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 Jeffrey 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 three of the projects I completed in the Applied Data Science program.

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.

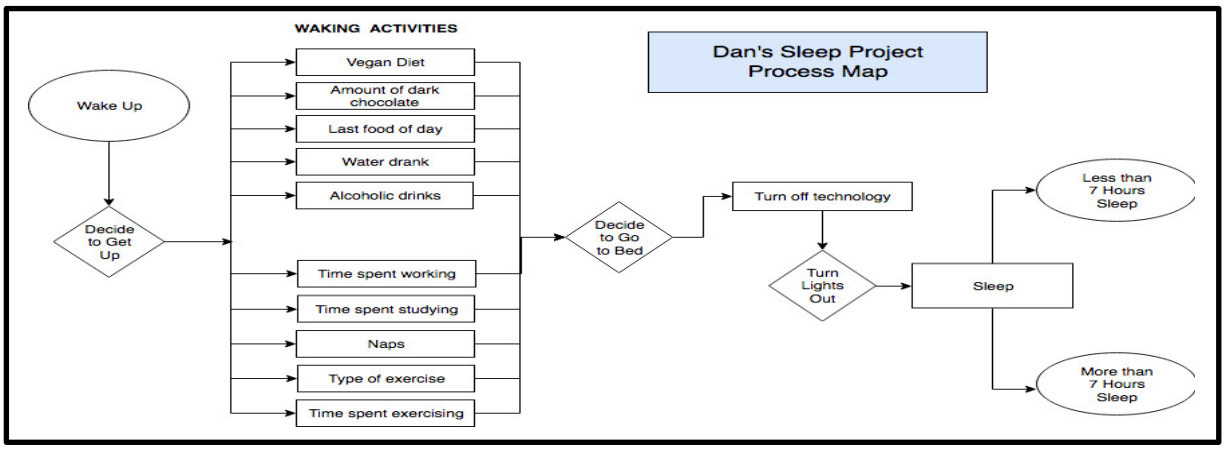


Figure 1. Process Map

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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	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)
2	4/15/19	68	2	n	50	walk	25	330	110	1	135	15	22:45	185
3	4/16/19	62	1	n	60	bike	30	315	25	0	125	35	21:10	315
4	4/17/19	72	1	n	75	bike	15	305	45	1	145	25	22:00	405
5	4/18/19	80	1.5	y	55	walk	0	180	120	1	225	60	22:30	360
6	4/19/19	74	2	n	85	walk	35	120	80	2	95	25	21:30	427
7	4/20/19	72	1	y	38	walk	13	185	135	2	185	10	22:45	460
8	4/21/19	63	1.25	n	61	bike	0	195	165	2	175	15	21:20	300
9	4/22/19	68	1.5	y	52	walk	0	360	45	0	133	10	22:20	345
10	4/23/19	61	0.5	y	56	bike	14	360	47	2	93	15	21:45	255
11	4/24/19	70	0.5	y	72	walk	36	165	75	0	225	20	22:30	195
12	4/25/19	72	0	y	50	walk	25	195	95	2	70	17	22:45	428
13	4/26/19	88	0	n	69	bike	20	255	63	0	175	27	22:22	431
14	4/27/19	85	0	n	65	walk	0	193	0	3	55	31	22:23	435
15	4/28/19	62	0.5	n	48	walk	0	175	0	3	96	28	21:38	285
16	4/29/19	84	1	y	59	bike	14	305	65	1	105	10	22:02	330
17	4/30/19	61	0	n	52	walk	24	247	75	0	235	30	22:39	325
18	5/1/19	68	1.5	y	94	walk	0	215	65	1	185	5	22:10	345
19	5/2/19	62	0.5	y	66	bike	13	255	0	2	115	35	22:01	449
20	5/3/19	63	1	n	48	walk	0	175	185	1	148	20	21:45	435
21	5/4/19	68	2	n	99	walk	12	145	125	2	107	44	23:15	229
22	5/5/19	54	0	n	47	walk	7	196	92	1	47	26	21:37	437
23	5/6/19	72	0.5	y	68	walk	0	315	45	0	116	12	22:42	195
24	5/7/19	69	1.5	y	67	bike	27	275	67	2	55	25	22:07	468
25	5/8/19	87	2	n	61	walk	9	236	32	0	252	24	23:15	358
26	5/9/19	64	1.25	y	52	walk	0	295	55	1	67	27	21:34	335
27	5/10/19	77	1.5	n	47	walk	0	255	63	1	285	4	22:52	370
28	5/11/19	58	1	n	51	walk	17	325	76	2	22	19	23:00	455
29	5/12/19	52	1.5	n	74	bike	0	165	95	0	75	16	21:40	335
30	5/13/19	73	1	y	61	walk	9	286	96	0	57	22	21:47	456
31	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).

