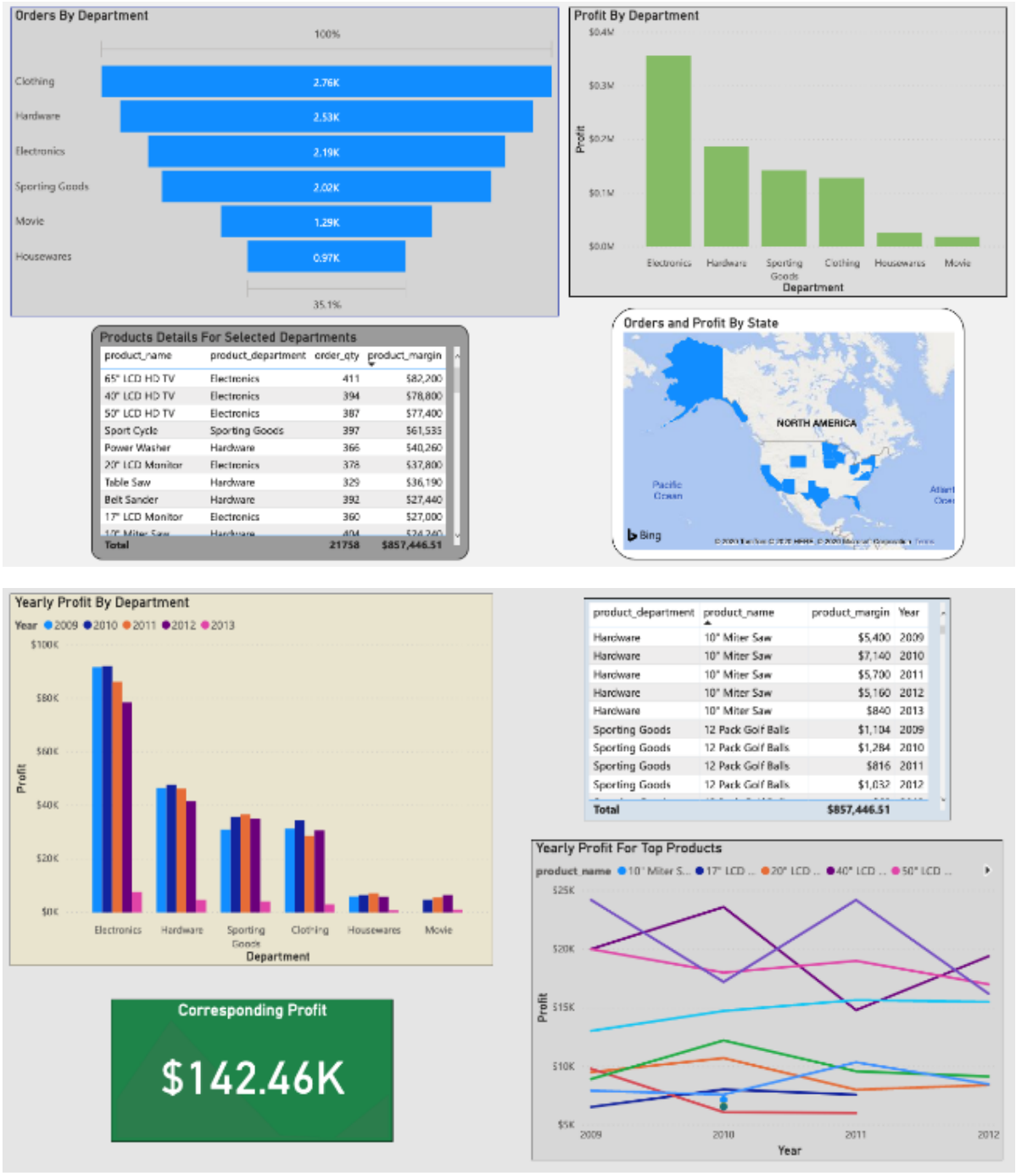


Figure 9. Fudegmart Inc. Sales BI

The Fudgemart Inc. project felt like the culmination of the program in many ways. The only part that was missing was the gathering of the data. The idea that a company with many divisions would need to combine their databases to create a centralized user interface to thoroughly understand the company operations is plausible.

Data Mining, Text Mining and Predictive Analytics

In the Applied Data Science program, I took a number of classes focused on predictive analytics and machine learning including Data Mining and Text Mining. It quickly became apparent that predictive analytics and machine learning are considered the sexy part of data science. In this section, I am going to review some of the highlights of my learning in this area.

