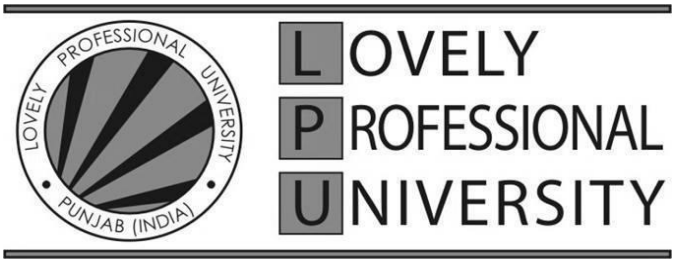


SIX WEEKS SUMMER TRAINING REPORT

on

Data Scientist with Python

Submitted by

Name :

Registration No :

Programme Name:

Under the Guidance of

DataCamp

School of Computer Science & Engineering
Lovely Professional University, Phagwara

(June-July, 2019)

DECLARATION

I hereby declare that I have completed my six weeks summer training at DataCamp from 2/06/2019 to 17/07/2019 under the guidance of DataCamp . I have declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfill the requirements of training for the award of degree of Bachelor of Technology in Computer Science, Lovely Professional University, Phagwara.

Date :

Rahul Rathore
11702112

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give towards our faculty of Computer Science, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my Summer Training especially in writing this report.

Furthermore I would also like to acknowledge with much appreciation the crucial role of the Mentors of the DataCamp, who gave the permission to use all required resources and the necessary to complete the Summer Training on their platform. Last but not least, many thanks goes to the Mentor allotted to me by the University who have invested his full effort in guiding me in achieving this goal. I have to appreciate the guidance given by my seniors as well as my parents that has improved my presentation skills thanks to their comment and advises.


#92,684
CERTIFICATE NUMBER
STATEMENT OF ACCOMPLISHMENT
HAS BEEN AWARDED TO
Rahul Rathore
FOR SUCCESSFULLY COMPLETING
Data Scientist with Python Track
DataCamp

Table of Contents

INTRODUCTION

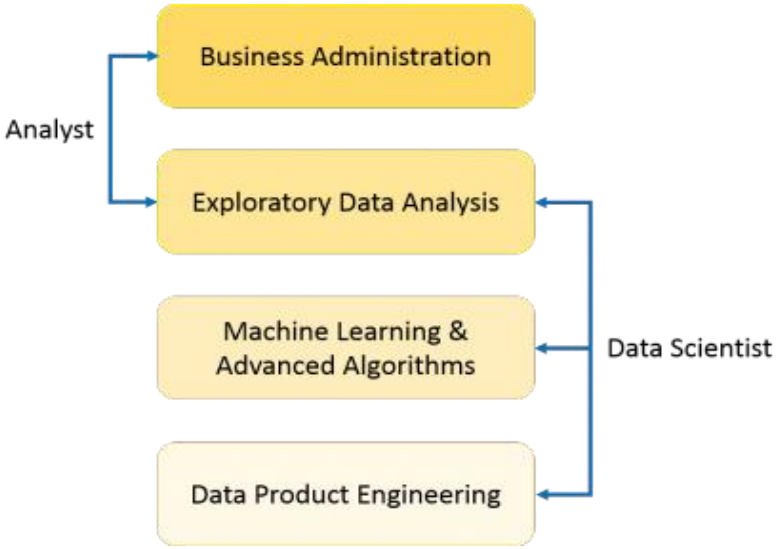
As the world entered the era of big data, the need for its storage also grew. It was the main challenge and concern for the enterprise industries until 2010. The main focus was on building framework and solutions to store data. Now when Hadoop and other frameworks have successfully solved the problem of storage, the focus has shifted to the processing of this data. Data Science is the secret sauce here. All the ideas which you see in Hollywood sci-fi movies can actually turn into reality by Data Science. Data Science is the future of Artificial Intelligence. Therefore, it is very important to understand what is Data Science and how can it add value to your business.

What is Data Science?

Use of the term Data Science is increasingly common, but what does it exactly mean? What skills do you need to become Data Scientist? What is the difference between BI and Data Science? How are decisions and predictions made in Data Science? These are some of the questions that will be answered further.

First, let's see what is Data Science. Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years?

The answer lies in the difference between explaining and predicting.