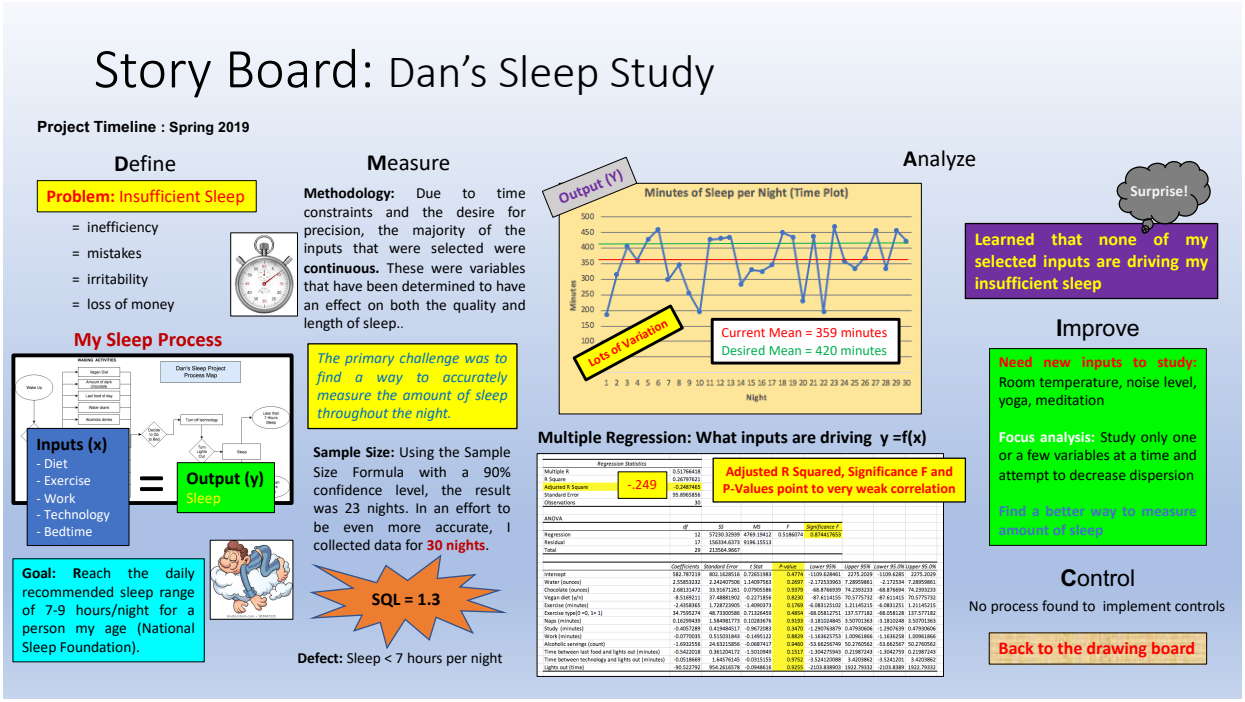


Figure 3. Dan's Sleep Study Story Board

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 Information Visualization class. The purpose of this class was to learn how to present information in an attractive, easy to comprehend display. I selected Global Literacy as the topic on which I would create my poster project.

This project moved towards 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. 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).

