Self-Created Course Plan:
How to Become a Data Analyst / Data Scientist

Bernardo Di Chiara

Scope

This is a self-created plan of on-line courses which provides a time-efficient, convenient and affordable way to achieve the necessary competences to become a data analyst / data scientist. The courses are from well-known Universities and are all contained in Coursera platform [1]. The plan makes uses of single courses and of specializations and is the result of heavy modifications to different proposals by Coursera, based on feedback got through job ads, job interviews and other sources.

Courses' Structure

The courses are instructor-led, have a perfect balance of theory and practice and their structure ensures that the student has mastered the contents after all the sections have been completed.

Each course contains at least video lessons with transcriptions and presentation material, weekly exams and a final peer-reviewed assignment. For almost all the courses an additional text book is available as well as practical exercises (like R or Python labs). The best courses have also additional practice quiz and in-video quiz, in addition to the graded ones.

Plan's Structure

The courses have been organized in four modules. All together there are 23 on-line courses, one face-to-face course and some extra self-studying activities. In practice the required commitment is about 1500 hours. The suggested planned calendar time is two years.

Most Important Courses

Topic	Course	By
Statistics	Inferential Statistics (in R)	Duke University on Coursera
	Linear Regression and Modeling (in R)	
	Bayesian Statistics: From Concept to Data Analysis	University of California, Santa Cruz
R	Introduction to Probability and Data (exploratory data analysis) (in R)	Duke University on Coursera
	R Programming	Johns Hopkins University on Coursera
	Practical Machine Learning (in R)	
	Getting and Cleaning Data (in R)	
Python	Applied Machine Learning in Python	University of Michigan on Coursera
	Applied Plotting, Charting and Data Representation in Python	
	Introduction to Data Science in Python (data cleansing and manipulation)	
	Getting Started with Python / Python Data Structures / Using Python to Access Web Data / Using Databased with Python	
	Applied Text Mining in Python	
Databases	SQL for Data Science	University of California Davis on Coursera
	Hadoop Platform and Application Framework	Univ. of California, San Diego on Coursera
Econometrics	Business Metrics for Data-Driven Companies	Duke University on Coursera
	Data Visualization and Communication with Tableau	

full chronological plan with completion status on the next page

[1] The only exception is the Big Data course of the second module which is a classroom course in Helsinki.

October 29th 2018

Basic: basic statistics, data analysis techniques and R (6,5 months)

- Data Science Math Skills (Duke) (4 weeks) *COMPLETED Sep. 2017*
  - Probability Theory
- Introduction to Probability and Data (Statistics with R, Duke) (5 weeks) *COMPLETED Oct. 2017*
  - Exploratory Data Analysis, Data Preparation, Data Visualization, RStudio
- Inferential Statistics (Statistics with R, Duke) (5 weeks) *COMPLETED Nov. 2017*
  - Hypothesis Tests, Confidence Intervals, ANOVA, Chi-square test, Bootstrapping, ...
- The Data Scientist's Toolbox (Data Science, Johns Hopkins) (4 weeks) *COMPLETED Dec. 2017*
  - Overview of Data Analysis Techniques, GitHub
- R Programming (Data Science, Johns Hopkins) (4 weeks) *COMPLETED Jan. 2018*
  - Creating R functions, using loop functions, debugging, profiling, ...
- Getting and Cleaning Data (Data Science, Johns Hopkins) (4 weeks) *COMPLETED Feb. 2018*
  - R Interfaces, Data Cleansing

Intermediate: machine learning, SQL and Big Data (4,5 months)

- Linear Regression and Modelling (Statistics with R, Duke) (4 weeks) *COMPLETED Mar. 2018*
  - Simple and Multiple Linear Regression, Logistic Regression
- More about Regression Models: studying Logistic Regression from the 3 videos in the Regression Models course of the Data Science specialization and from chapter 8.4 of OpenIntro Statistics book *DONE*
- More about EDA: studying Hierarchical Clustering, K-Means Clustering and Dimension Reduction (SVD and PCA) from the Exploratory Data Analysis course of the Data Science specialization (8 videos and 4 R labs) *DONE*
- Practical Machine Learning (Data Science, Johns Hopkins) (4 weeks) *COMPLETED Apr. 2018*
  - Clustering, Principal Component Analysis, Singular Value Decomposition, Decision Trees, Bagging, Random Forests, Boosting, Forecasting, ...
- SQL for Data Science (University of California, Davis) (4 weeks) *COMPLETED May 2018*
  - Data modelling, ER diagrams, retrieving data from multiple tables (subqueries, joins, unions), filtering, sorting, calculations, aggregations, manipulating text and dates, case statements, creating tables, SQLite
- Self-study: how to retrieve and modify data from an SQLite database in R by using RSQLite R package *DONE*
- Big Data (classroom course with instructors by TalentGate/Haaga-Helia University of Applied Sciences) (3 days)
  - Handling data volume, velocity and variety, supervised and unsupervised machine learning hands-on with KNIME

  COMPLETED May 2018
- Hadoop Platform and Application Framework (University of California, San Diego) (5 weeks) *COMPLETED Jun. 2018*
  - HDFS, MapReduce, Spark, HiveQL, HBase, Pig, …

