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

Bernardo Di Chiara

## Scope

This is a self-created course plan 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. 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 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 five modules.

The "Statistics with R" specialization by Duke University together with selected courses of the "Data Science" specialization by Johns Hopkins University allow acquiring 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, which is organized in two modules. The second module contains also 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. [1]

The third module provides knowledge about big data tools and it allows learning how to visualize data and developing prediction models by using a second language: Python. Also, more advanced machine learning techniques (including vector machines, neural networks) are included. This is achieved with the first part of the specialization "Applied Data Science with Python" by the University of Michigan. The module also includes the first course of the specialization "Excel to MySQL: Analytic Techniques for Business" from Duke University, which gives an introduction about business metrics. The fourth module uses the same specializations of the third module and is meant as a continuation.

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

## 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, it is necessary to enroll to the premium version. The cost is about 40€ per month for specializations. For single courses the allowed calendar time to complete the course with no extra fee might be slightly longer

## 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 order described below is optimal for learning since it allows to go back to a major topic and refresh it after a while. On the other hand,

[1] The first course of the first module is suggested for those who do not have a good mathematical background. Otherwise, it can be skipped.

[2] The (11 week) Machine Learning course from Stanford University is a good alternative to the specialization from the University of Michigan for the third module but it might make harder to follow the two Python courses in the fourth module.

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.

## 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, Bagging, Random Forests, Boosting, Forecasting, ...
- SQL for Data Science (University of California) (4 weeks) *ongoing!*
- Bayesian Statistics (Statistics with R, Duke) (5 weeks)

## Advanced: big data, Python, more machine learning and business metrics (5 months)

- Hadoop Platform and Application Framework (University of California, San Diego) (5 weeks) *high priority!*
  - Hadoop, Spark and MapReduce
- Introduction to Data Science in Python (Applied Data Science with Python, University of Michigan) (4 weeks) *high priority!*
- Applied Plotting, Charting & Data Representation in Python (Applied Data Science with Python, University of Michigan) (4 weeks) *high priority!*
- Applied Machine Learning in Python (Applied Data Science with Python, University of Michigan) (4 weeks) *high priority!*
- Business Metrics for Data-Driven Companies (Excel to MySQL: Analytic Techniques for Business, Duke) (4 weeks) *important!*

## Pro: more Python, more business metrics and fancy visualization tools (6 months)

- 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) *important!*
- Managing Big Data with MySQL (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)

## 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)