Glo·bal Lit·er·a·cy

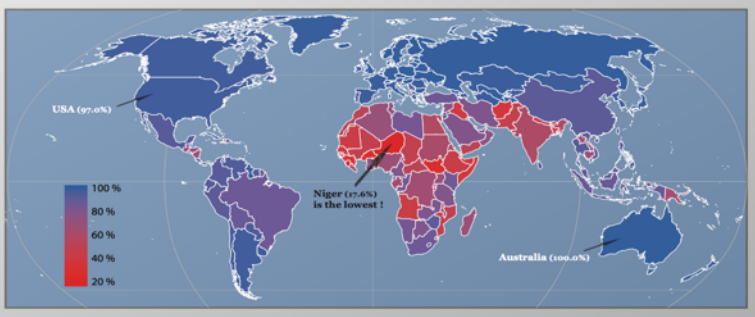
Daniel Scholnick
IST 719

"Literacy is a bridge from misery to hope."
- Kofi Annan (Former Secretary General of U.N.)

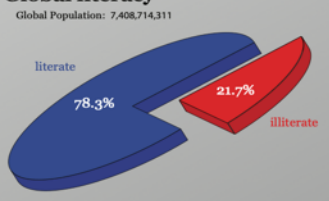
The ability to read and write are the building blocks of society. Studies by UNESCO and the World Bank have concluded that the benefits of literacy include: 1) less poverty, 2) stronger economies, 3) lower infant mortality rate, and 4) increased community involvement.

How do literacy rates compare at the country, global, and regional levels?

Literacy rate by country



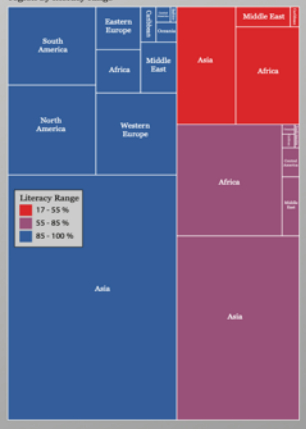
Global literacy



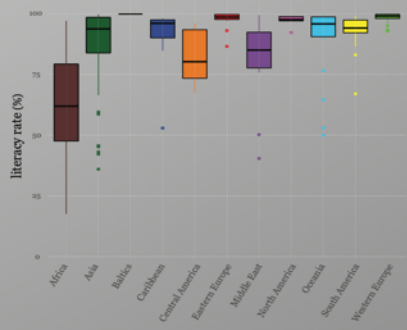
1.6 billion people are illiterate. Even more troubling, 3.1 billion people live in countries with literacy rates below 85%.

Global literacy by country

Each box represents the population of countries within each region by literacy range



Literacy rate by region

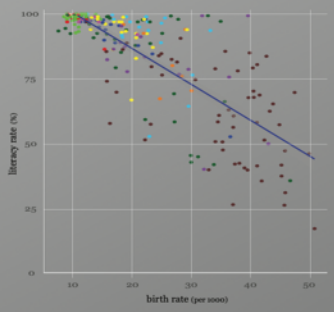


Countries with the lowest literacy rate

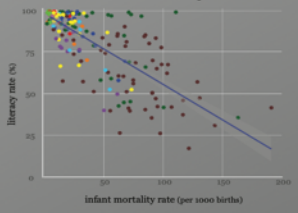
Africa has the lowest literacy as a region. Sixteen of the 22 most illiterate countries are located there. It also has the country with the lowest literacy rate, Niger. It has a literacy rate of 17.6%.

How is literacy rate correlated with other key developmental metrics?

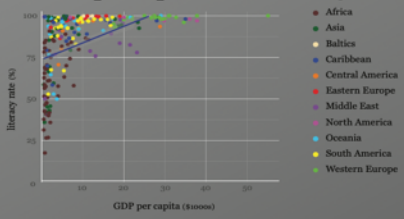
Literacy vs birth rate



Literacy vs infant mortality rate



Literacy vs GDP per capita



As predicted by UNESCO and the World Bank, countries with lower literacy rates perform extremely poorly across other key development metrics. Parts of Africa, the Middle East and Asia are really sturggling. While this does not prove literacy is the root cause, it makes the case for more research in to how literacy impacts the quality of human life. This is a concern for all humanity.

