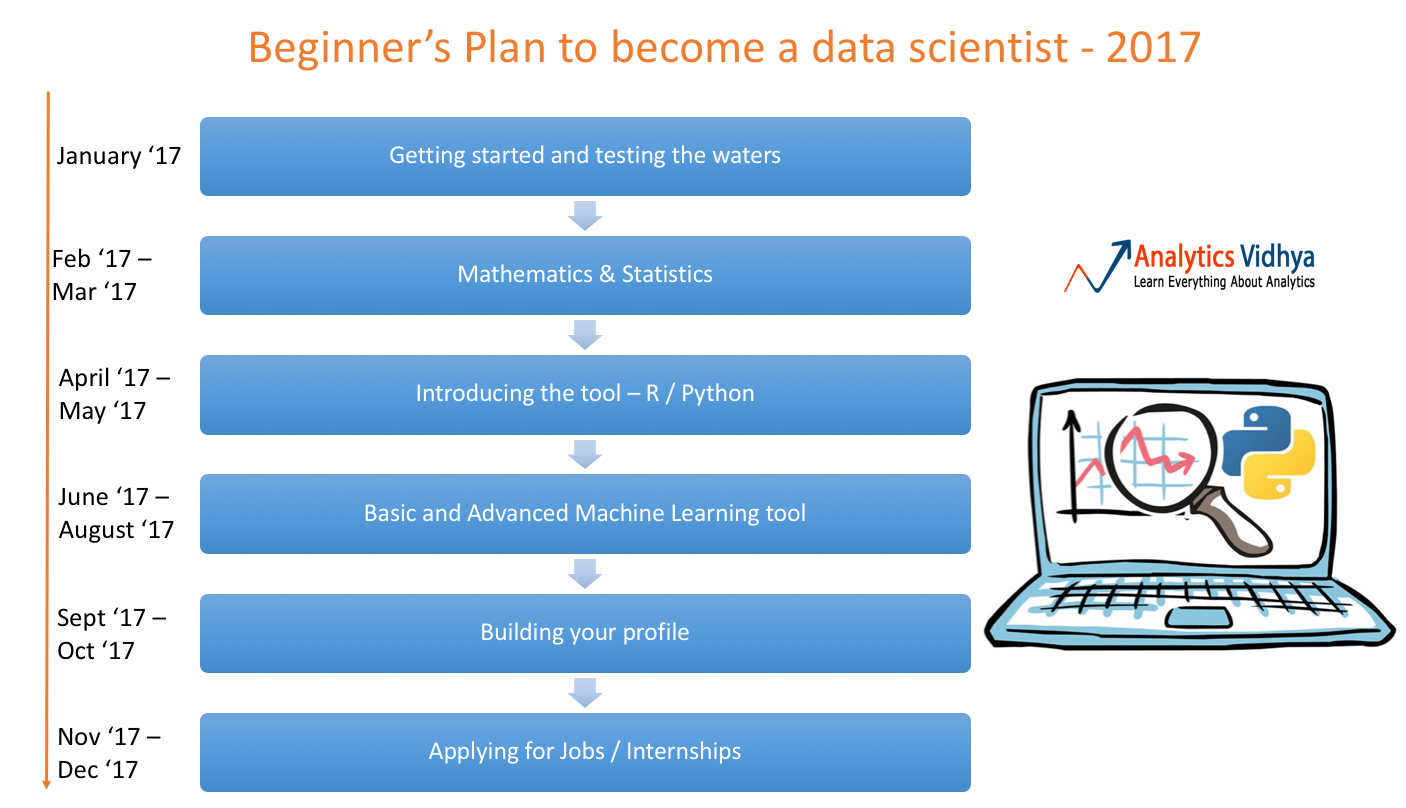
### Transitioner Data Scientist

* Learn basic mathematics and statistics required for data science
* Develop a basic understanding of machine learning algorithms
* Work on projects and create a portfolio of projects
* Skills required to land your first data science internship / job.
* Time spent ~ 5 hours / day

### Structure for your 2017 journey:

The structure of the path is similar, but you will need to accelerate your learning in the first half of the plan. Start by going through [this article](https://www.analyticsvidhya.com/blog/2014/05/planning-late-career-shift-analytics-big-data-prepared/) and go through a few [success stories](https://www.analyticsvidhya.com/blog/category/stories/) to understand what a transition would entail. Once you are set for the journey, follow the plan by sticking to these timelines.

