Data Analyst (Python) course

1. Overview

Outcome: at the end of this course, participants will know how to collect data from multiple sources, analyse it to find correlation and pattern, create reports and compelling visualisations.

Prerequisites: basic knowledge in Programming and Mathematics

Admission: Assessments (in Programming and Mathematics)

Preparation work: est. 40 hours online

Duration: 20 units (total 160H)

2. Objectives

A Data Analyst gathers data from multiple sources, analyses it to find correlation and pattern, and creates reports and compelling visualisations. This course is designed to learn skill & key competencies (as defined below) adequate to the Data Analyst position.

Target skills & competencies

- Comfortable using Python and SQL autonomously

- Comfortable analyzing, visualizing & communicating data with a notebook

- Comfortable creating dashboards and reports

- Comfortable using and collaborating with a version control system

- Advanced using descriptive statistics

- Comfortable using statistical modelling techniques (regression, clustering, classification)

- Knowledge of main database technologies & data collection/engineering methods

3. Syllabus

1. **Admission and Preparation - 40 hours online**

Pre-requisites and Assessment

Our Data Analyst course requires pre-requisite skills in Programming and basic concepts of Mathematics. After applying, we will communicate a technical assessment to each candidate and determine if they can enroll into the course.

Preparation Work
If they succeed in their technical assessment, students will then have to complete an online preparation work before starting the course. This work takes a maximum of 40 hours and covers the basics of Python, the pre-requisite language of the course, and some topics used every day by data analysts.

Concepts & Skills or Preparation Work

- Python Programming Basics

- Terminal

- Source control (Git, Github)

- Mathematics (statistics, probability)

2. Data Science Toolkit - 10 units

2.1. Python for Data Science - 5 units

Module Description

Learn programming in Python, how to work with Jupyter Notebook and to use powerful Python libraries like *Pandas* and *NumPy* to explore and analyze big data sets. Collect data from various sources, including CSV files, SQL queries on relational databases, data warehouses, APIs and Web scraping.

Concepts & Skills Learned

- Using a version control system
- Using Jupyter Notebook
- Loading and exploring a dataset
- Data collection methods from different sources
- *Pandas* and *NumPy*
- Web scraping

2.2. Relational Database & SQL - 2 units

Module Description

Learn how to formulate a good question and how to answer it by building the right SQL query. This module will cover schema architecture and then dive deep into the advanced manipulation of *SELECT* to extract useful information from a stand-alone database or using a SQL client software like DBeaver.

Concepts & Skills Learned

- Database schema architecture
- Translate a business question into a SQL query
- Advanced manipulations of *SELECT*
- SQL client software like DBeaver or Metabase

2.3. Data Visualization - 1 unit

Module Description

Make data analysis more visual and understandable by including data visualizations in a Notebook. Learn how to plot data frames using Python libraries such as *matplotlib* and *seaborn* and transform data into actionable insights.

Concepts & Skills Learned

- Turn data into insights with data visualizations
- Different categories of charts
- *matplotlib* and *seaborn*

2.4. Statistics, Probability, Linear Algebra - 2 units

Module Description

Understand the underlying math behind all the libraries and models used in the bootcamp. Become comfortable with the basic concepts of statistics & probabilities (mean, variance, random variable, Bayes's Theorem, etc.) and with matrix computation, at the core of numerical operations in libraries like Pandas and Numpy.

Concepts & Skills Learned

- Statistics (mean, variance, standard deviation, distribution, etc.)
- Probability (Bayes's Theorem)
- Matrix calculus

3. Decision Science - 5 units

Module Description

Time to solve a real-life problem: "as a data analyst working for a major e-commerce company (OLIST), how can I find interesting recommendations to improve our website's performance?".

Learn how to structure a Python repository with object-oriented programming in order to collaborate efficiently, how to survive the data preparation phase of a vast dataset provided, how to find and interpret meaningful statistical results quickly before making advanced predictions, and how to explain results to a non-technical audience thanks to cost/benefits analysis. This phase is organised in groups of 3-4 people.

Concepts & Skills Learned

- Starting from an open-ended problem
- Exploring a large dataset
- Preparing the data
- Statistical inference
- Multivariate Linear Regression
- Logistic Regression
- Presenting results with a convincing cost/benefits analysis

4. Capstone Data Analytics Project - 5 units

Module Description

Wrap up and reflect on the whole Data Analytics journey, applying what we've learned throughout the course to one of the company's datasets.

Bring results as each group must create a presentation including a summary of the challenge, a description of the data acquisition, cleaning, and parsing steps; a clear visualisation or dashboard that displays insights, numerically and graphically; a description of high-level conclusions and recommendations for the management. It will be a great way to practise all the tools, techniques and methodologies covered in the past and make everyone realise how autonomous they have become.

Concepts & Skills Learned

- Methodical and logical approach to a Data Analytics project
- Statistical modelling techniques
- Communicate findings