Sources

1) https://www.kaggle.com/fernandol/coutries-of-the-world (Fernando Lasso)
2) https://www.cia.gov/library/publications/the-world-factbook/
3) unesco.org/fileadmin/MULTIMEDIA/HQ/ED/GMR/pdf/gmr2010/MDG2010_Facts_and_Figures_EN.pdf ;
4) data.worldbank.org/indicator/SP.DYN.IMRT.IN
5) literacypartners.org/literacy-in-america/literacy-facts;

Dataset

The data was collected from CIA World Factbook from years prior to 2018. It is a subset of categories of the much larger CIA World Factbook categories. The dataset used for analysis consisted of 225 rows and 20 columns. New regions were added and population information was cleaned and corrected. The country names needed to be matched with the names in the map_data() package in R. The data was aggregated to create the tree map as well as determine some of the population calculations for the literacy ranges.

Figure 4. Global Literacy Poster

This project emphasized the power of visualization when 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 last project I will review was from Data Warehouse class. The project was to build a data warehouse from the merging of two fictitious databases. The next step, using this data warehouse, was to generate a data mart and produce analytics with PowerBI and Excel to 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 the 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. Which database tables would be required to build and populate the data mart? Customers, Products and Date tables were utilized. Which level of granularity would our process represent? We selected one row per product order. This would enable us to explore the data from an individual product, a single department or at a company-wide level. How could we compare the services of the streaming division with the products of the online retail sales division? We accomplished this by treating each service package as a unique product.

Business Process Name	Fact Table	Fact Grain Type	Granulairty	Facts	Custome	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.

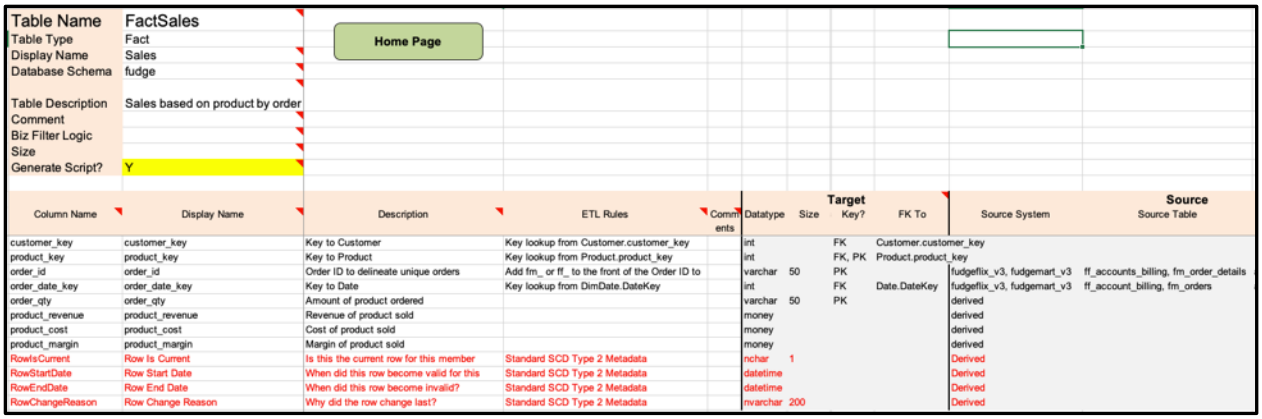


Table Name	FactSales
Table Type	Fact
Display Name	Sales
Database Schema	fudge
Table Description	Sales based on product by order
Comment	
Biz Filter Logic	
Size	
Generate Script?	Y

Home Page

Column Name	Display Name	Description	ETL Rules	Comments	Target 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 better understand the historical sales analytics and trends for Fudgemart Inc. The cube created another option for analyzing the Fudgemart Inc. sales data that would be fetched from a source outside of 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.

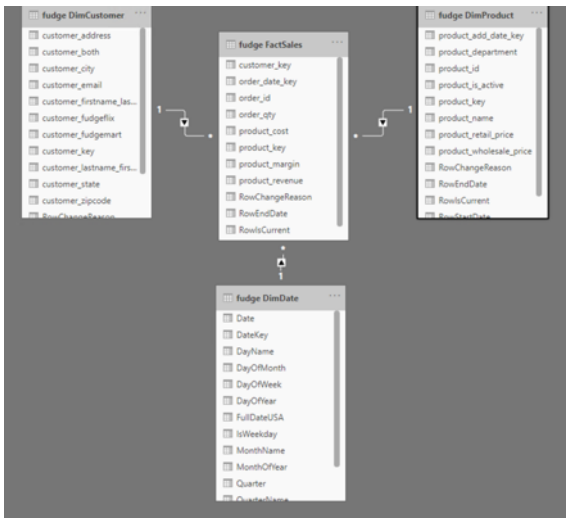


Figure 7. Data Warehouse Schema (ROLAP)