* Step 1: Getting started and testing the waters (1 week in January ’17)
* Step 2: Mathematics & Statistics (Jan ’17 – March ’17)
* Step 3: Introducing the tool – R / Python (March ’17 – April ’17)
* Step 4: Basic & Advanced machine learning tools (May ’17 – July ’17)
* Step 5: Building your profile (Aug ’17 – Oct ’17)
* Step 6: Applying for Jobs  (Nov ’17 – Dec ’17)



3.1: Getting started and testing the waters

**Time suggested: 4 weeks (January 2017)**

At this stage, it is important to understand why you want to become a data scientist? What are your strengths and weaknesses? Do you know what it takes to be a Data Scientist? You must answers these questions before jumping on the boat of Data Science journey.

Watch [this](https://www.youtube.com/watch?v=rIofV14c0tc)[excellent](https://www.youtube.com/watch?v=rIofV14c0tc)[video](https://www.youtube.com/watch?v=rIofV14c0tc) where **Tetiana Ivanova** describes how she became a Data Scientist without going through a Masters or doctorate program in data science and with help of Meetups.

Here are some additional resources you can use to answer these questions:

1. [What is Data Science?](https://datajobs.com/what-is-data-science) – This article by Data Jobs will give you a broad perspective of how data science is being used in Netflix and Amazon. Also, it will highlight the skill set required for Data Science.
2. [Should I become a Data Scientist?](https://www.analyticsvidhya.com/blog/2014/09/become-data-scientist-business-analyst/) This article points out some questions for you to decide whether you are fit for a Data Scientist role. I suggest you must go through this article before proceeding further.
3. Next, you should attend local meetups in your area. Go out and find out what people are talking about Data Science / Machine Learning. Meetups not only help you learn the tools and techniques, they provide you with a network of people in similar industry which helps you in finding the right jobs and internships later on.

Go ahead and think through these aspects of choosing a career in data science. This decision is going to decide the next 11 months of your life.

3.2: Basics of Mathematics and Statistics

**Time suggested: 8 weeks (February 2017 – March 2017)**

Topics to be covered:

* Descriptive Statistics – 1 week
* Probability – 2 weeks
* Inferential Statistics – 2 weeks
* Linear Algebra – 1 week
* Structured Thinking – 2 weeks

Descriptive Statistics – 1 week

* **Course (mandatory) –**[Descriptive Statistics from Udacity](https://www.udacity.com/course/intro-to-descriptive-statistics--ud827) is a basic and must do course to get started.
* **Books (optional)**– Supplement your online course with [online stats book](http://onlinestatbook.com/2/index.html). A good book for any one looking for learning basic statistics.

Probability – 2 weeks

* **Course (mandatory)** – [Introduction to probability – The science of uncertainty](https://courses.edx.org/courses/MITx/6.041x_1/1T2015/info) is an excellent course on edX to learn concepts of probability like conditional probability and probability distributions.
* **Books (optional)** – The textbook [Introduction to probability – Berkley’s stats 134 standard textbook](http://www.stat.berkeley.edu/~aldous/134/grinstead.pdf) will supplement the course above and can be used as a good reference material.

Inferential Statistics – 2 weeks

* **Course (mandatory)**– [Intro to Inferential Statistics from Udacity](https://www.udacity.com/course/intro-to-inferential-statistics--ud201) – Once you have gone through the descriptive statistics course, this course will take you through statistical modeling techniques and advanced statistics.
* **Books (optional)** – [Online Stats Book](http://onlinestatbook.com/2/index.html) – This online book can be used for a quick reference for inference tasks.

Linear Algebra – 1 week

* **Course (mandatory)**
  + [Linear Algebra – Khan Academy](https://www.khanacademy.org/math/linear-algebra) : This concise and an excellent course on Khan Academy will equip you with the skills necessary for Data Science and Machine Learning.
* **Books (optional)**
  + [Linear Algebra/ Levandosky](https://www.amazon.com/gp/product/0536667470/ref=as_li_qf_sp_asin_il_tl?ie=UTF8&camp=1789&creative=9325&creativeASIN=0536667470&linkCode=as2&tag=theopesoudats-20&linkId=YLLIM4C4LJSOZIQR) – This is an often cited book to Stanford graduates for Linear Algebra.
  + [The Manga guide to Linear Algebra](https://www.amazon.com/Manga-Guide-Linear-Algebra/dp/1593274130/ref=as_li_ss_tl?ie=UTF8&qid=1452829566&sr=8-1&keywords=The+Manga+Guide+to+Linear+Algebra&linkCode=sl1&tag=theopesoudats-20&linkId=3d46150bdab2ef0ba9805cee31eb4d2d) – This is a fun filled Linear Algebra book which keeps Machine Learning in context. You will never forget these Algebra lessons for sure.