Advanced: Python, more machine learning and data visualization, econometrics (5 months)

- Getting Started with Python (Programming for Everybody, Univ. of Michigan) (7 weeks > 1 week) *COMPLETED Sep. 2018*
  - Conditional statements, iterations, functions
- Python Data Structures (Programming for Everybody, Univ. of Michigan) (7 weeks > 1 week) *COMPLETED Sep. 2018*
  - Files, lists, dictionaries, tuples
- Introduction to Data Science in Python (Applied Data Science with Python, University of Michigan) (4 weeks)
  - Loop functions, NumPy, arrays, Pandas, series, data frames, dates, data manipulation, statistical analysis

  COMPLETED Oct. 2018
- Applied Plotting, Charting & Data Representation in Python (Applied Data Science with Python, Univ. of Michigan) (4 weeks)
  - Graphical heuristics (Edward Tufte), Matplotlib architecture, scatterplots, line plots, bar charts, histograms, box plots, heatmaps, subplots, interactive charts, animations *ONGOING*
- Applied Machine Learning in Python (Applied Data Science with Python, University of Michigan) (4 weeks)
- Using Python to Access Web Data (Programming for Everybody, Univ. of Michigan) (6 weeks > 1 week)
  - Regular expressions, HTTP, Web services
- Using Databased with Python (Programming for Everybody, Univ. of Michigan) (5 weeks > 1 week)
- Business Metrics for Data-Driven Companies (Excel to MySQL: Analytic Techniques for Business, Duke) (4 weeks)

continues on the next page

Pro: more advanced statistics, more econometrics, more Python based analysis (6 months)

- Bayesian Statistics: From Concept to Data Analysis (University of California, Santa Cruz) (4 weeks)
- Mastering Data Analysis in Excel (Excel to MySQL: Analytic Techniques for Business, Duke) (6 weeks)
- Data Visualization and Communication with Tableau (Excel to MySQL: Analytic Techniques for Business, Duke) (5 weeks)
- Applied Text Mining in Python (Applied Data Science with Python, University of Michigan) (4 weeks)
- Applied Social Network Analysis in Python (Applied Data Science with Python, University of Michigan) (4 weeks)

More About the Plan's Structure

The first module contains the first courses of the "Statistics with R" specialization by Duke University as well as selected courses of the "Data Science" specialization by Johns Hopkins University. It allows acquiring the basic competences to start to work as a data analyst/data scientist: statistics, R programming, data analysis techniques.

The second module uses courses from the above-mentioned specializations to add other essential skills like fitting regression models and using machine learning algorithms to make previsions. It also contains courses that teach how to gather information from different types of databases.

The second module also contains selected lessons and exercises from the statistical courses of the Data Science specialization which cover arguments not contained in the corresponding courses from Duke University. The statistical courses from Duke University are preferred since they are of better quality and are definitely superior didactically. The second module contains also the only course which is not an on-line Coursera course. It is a classroom course about Big Data that, while it is not essential to proceed with the study plan, it adds a lot of value since it allows putting different tools and methods into context.

The third module focuses on a new programming language: Python. Two basic programming courses are followed by the first three courses of the specialization "Applied Data Science with Python" by the University of Michigan, which allows learning how to visualize data and developing prediction models by using this new language. Also, more advanced machine learning techniques (including vector machines, neural networks) are included. Other two Python courses are added to learn how to interface with Web Data and Databases in Python. The module includes also a course of econometrics from the specialization "Excel to MySQL: Analytic Techniques for Business" from Duke University.

The fourth module completes the specializations started in the previous modules and includes a more advanced statistical course, more econometrics related courses and more Python based data analysis courses.

Plan's Flexibility

Note that the courses could be arranged in different order, provided that the courses within a specialization are taken in order and that statistics courses are completed before the first machine learning course. Also, the Python specialization assumes some programming and statistical knowledge and therefore it cannot be taken if the first two modules are not completed [2]. The Hadoop course requires some Python knowledge and therefore postponing it after the first two Python coursee of the third module could be an alternative.

The order described above is optimal for learning since it allows going back to a major topic and refresh it after a while. On the other hand, sticking to a certain specialization till all the wished courses are done and then moving to another one would be more efficient cost wise. There are many sessions per year for each course (monthly, be-weekly or weekly, depending on the course).

The duration of the courses indicated here is the duration suggested by Coursera, which does not require a full-time commitment. Courses can be run slower or quicker but the peer-reviewed final project shall be submitted according to fixed schedules.

Costs

Access to the video lessons is free (but for some courses is limited to the first week only). In order to access to the graded exams, the final peer-reviewed project and to obtain the certificate, normally it is necessary to enroll to the premium version. The cost is about 40€ per month for each specializations. For single courses the allowed calendar time to complete the course with no extra fee might be slightly longer.

[2] The (11 week) Machine Learning course from Stanford University could be an alternative to the first 3 courses of the specialization "Applied Data Science with Python" from the University of Michigan for the third module but it might make harder to follow the other two Python courses in the specialization.