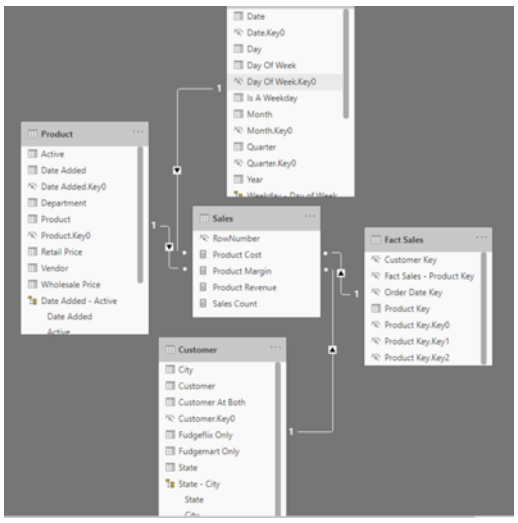
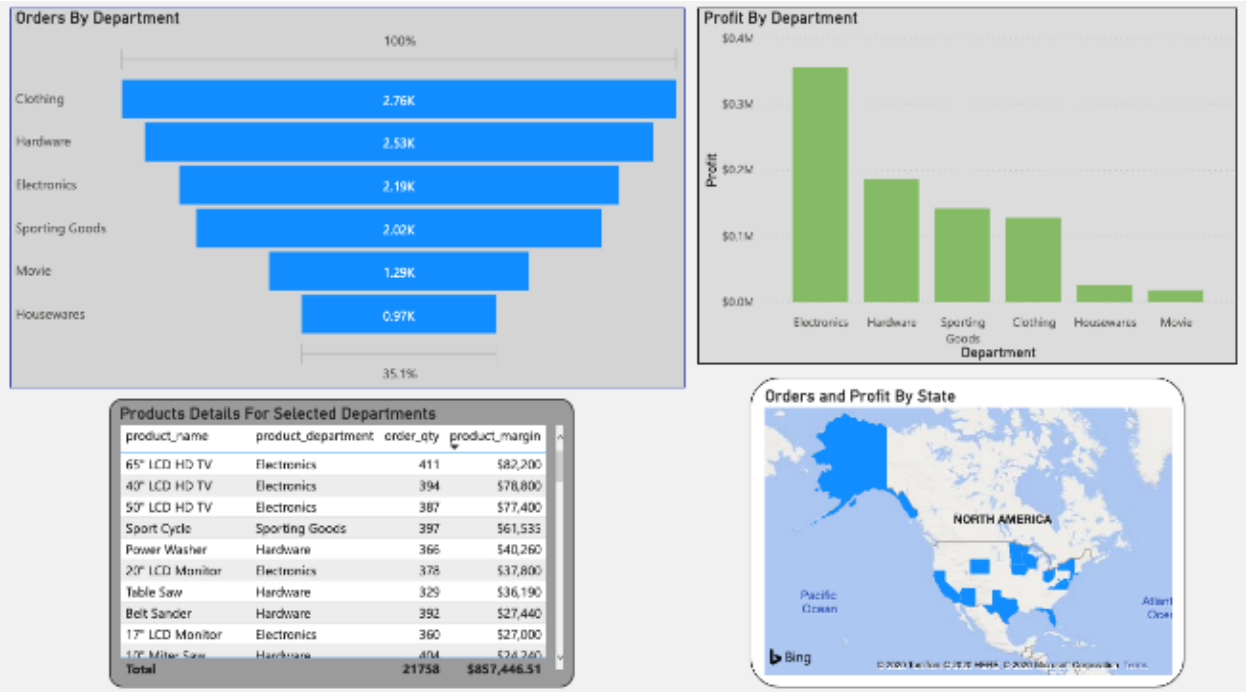


Figure 8. MOLAP Schema (Multidimensional Cube)

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.



