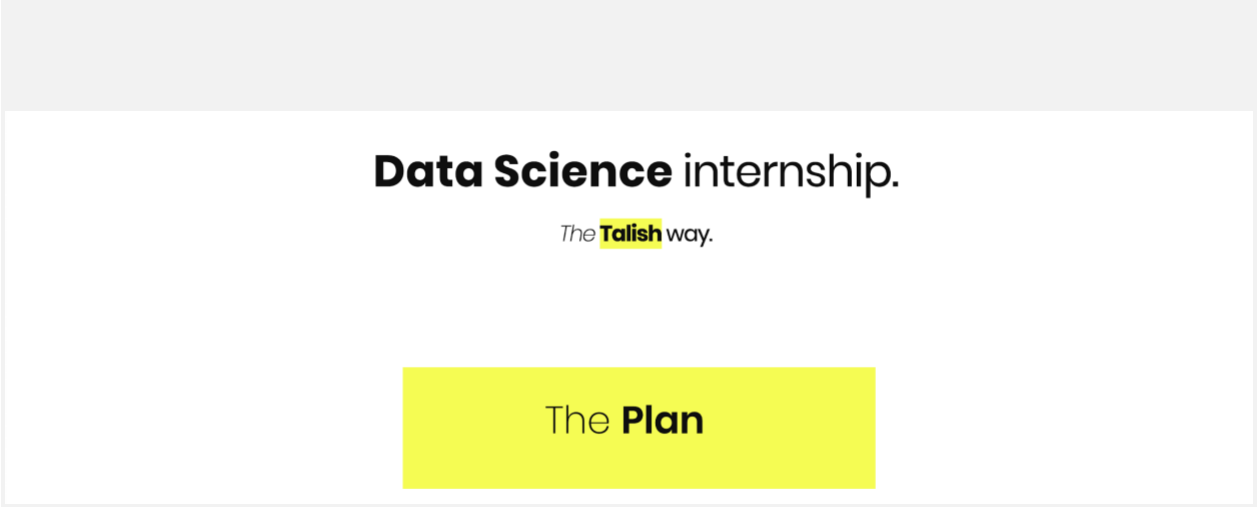
How I crafted my own Data Science Internship!

 Talish Barmare



At the beginning of 2019, I was struggling to find an internship for the summer! After some serious self-reflection, I committed to learning everything I could about machine learning, data science, and the tech industry. Thus, I decided to construct my very own summer internship to help me fill the gaps in my knowledge and learn more! I built my own curriculum, a hodgepodge of hundreds of websites and forums, and committed between 4–6 hours each day.

When I first started, there was always this one question- **"What's the best path to becoming a data scientist? " And I realized the number of resources available is ginormous! Thus, I had to develop a structure and choose a combination of platforms to learn as much as possible in the most concise yet affordable way.**

Before I began my learning journey, I spent a week to structure my thinking process as a Data Scientist. Being a marketer always put me in an innovative/creative mindset so this activity was fairly easy. Yet, it was important to lay in a solid foundation for what came ahead.

Structured Thinking — 2 weeks

- These blog posts were the best material to get me started
- How to train your mind for analytical thinking?
- Tools for improving structured thinking
- The art of structured thinking and analyzing

I then solved a couple of business case studies to help just solidify these thoughts.