In all of these classes, the critical role of the data science process became clear. When a scientist does not create a well-defined objective, it is challenging to deliver value for the stakeholders. The results are often extraneous insights. In my team project for Text Mining, this axiom was tested. The initial goal was to classify people by Star Wars characters as defined by their Meyers Briggs personality types using the dialogue from the movies. The first step was to determine the personality type that matched with each character from the saga by using collaborative online classifications. Some were done by psychologists and others just by fans. Through exploration, we found that there were widely divergent views and that this task would not be tenable without input from many more experts. This experience taught me the value of domain knowledge and staying focused on the objective. If the design of the project is not actionable, it is necessary to stop the process and determine how to move forward.

The other element that became clear is that data science is all about the data, particularly the quality of the data. Without quality data, the "garbage in, garbage out" adage will apply. This is why the scientist needs to determine the type of data needed to accomplish their objective and then explore the available data thoroughly. The team can then make sure it is of high enough quality to produce valuable results and is feasible to collect. If the data is not sound, a search for more or better data is necessary.

The exploration of the data is essential to success. As previously mentioned, in the initial stages of the Star Wars project, we were having difficulty grouping the characters into Myers Briggs types. They were not aligning in a pattern we could recognize. We utilized Topic Modeling, an unsupervised machine learning technique, to gain a clearer picture of the relationships between characters. As you can see in the strong results of the inter topic distance model, the topics, represented by the circles, are all distinct, good size and spread throughout the quadrants (figure 10). The model placed the dark side in its own quadrant, Luke who is deeply conflicted between good and evil is off by himself and the least conflicted Allegiance group is separate from the others.

Using these results, we found that many of the characters did align, just not in the ways we had expected. In addition, we realized the Star Wars saga simply did not have a large enough pool of well-defined characters to represent the 16 personality types of the Meyers Briggs model. The utilization and deeper exploration of the data using topic modeling proved invaluable in helping us understand the data.

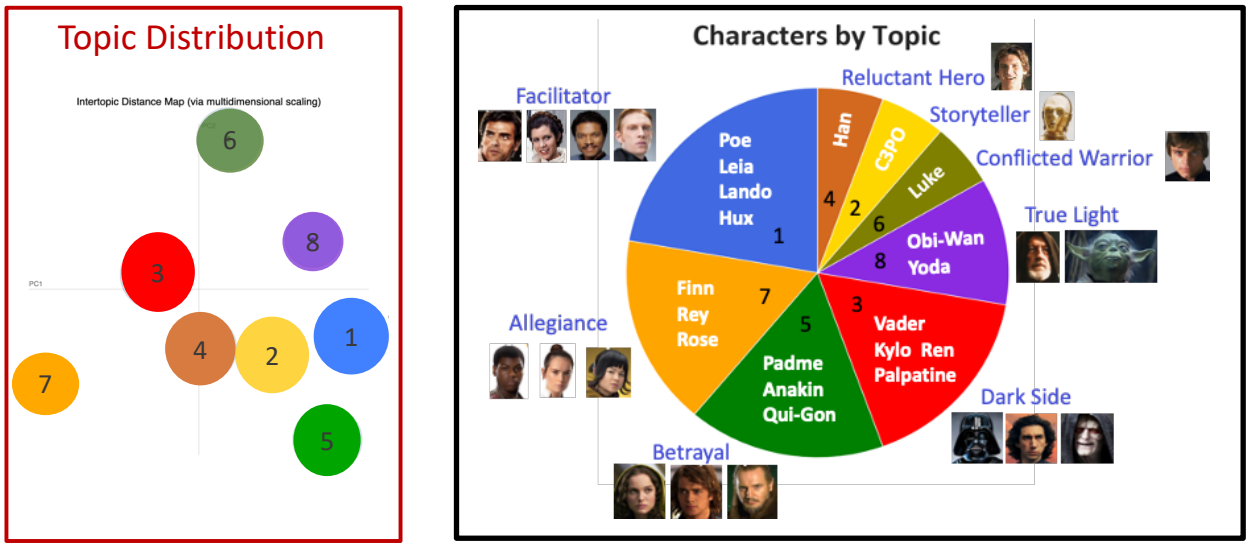


Figure 10. Intertopic distance map and character pie chart based on topic modeling.

Once the data has been collected and explored, it needs to be prepared for further analysis. Some of the critical questions I have encountered through work in these classes are: Does the data need to be put into a document term matrix and/or normalized? Are some of the statistics collinear or insignificant? Is the variance of some of the data too limited and therefore of little value? Is all of the data temporally the same? What percentage of data is best for training and testing? Are the splits representative of the data? Is cross validation required? If so, what type?