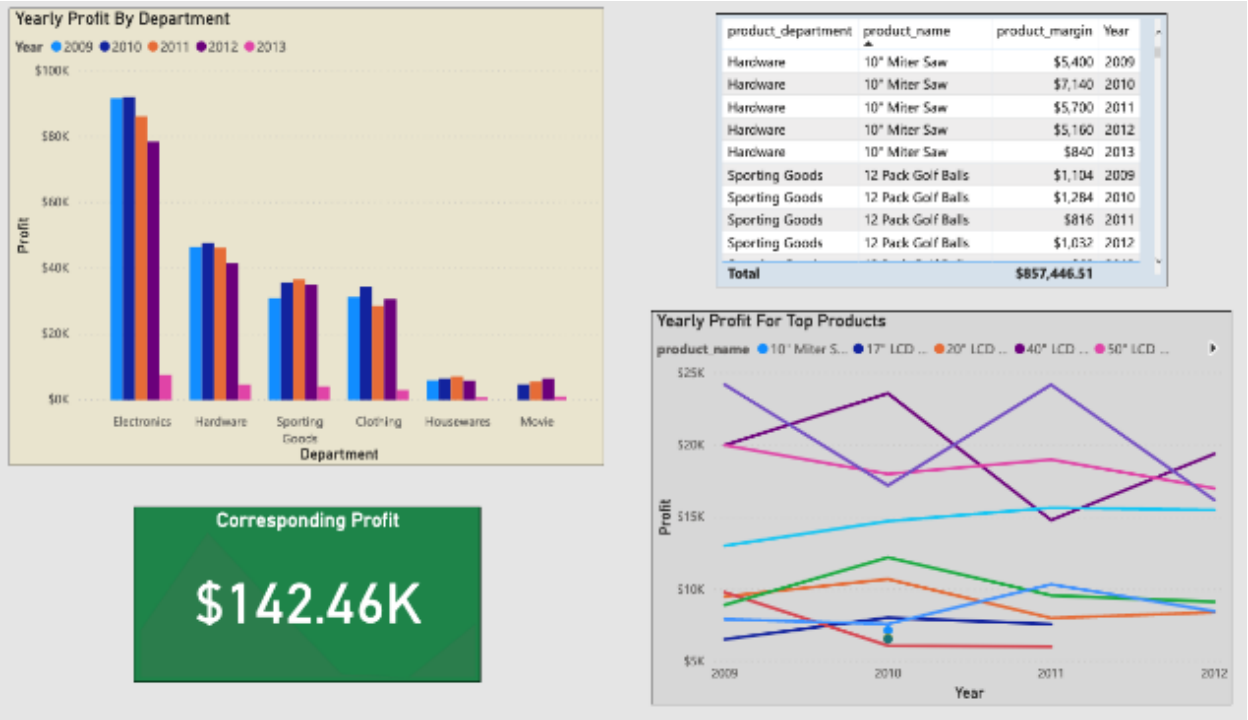


Figure 9. Fudegmart Inc. Sales BI

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 a person to combine their databases to help create a centralized user interface to gain a better understanding of the company is plausible.

Other Thoughts

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 with numerous algorithms, big data, data visualizations, information security and solver for supply chain management. While this was a valuable start, I am far from being an expert. I intend to improve these skills with more study, such as AWS, as well as more hands-on practice in a real-world environment.

Conclusion

My primary takeaway from the program is learning the data science process and how to think like a data scientist. Quality science needs to be well thought out, defined, repeatable and include extensive visuals. Without the data science process, it is next to impossible to ensure that the information attained will be valuable.

Start by creating a deep understanding of the process/business that you are going to study. Create a well-defined objective that creates value for the stakeholders and is not just an extraneous insight. Make sure that everyone including the stakeholders are involved in the process and keep the focus on the objective.

Determine the type of data needed to accomplish your objective. Explore the available data thoroughly. Make sure it is of high quality to produce valuable results. Search for more or better data if necessary. Finally, prepare the data for analytical use.

Then, process your data using tools such as machine learning, linear regression and solver. When statistically significant predictors are found, produce models to predict or prescribe. It is imperative to keep an open mind and not predetermine the results. Throughout the process, ensure that you stay focused on the objective and are using an iterative process.

And most significant of all, make communication central to the process. This starts at the beginning and should continue beyond the completion of the project. Listening to critiques will only make your work better.