Northwestern Kellog Casestudy: Analyse the IMC of Hunger Games Catching Fire

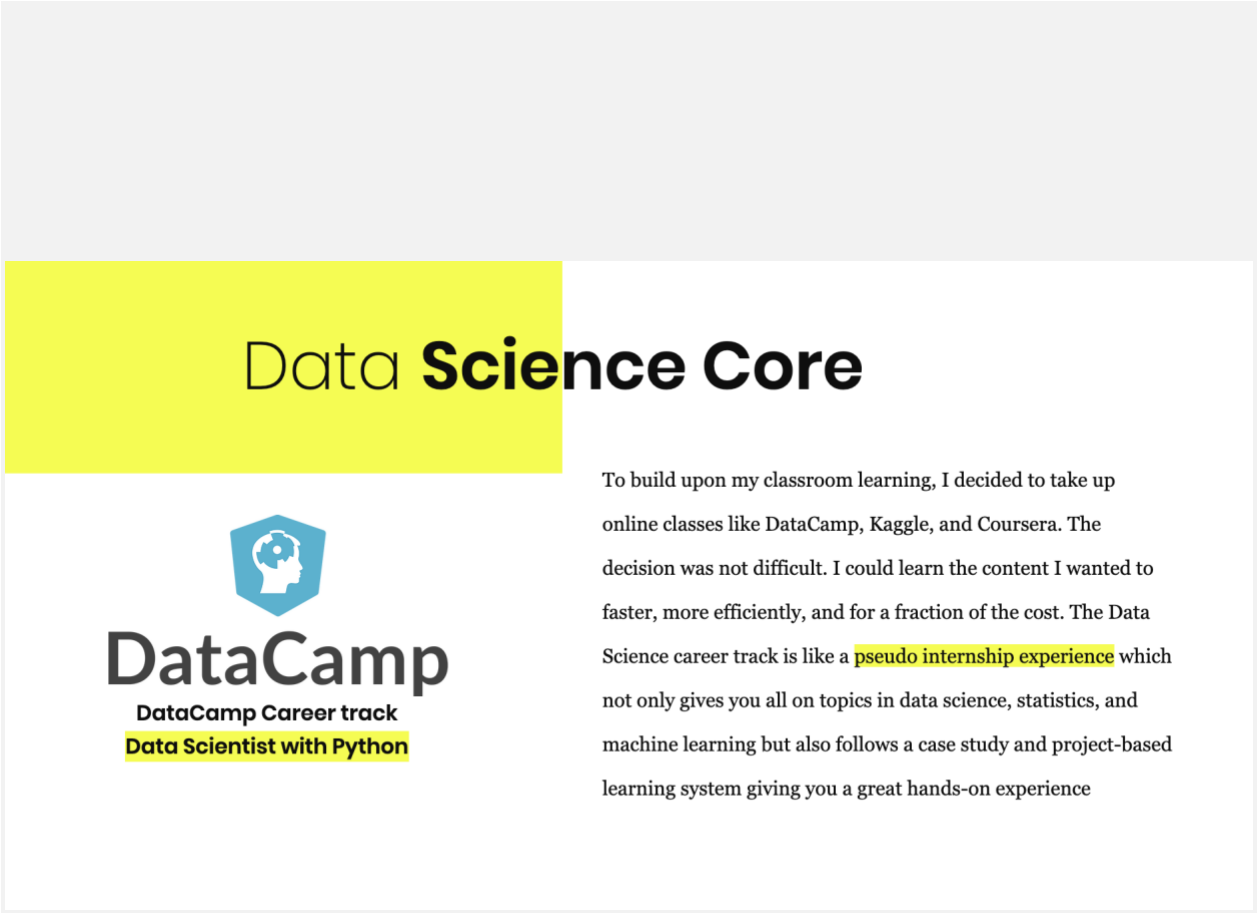
Kirin Case, Conjoint Analysis (using Enginius, a marketing analytics tool)

Moving forward, I divided my learning into cores:

Data Science Core:

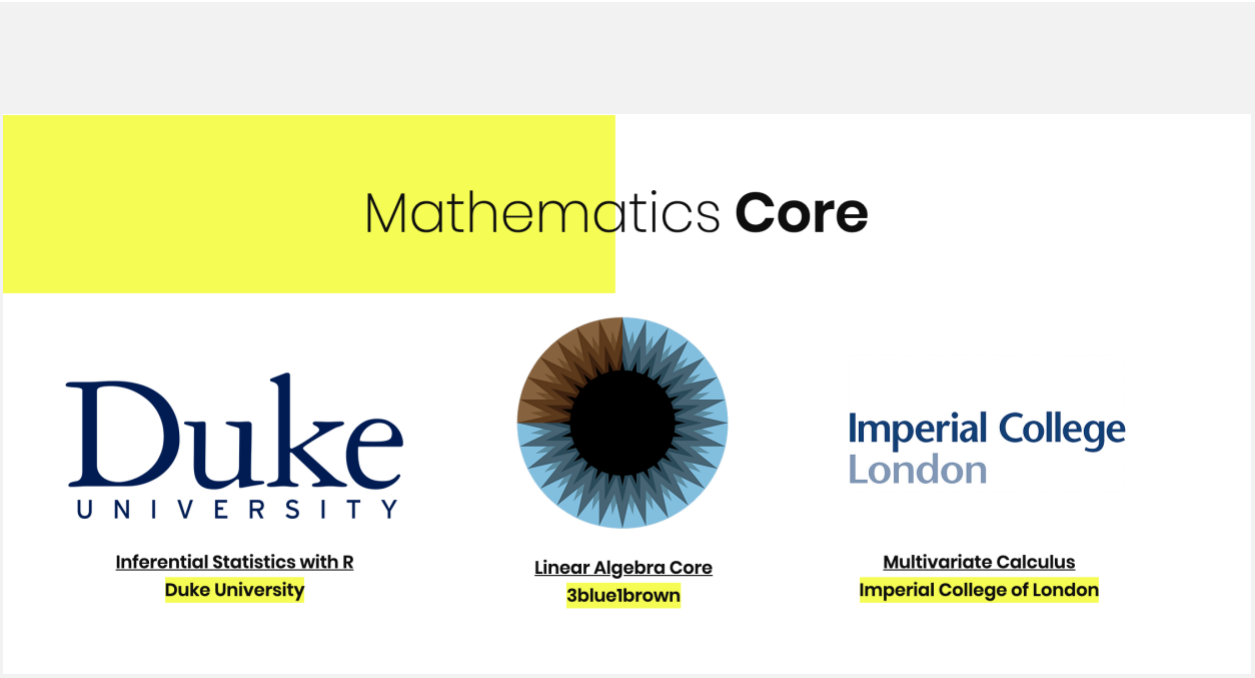
Focused on learning all the key components a data scientist needs to know! From procuring the data to visualizing it, the idea was to truly understand what exactly a data scientist does and learn the necessary tools needed for the job. Discovering the DataCamp's Data Scientist with Python track was a blessing, for I was able to learn all the various concepts and techniques in an application-based format. The learning structure is