After the data has been prepared, it is time to use the machine learning algorithms to classify the data or make predictions and prescriptions. The type of machine learning method needs to be determined. The type of data or the size of the data set can help clarify the best choice. Experience with previous data sets is also an incredibly helpful guide.

After running the machine learning models, the interpretation of the results, the tuning the of the model parameters and validation become the focus. It is imperative to keep an open mind and not predetermine the results. Make sure the objective is achieved. Think about whether the results are accurate, insightful, surprising and ponder adjustments to improve the results.

In my team's project for Data Mining, we set out to predict fantasy football outcomes and identify their associated drivers. After the preparation of our dataset, associative rules mining and Boruta were utilized to determine critical features. Then, we generated models using regression, Naïve Bayes, decision tree, support vector machines, kNN and random forest. The model results varied between 75% and 89% accuracy. The decision tree model performed the best (89%), Naïve Bayes was the least accurate (75%) and the other models in the range of 82-84% (Figure 11, 12). The clearest theme was that Market Share of Offense, which refers to how many touches a player has in a game, has the most impact on fantasy football scoring. This is demonstrated in the variables of importance plots generated from the random forest and SVM packages. The SVM plot is displayed in figure 13.

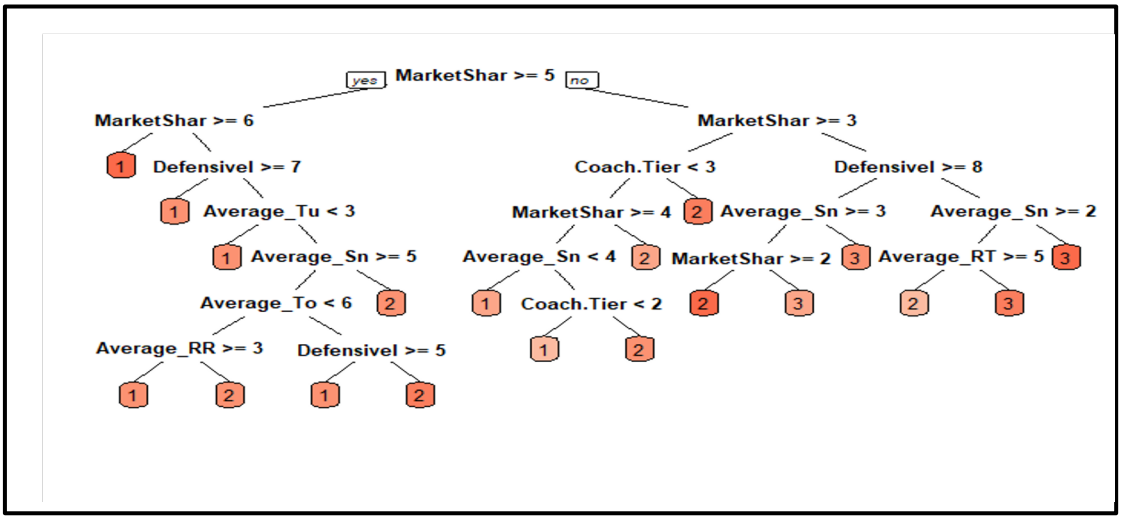
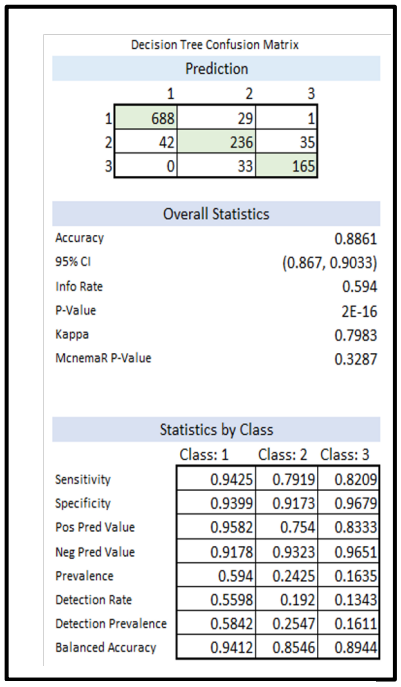