Structured Thinking – 2 weeks

* **Articles (mandatory):**These articles will guide you to structure your thinking process to approach problems in a better way so as to improve your efficiency.
  + [How to train your mind for analytical thinking?](https://www.analyticsvidhya.com/blog/2014/01/train-mind-analytical-thinking/)
  + [Tools for improving structured thinking](https://www.analyticsvidhya.com/blog/2014/02/tools-structured-thinking/)
  + [The art of structured thinking and analyzing](https://www.analyticsvidhya.com/blog/2013/06/art-structured-thinking-analyzing/)
* **Competitions (mandatory):**No amount of theory can beat practice. This is a [strategic thinking problem](https://datahack.analyticsvidhya.com/contest/practice-problem-strategic-thinking-ii/) which will test you on your thinking process. Also, keep an eye on business case studies as they help in structuring your thoughts tremendously.

3.3: Introducing the tool – R / Python

**Time suggested: 8 weeks (April 2017 – May 2017)**

**Topics to be covered:**

* Tools (R/Python) – 4 weeks
* Exploration and Visualization (R/Python) – 4 weeks
* Feature Selection/ Engineering

Tools

1. R

* **Course**– [Interactive Intro to R Programming Language by DataCamp](https://www.datacamp.com/courses/free-introduction-to-r) – An excellent course by DataCamp to give you hands-on experience in R. The course includes interactive examples You will never feel bored while learning R.
* **Books** – [R for Data Science](http://r4ds.had.co.nz/) – This is your one stop solution for referencing basic materials on R.
* **Blogs/Articles**
  + [This](https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/) article will serve a great point for collating the entire process of model building starting from installation of RStudio/R.
  + [R-bloggers](https://www.r-bloggers.com/) – This is one of the most recommended blog for R- users. Every R practitioner should keep this blog bookmarked. It has some of the most effective and practical R tutorials. Bookmark it now.

2. Python

* **Course (mandatory)** – [Intro to Python for Data Science](https://www.datacamp.com/courses/intro-to-python-for-data-science) – An interactive course developed by DataCamp to facilitate Data Science learning using Python.
* **Books (mandatory)** – [Python for Data Analysis](http://www3.canisius.edu/~yany/python/Python4DataAnalysis.pdf) – This book covers various aspects of Data Science including loading data to manipulating, processing, cleaning and visualizing data. Must keep reference guide for Pandas users.
* **Blogs/Articles (optional)**
  + [A Complete Tutorial to Learn Data Science with Python from Scratch](https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/): This article will serve as a quick guide to learning Data Science using Python.

Exploration and Visualization

1. R

* **Course**
  + [Exploratory Data Analysis](https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/) – This is an awesome course by Johns Hopkins University on Coursera. You will need no other course to perform visualization and exploratory work in R.
* **Blogs/Articles**
  + [Comprehensive guide to Data Exploration in R](https://www.analyticsvidhya.com/blog/2015/04/comprehensive-guide-data-exploration-r/) – This will be a one-stop article that I will suggest you to go through carefully and follow every step. This is because the steps mentioned in the article are the same steps you will be using while solving any data problem or a hackathon problem.
  + [Cheat sheet – Data Exploration in R](https://www.analyticsvidhya.com/blog/2015/10/cheatsheet-11-steps-data-exploration-with-codes/) – This cheat sheet contains all the steps in data exploration with codes. I suggest you to take out a print and paste it on your wall for quick reference.