beautiful because as soon as you watch a short video, you are thrown into a place to immediately write code and apply that concept to solve a short case study.



Mathematics Core:

Mathematics is truly a game-changer for a data scientist! Understanding the math behind the machine learning algorithms I was learning was helping me understand the best use case for them. For example, understanding the difference between a Bernoulli Trial and a Poisson Process will help you model your samples accordingly for the best testing of the scenario. Thus, I focused on learning and refreshing linear algebra,calculus, probability, and statistics needed for machine learning and modeling.



Machine Learning Core:

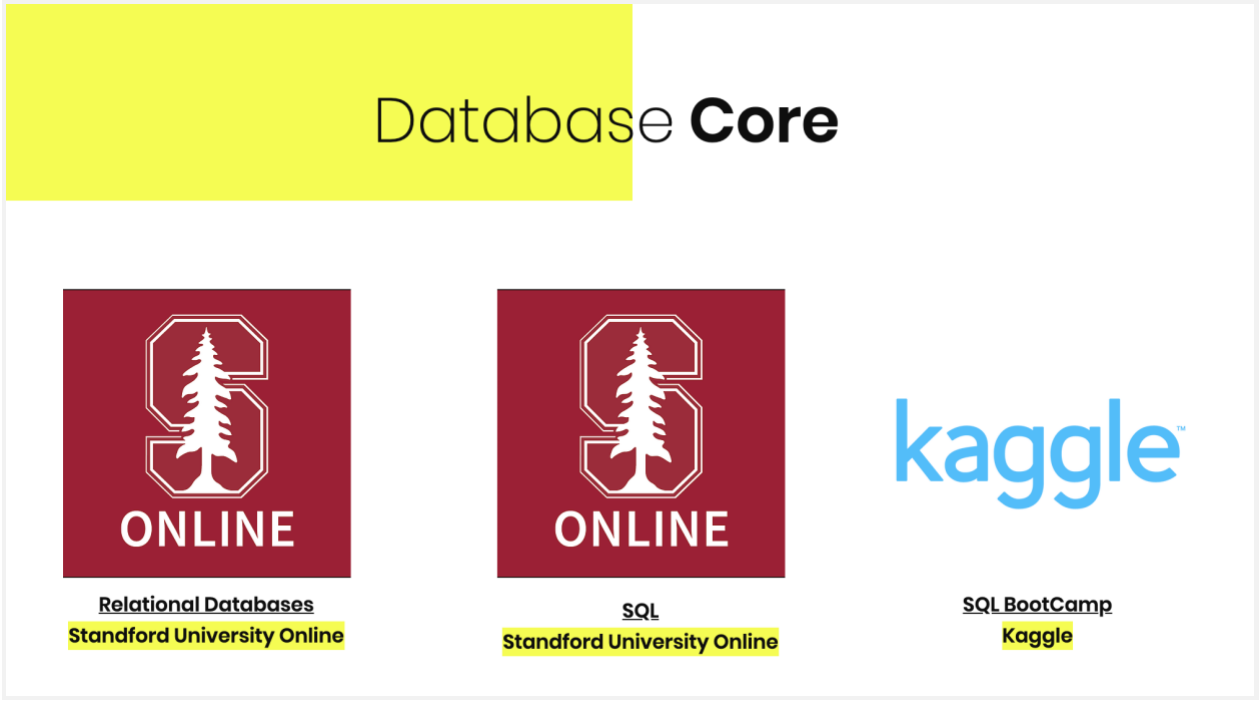
Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Today, ML is in high demand to solve business problems! The idea was to learn about the most effective machine learning techniques and gain practice by applying them on projects. I took the ML by Standford, a famous Coursera Course. Taught by the famous Andrew Ng, Google Brain founder and former chief scientist at Baidu, Stanford University's Machine Learning covers all aspects of the machine learning workflow and several algorithms. Using this course and many other YouTube sources, I learned the following algorithms