Figure 11. Decision Tree for Fantasy Football Project



Decision Tree Confusion Matrix

Prediction

	1	2	3
1	688	29	1
2	42	236	35
3	0	33	165

Overall Statistics

Accuracy	0.8861
95% CI	(0.867, 0.9033)
Info Rate	0.594
P-Value	2E-16
Kappa	0.7983
McnemaR P-Value	0.3287

Statistics by Class

	Class: 1	Class: 2	Class: 3
Sensitivity	0.9425	0.7919	0.8209
Specificity	0.9399	0.9173	0.9679
Pos Pred Value	0.9582	0.754	0.8333
Neg Pred Value	0.9178	0.9323	0.9651
Prevalence	0.594	0.2425	0.1635
Detection Rate	0.5598	0.192	0.1343
Detection Prevalence	0.5842	0.2547	0.1611
Balanced Accuracy	0.9412	0.8546	0.8944

Figure 12. Decision Tree Confusion Matrix

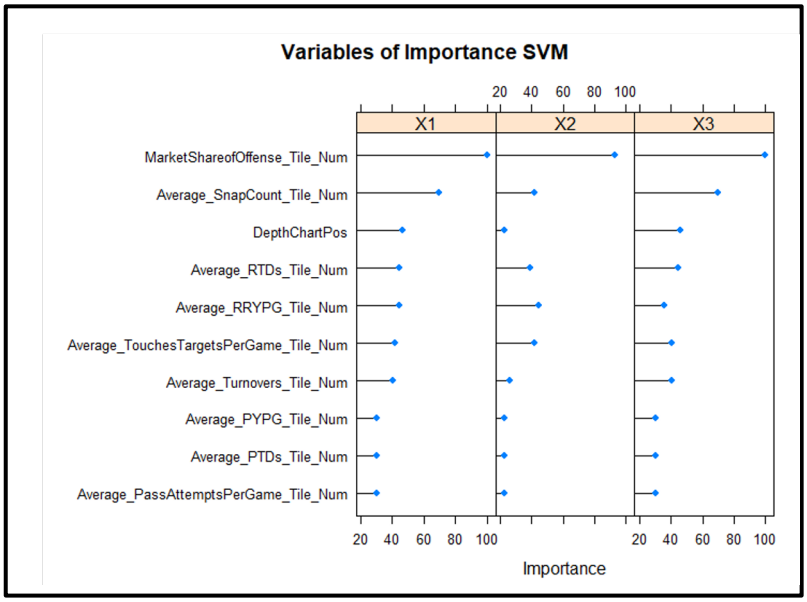


Figure 13. SVM Variables of Importance

And the final component is communication. This starts at the beginning and continues beyond the completion of a project. Fellow scientists need to discuss the process and results. These ideas should then be synthesized and shared with the stakeholders. It's critical to document feedback from all the stakeholders for use in future projects. Possible next steps should be discussed as well. The group projects in the program made value of communication exceptionally clear. If the prospective data scientists are not able to work through the process and communicate their ideas and concerns with others, time is wasted, and the results will be diminished.

Ethical Considerations

Data science is intrinsically reliant on the collection of data. The more quality data, the better the potential results of the science. In many cases, data science has no ethical constraints such as weather prediction. In others, the studies directly affect humans or animals. These studies can lead to ethical gray areas. How should a data scientist determine what is acceptable to collect and how to use it? Are facial recognition and movement tracking ethical without consent? How about research that requires animal testing? I tend to favor the idea that the balance should lean towards the greater good of society over the individual's concerns. However, this must be tempered by attempts to get the individual's consent and the laws regarding privacy. In order to achieve this balance, I believe it is imperative that data scientists stay current on federal and state laws. If a scientist disregards ethical concerns and chooses to blindly follow the policy of an employer, they do so to their own detriment and will potentially harm their reputation, people and companies and the field of data science. I believe it would be beneficial to allocate more time in the program to the ethics of data science.

Conclusion

The Applied Data Science program introduced me to numerous data science techniques. Some of the most memorable were data exploration, data cleaning, data quality, regression, machine learning, data visualizations and solver for supply chain management. Each of these emphasized the data science process.

What I learned by focusing on the data science process was how to think like a data scientist. Quality science needs to be well thought out, defined, repeatable and include extensive visualizations. Without the data science process, it is next to impossible to ensure that the information attained will be valuable.

While this program was a sound start, I am far from being an expert. I intend to improve these skills with more study and exploration. And finally, as is often the case, the best learning will come from the experience of doing data science in a real-world environment.

Other Documents

Supporting documents and resume can be accessed at:

https://github.com/Scholbandit/ChefD/tree/master/Scholnick_Portfolio_Project