2. Python

* **Course (optional)**
  + [Intro to Data Analysis](https://www.udacity.com/course/intro-to-data-analysis--ud170) – This is an excellent course by Udacity on Data Exploration using Numpy and Pandas.
* **Blogs/Articles (mandatory)**
  + [Comprehensive guide to Data Exploration using Python NumPy, Matplotlib and Pandas](https://www.analyticsvidhya.com/blog/2015/04/comprehensive-guide-data-exploration-sas-using-python-numpy-scipy-matplotlib-pandas/) – This is a sufficient and comprehensive article which uses the most popular Python libraries for exploration and visualization purposes.
  + [9 popular ways to perform Data Visualization in Python](https://www.analyticsvidhya.com/blog/2015/05/data-visualization-python/) – This article presents the most commonly used graphs and plots used in Data Exploration along with Python codes. This is a must bookmarked article for people working in Data Science using Python.
* **Books (optional)** – [Python for Data Analysis](http://www.cin.ufpe.br/~embat/Python%20for%20Data%20Analysis.pdf) – A one stop solution for your Data Exploration and Visualization in Python.

Feature Selection/ Engineering

* **Blog** – [A Comprehensive Guide to Data Exploration](http://www.cin.ufpe.br/~embat/Python%20for%20Data%20Analysis.pdf): This article will explain underlying techniques of feature engineering and different methods for feature creation
* **Books (optional)** –[Mastering Feature Engineering](https://www.amazon.com/Mastering-Feature-Engineering-Principles-Techniques/dp/1491953241): This book is master piece to learn feature engineering. Not only will you learn how to implement feature engineering in a systematic way. You will also learn different methods involved in feature engineering.

3.4: Basic & Advanced machine learning tools

**Time suggested: 12 weeks (June 2017 – August 2017)**

Topics to be covered (June 2017 – July 2017):

* Basic Machine Learning Algorithms.
  + Linear Regression
  + Logistic Regression
  + Decision Trees
  + KNN (K- Nearest Neighbours)
  + K-Means
  + Naïve Bayes
  + Dimensionality Reduction
* Advanced algorithms (August 2017)
  + Random Forests
  + Dimensionality Reduction Techniques
  + Support Vector Machines
  + Gradient Boosting Machines
  + XGBOOST

Linear Regression

* **Course**
  + [Machine Learning by Andrew Ng](https://www.coursera.org/learn/machine-learning) – There is no better resource to learn Linear Regression than this course. It will give you a thorough understanding of linear regression and there is a reason why Andrew Ng is considered the rockstar of Machine Learning.
* **Blogs/Articles**
  + [This](https://onlinecourses.science.psu.edu/stat501/node/250) lesson out of PennState Stat 501 course outlines the main features of Linear Regression ranging from a simple definition of a Linear Regression to determining the goodness of fit of a regression line.
  + [This](https://www.analyticsvidhya.com/blog/2015/10/regression-python-beginners/) is an excellent article with practical examples to explain Linear Regression with code.
* **Books**
  + [The Elements of Statistical Learning](http://www-stat.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf) – This book is sometimes considered the holy grail of Machine Learning and Data Science. It explains Machine Learning concepts mathematically from a Statistics perspective.
  + [Machine Learning with R](https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r) – This is a book I personally use to have a brief understanding of Machine Learning algorithms along with their implementation code.
* **Practice**
  + [Black Friday](https://datahack.analyticsvidhya.com/contest/black-friday/) – Like I already said – No amount of theory can beat practice. Here is a regression problem that you can try your hands on for a deeper understanding.

Logistic Regression

* **Course (mandatory)**
  + [Machine Learning by Andrew Ng](https://www.coursera.org/learn/machine-learning)– The week 3 of this course will give you a deeper understanding of the one of the most widely used classification algorithm.
  + [Machine Learning: Classification](https://www.coursera.org/learn/ml-classification) – Week 1 and 2 of this practical oriented Specialization course using Python will satiate your knowledge thirst about Logistic Regression.
* **Blogs/Articles (optional)**
  + [Logistic Regression by Machine Learning Mastery](http://machinelearningmastery.com/logistic-regression-for-machine-learning/) – This is an excellent non-code based approach to Logistic regression to deepen your knowledge. I suggest you to have a look at it.
* **Books (optional)**
  + [Introduction to Statistical Learning](http://www-bcf.usc.edu/~gareth/ISL/) – This is an excellent book with a quality content on Logistic Regression’s underlying assumptions, statistical nature and mathematical linkage.
* **Practice (mandatory)**
  + [Loan Prediction](https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/) – This is an excellent competition to practice and test your new Logistic Regression skills to predict whether loan status for a person was approved or not.

