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

Bernardo Di Chiara

This is a self-created course plan to achieve the necessary competences to become a data analyst. The courses are all contained in Coursera platform. 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.

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

The courses have been organized in five groups.

The "Statistics with R" specialization by Duke University together with selected courses of the "Data Science" specialization by Johns Hopkins University allow to acquire the necessary competences to start to work as a data analyst/data scientist: statistics, R programming, data analysis techniques. A SQL course completes this first set of ten courses, as well as some 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 didactically are definitely superior.

The third block of courses allows to specialize as a Business Analyst and uses a third specialization from Duke University: "Excel to MySQL: Analytic Techniques for Business". It also includes a course that teaches about the most used tools for big data.

The fourth block consists of more advanced courses and is dedicated to performing exploratory data analysis, inferential statistics and prediction models by using a second programming language, Python. Also, more advanced machine learning techniques (including vector machines, neural networks) are included. The courses are part of the "Applied Data Science with Python" specialization by the University of Michigan.

The last optional block consists in the final projects of a couple of specializations.

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 done 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. The order described below is optimal for learning since it allows to go back to a major topic 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 (at least a couple per month, if not weekly).

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.

April 6th 2018

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

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

Intermediate: more advanced statistics, machine learning and SQL (4,5 months)

- Linear Regression and Modelling (Statistics with R, Duke) (4 weeks) DONE!
  - 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) DONE!
  - Clustering, Principal Component Analysis, Singular Value Decomposition, Decision Trees, Random Forests, Boosting, Forecasting, ...
- SQL for Data Science (University of California) (4 weeks) *ongoing!*
- Bayesian Statistics (Statistics with R, Duke) (5 weeks)

Advanced: business metrics, fancy visualization tools and big data (6 months)

- Hadoop Platform and Application Framework (University of California, San Diego) (5 weeks) *high priority!*
  - Hadoop, Spark and MapReduce
- Business Metrics for Data-Driven Companies (Excel to MySQL: Analytic Techniques for Business, Duke) (4 weeks) *high priority!*
- 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) *high priority!*
- Managing Big Data with MySQL (Excel to MySQL: Analytic Techniques for Business, Duke) (5 weeks)

Pro: more machine learning, Python (5 months)

- Introduction to Data Science in Python (Applied Data Science with Python, University of Michigan) (4 weeks) *important!*
- Applied Plotting, Charting & Data Representation in Python (Applied Data Science with Python, University of Michigan) (4 weeks) *important!*
- Applied Machine Learning in Python (Applied Data Science with Python, University of Michigan) (4 weeks) *important!*
- 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)
- *As an alternative, check the Machine Learning course from Stanford University (11 weeks)*

Completing: final projects of two specializations (4 months)

- Increasing Real Estate Management Profits: Harnessing Data Analytics (Excel to MySQL: Analytic Techniques for Business, Duke) (8w)
- Statistics with R Capstone (Statistics with R, Duke) (8w)