- Linear Regression
- Logistic Regression
- Decision Trees
- KNN (K- Nearest Neighbours)
- K-Means

- Naïve Bayes
- Dimensionality Reduction (Principal Components Analysis)
- Random Forests
- Dimensionality Reduction Techniques
- Support Vector Machines
- Gradient Boosting Machines
- XGBOOST

Database Core:

Learning Databases and SQL was one of my top priorities. I started refining my SQL course by watching a 6-hour video by Mosh Amedani who goes through SQL in a very simple yet informative way. He also gives you exercises for practice with a cheat sheet. Stanford University's Introduction to Databases covers database theory comprehensively while introducing several open-source tools. I learned Importing & Cleaning Data with Python from DataCamp. This track focused on teaching the mechanics of preparing data for analysis and/or visualization. The final key component was going through a SQL Bootcamp organized by Kaggle. All these together have given me a good command over this core.



Application Core

This core was all about solving case studies, working on personal projects using open-source data and publishing the work.

DataCamp's track introduced me to a good collection of case studies or small projects. I solved a number of these before I began my first project- Exploring the Titanic Data Set and predicting the survival rate. It was a simple yet very powerful project that helped me apply the various techniques I learned.

I then worked on a topic of my interest to understand the patterns in Global Terrorism. This was a huge data set with around 180k data instances. The set was published on Kaggle and it had issues that needed to be addressed. I conducted a good EDA and solved a couple of those problems using ML

It is very important for a Data Scientist to have a GitHub profile to host all the codes of the project he/she has undertaken. Potential employers not only see what you have done, how you have coded and how frequently / how long you have been practicing data science. I thus set up my own Github Profile and constructed a website to present my work.

Predicting Unknown terrorist groups behind attacks
Global Terrorist Dataset - Kaggle

Predicting the Survival on the Titanic
Titanic Dataset- Kaggle

Analyzing a Dying Piano industry using Twitter data
Dataset scrapped using Twitter API

Key Take-aways:-

1. **Never give up when life knocks you down!** (Super Important)
2. To truly grasp any subject, its important to **learn actively.** Attending classes on campus was fun, but I realized there was so much more I needed to learn.
3. **Understanding the Problem** — Before solving any case study, I made it a point to truly question what exactly was going on? This has now installed me a healthy habit of diving into the topic so then construct a good plan for the solution.
4. **Real-world data is dirty —** I learned this the hard way when I was dealing with the scrapped Twitter data and Global Terrorism data sets(where 1993 data was missing) Therefore, data preprocessing is so critical to master. It is the most important stage as it could occupy 50%-70% of the whole workflow, just to clean the data to be fed to your models.
5. When you don't know what you don't know, and you think the data preprocessed is already clean enough and ready to feed to your models, therein lies a risk of **building the correct models with the wrong data**. In order words, always try to question yourself if the data is technically correct with the domain knowledge that you have, scrutinize the data with a stringent threshold to check for any other outliers, missing or inconsistent data in the whole

datasets. I was particularly careful about this after I made a mistake of feeding the models with the wrong data, just because of a flaw in one of the preprocessing steps.

6. **Application is key!** Knowing different algorithms is great? But understanding their use cases and scenario is extremely important too! Eg. KNN might not be that useful when we have a large number of features or variables (#curseofdimensionality). Building different models was a steep learning curve for me as a person who was still learning from MOOCs and textbooks. Fortunately, Scikit-learn came to my rescue as it is easy to learn for fast models prototyping and implementation in Python. In addition, I also learned how to optimize the models and fine-tuned the hyperparameters for each model using several techniques.

End of Summer!

Overall, it was a very productive summer for me. Sometimes, when life throws you in a spot, it only makes you emerge as a stronger individual. Not getting an internship was quite disheartening, but I take it as a turning point in my life to achieving my career of becoming a Data Scientist. The pseudo internship has definitely reaffirmed my passion for Data Science. The research and development phase, the curiosity and passion to solve business problems using data have all contributed to my interest & passion for this beautiful field.I now have a set plan for the Fall, where I will work on my pending projects, learn new techniques like Big Data Analytics, Advanced ML, and Deep learning !