Decision Trees

* **Course (mandatory)**
  + [Machine Learning: Classification](https://www.coursera.org/learn/ml-classification) – Week 3 and 4 in this course is about the working of decision trees, preventing overfitting and handling missing values
* **Blogs/Articles (mandatory)**
  + [Technical Overview of decision trees](http://www.kdnuggets.com/2016/10/decision-trees-concise-technical-overview.html) – This is a quick overview of decision trees and a must read for anyone new to decision trees.
  + [Complete tutorial on tree based modeling](https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/) – This is a python based tutorial on decision trees. For the sake of decision trees, read only sections 1-6 in this article.
* **Books (mandatory)**
  + [Introduction to Statistical Learning](http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf) – Section 8.1 and 8.3 explain the basics of decision trees through theory and practical examples.
  + [Machine Learning with R](https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r) – Chapter 5 of this book provides you the best explanation of Machine Learning Algorithms available in the market. Here, the decision trees are explained in an extremely non-intimidating and easier style.
* **Practice (mandatory)**
  + [Loan Prediction](https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/) – This is an excellent competition to practice and test your new Logistic Regression skills to predict whether loan status for a person was approved or not.

KNN (K- Nearest Neighbors)

* **Course (mandatory)**
  + [Machine Learning – Clustering and Retrieval](https://www.coursera.org/learn/ml-clustering-and-retrieval): Week 2 of this course progresses to k-nearest neighbors from 1-nearest neighbor and also describes the best ways to approximate the nearest neighbors. It explains all the concepts of KNN using python.
* **Blogs/Articles (mandatory)**
  + [Introduction to k-nearest neighbors: simplified](https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/) – This basic article describes when to use KNN, the ways in which k can be chosen and the way in which KNN algorithm works.
  + [Learning KNN algorithm using R](https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/) – This article is a comprehensive guide to learning KNN with hands-on codes for future references.

K-Means

* **Course**
  + [Machine Learning Course – Unsupervised Learning with K-means algorithm](https://www.coursera.org/learn/machine-learning): Week 8 of this discusses how to use course how K-means algorithm is used for handling unstructured data.
* **Blog**
  + [An Introduction to Clustering and different methods of clustering](https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/): In this article, you will learn what is k-means clustering and the intricacies involved in that. It will give you a step by step approach how K-means algorithm works.

Naive Bayes

* **Course**
  + [Intro to Machine Learning](https://www.udacity.com/course/intro-to-machine-learning--ud120): Take this course to see Naive Bayes in action. In this course, Sebastian Thrun has explained Naive Bayes in Simple English.
* **Blog / Article**
  + [6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python)](https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/) : This article will take you through Naive Bayes algorithm in detail. In this guide, you will learn how Naive Bayes algorithm works, applications and many more. It will also give you hands-on knowledge of building a model using Naive Bayes.
  + [Naive Bayes for Machine Learning](http://machinelearningmastery.com/naive-bayes-for-machine-learning/) : This is one of the most comprehensive articles I have come across. Go through this article to have a complete understanding of why naive bayes algorithm is important for machine learning.

Dimensionality Reduction

* **Course**
  + [Machine Learning – Dimensionality Reduction:](https://www.coursera.org/learn/machine-learning) Week 8 of this course will walk you through dimensionality reduction and how Principal Components Analysis can be used for data compression of complex data.
* **Blog / Article**
  + [Beginners Guide To Learn Dimension Reduction Techniques:](https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/)In this article, you will learn why dimension reduction is important in machine learning and the various techniques of dimension reduction.

Random Forests

* **Videos (mandatory)**
  + [How Random Forest algorithm works?](https://www.youtube.com/watch?v=loNcrMjYh64) – Watch this video to have a visual perspective of how the Random Forest algorithm works.
* **Books (optional)**
  + [Introduction to Statistical Learning](http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf) – Section 8 explains the basics of Random Forests including bagging and boosting through theory and practical examples.
  + [Applied predictive modeling](http://appliedpredictivemodeling.com/) –  Chapter 8
* **Blogs/Articles (mandatory)**
  + [A tutorial on tree based modeling from scratch](https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/) – This is an excellent article on trees based modeling using python. I suggest you to bookmark it right now.
  + [Random Forests](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm) – This blog explains the entire working, nuts and bolts of Random Forest.

Gradient Boosting Machines

* **Blogs/Articles (mandatory)**
  + [Guide on Boosting methods](https://www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-methods/)
  + [Parameter tuning GBM](https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/)
  + [Machine Learning Mastery- GBM](http://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/)
* **Presentation (mandatory):**[Here](http://www.slideshare.net/mark_landry/gbm-package-in-r) is an excellent presentation on GBM. It contains the prominent features of GBM and the advantages and disadvantages of using it to solve real-world problems. It is must see article for somebody trying to understand GBM.

XGBOOST

* **Blogs /Articles (mandatory)**
  + [Official Introduction XGBOOST](http://xgboost.readthedocs.io/en/latest/model.html) – Read the documentation of hackathons winning algorithm. It is an improvement over GBM and is right now the most widely used algorithm for winning competitions.
  + [Using XGBOOST in R](https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/) – An excellent article on deploying XGBOOST in R using a practical problem at hand.
  + [XGBOOST for applied Machine Learning](http://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/) – An article by Machine Learning Mastery to evaluate the performance of XGBOOST over other algorithms.

Support Vector Machines

* **Course (mandatory)**
  + [Machine Learning by Andrew Ng](https://www.coursera.org/learn/machine-learning/home/week/7) – Week 7 of this course is an interesting place to start your SVM journey.
* **Books (mandatory)**
  + [Introduction to Statistical Learning](http://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/)– Chapter 9 of the book contains a detail discussion about SVMs and the ways to deploy them.
* **Blogs/Articles (optional)**
  + [Understanding support vector machines](https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/) – This is an excellent article to understand an algorithm practically using examples.
  + [SVM by Machine Learning Mastery](http://machinelearningmastery.com/support-vector-machines-for-machine-learning/) – This article discusses the different types of kernels employed in SVM and their uses.

3.5: Building your profile

**Time suggested: 8 weeks (September 2017 – October 2017)**

Topics to be covered:

1. GitHub Profile Building
2. Practice via competitions
3. Discussion Portals

GitHub Profile Building (mandatory)

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.

Also, codes on GitHub open up avenues for open source projects which can highly boost your learning. If you don’t know how to use Git, you can learn from [Git and GitHub](https://www.udacity.com/course/how-to-use-git-and-github--ud775" \t "_blank) on Udacity. This is one of the best and easy to learn course to manage the repositories through terminal.

Practice via competitions (mandatory)

Time and again, I have stressed on the fact that practice beats theory. Moreover coding in hackathons brings you closer to developing data products in real life for solving real world problems. Below are most popular platforms to participate in Data Science/ Machine Learning Competitions.

1. [Analytics Vidhya Datahack](https://datahack.analyticsvidhya.com/contest/all/)
2. [Kaggle competitions](https://www.kaggle.com/competitions)
3. [Crowd Analytix human layer](https://www.crowdanalytix.com/community)

Discussion Forums (optional)

Discussions are a great way to learn in a peer-to-peer setup from finding an answer to a question you stuck to providing answers to someone else’s questions. Below are some of the discussion rich platforms which you should keep a tab on to clear your doubts.

1. [Analytics Vidhya Discussion Portal](https://discuss.analyticsvidhya.com/" \t "_blank)
2. [Kaggle Discussion](https://discuss.analyticsvidhya.com/)
3. [StackExchange](http://stackexchange.com/)

3.6: Apply for Jobs & Internships

**Time suggested: 8 weeks (November 2017 – December 2017)**

**Topics to be covered: Jobs / Internships**

If you are here after diligently following the above steps, then you can be sure that you are ready for a Job / Internship position at any Data Science / Analytics or Machine Learning firms. But it becomes quite difficult to identify the right jobs. So, for the purpose of saving the trouble, I have created a list of portals which lists down Data Science/ Machine Learning jobs and Internships.

1. [Analytics Vidhya Job Portal](https://www.analyticsvidhya.com/jobs/" \l "/" \t "_blank)
2. [Datajobs](https://datajobs.com/)
3. [Kaggle Job portal](https://www.kaggle.com/jobs)
4. [Internshala](https://internshala.com/)

In order to prepare for these interviews, you should go through this [Damn Good Hiring Guide](https://www.analyticsvidhya.com/blog/2015/09/damn-good-hiring-path-hired-data-scientist/)