As you can see from the above image, a Data Analyst usually explains what is going on by processing history of the data. On the other hand, Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future. A Data Scientist will look at the data from many angles, sometimes angles not known earlier.

So, Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

I. Predictive causal analytics
II. Prescriptive analytics:
III. Machine learning for making predictions
IV. Machine learning for pattern discovery

Let's say you are working in a telephone company and you need to establish a network by putting towers in a region. Then, you can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength.

TECHNOLOGY LEARNT

I choose **Data Science with Python** for my summer training and completed it under the supervision of online platform called DataCamp.
In this course there were 26 different modules. The course started with python basics and through out the course I observed how a Data Scientist can use Python to manipulative, visualize the data and extract meaningful information from that data.

1. Introduction to python

Python is a general-purpose programming language that is becoming more and more popular for doing data science. Companies worldwide are using Python to harvest insights from their data and get a competitive edge. Unlike any other Python tutorial, this course focused on Python specifically for data science.

- ✓ Python Basics
- ✓ Python Lists
- ✓ Functions and Packages

2. Intermediate Python for Data Science

The intermediate python course was crucial to data science curriculum. I learnt to visualize real data with matplotlib's functions and get to know new data structures such as the dictionary and the Pandas DataFrame. After covering key concepts such as boolean logic, control flow and loops in Python.

- ✓ Matplotlib
- ✓ Dictionaries & Pandas
- ✓ Logic, Control Flow and Filtering
- ✓ Loops
- ✓ Case Study: Hacker Statistics

3. Python Data Science Toolbox (Part 1)

It's was the time to push forward and develop Python chops even further. There were lots and lots of fantastic functions in Python and its library ecosystem. However, a Data Scientist, constantly need to write own functions to solve problems that are

dictated by the data. The art of function writing is what I learnt in this first Python Data Science toolbox course.

- ✓ Writing Python functions
- ✓ Default arguments, variable-length arguments and scope
- ✓ Lambda functions and error-handling.

4. Python Data Science Toolbox (Part 2)

In this second course in the Python Data Science Toolbox, I continued to build Python Data Science skills. First I learnt the wonderful world of iterators, objects that I have already encountered in the context of for loops without having necessarily known it. Then I learnt about list comprehensions, which are extremely handy tools that form a basic component in the toolbox of all modern Data Scientists working in Python. I ended the course by working through a case study in which I applied all of the techniques I learned both in this course as well as the prequel.

- ✓ Using iterators in PythonLand
- ✓ List comprehensions and generators
- ✓ World Bank's World Development Indicators dataset!

5. Importing Data in Python (Part 1)

As Data Scientist, on a daily basis he will need to clean data, wrangle and munge it, visualize it, build predictive models and interpret these models. Before doing any of these, however, he will need to know how to get data into Python. In this course, I learnt the many ways to import data into Python.

- ✓ Introduction and flat files
- ✓ Importing data from other file types
- ✓ Working with relational databases in Python

6. Importing Data in Python (Part 2)

This second course on importing data helped me learn about other varoius file formats.

- ✓ Importing data from the Internet
- ✓ Interacting with APIs to import data from the web
- ✓ Diving deep into the Twitter API

7. Cleaning Data in Python

It is commonly said that data scientists spend 80% of their time cleaning and manipulating data, and only 20% of their time actually analyzing it. This course equipped me with all the skills I needed to clean data in Python, from learning how to diagnose data for problems to dealing with missing values and outliers.

- ✓ Exploring your data
- ✓ Tidying data for analysis
- ✓ Combining data for analysis
- ✓ Cleaning data for analysis
- ✓ Dataset obtained from the Gapminder Foundation

8. pandas Foundations

Pandas DataFrames are the most widely used in-memory representation of complex data collections within Python. Whether in finance, scientific fields, or data science, a familiarity with Pandas is essential. This course taught me to work with real-world data sets containing both string and numeric data, often structured around time series. I learnt powerful analysis, selection, and visualization techniques in this course.

- ✓ Data ingestion & inspection
- ✓ Exploratory data analysis
- ✓ Time series in pandas
- ✓ Case Study - Sunlight in Austin

9. Manipulating DataFrames with pandas

In this course, I learn how to leverage pandas' extremely powerful data manipulation engine to get the most out of data. It is important to be able to extract, filter, and transform data from DataFrames in order to drill into the data that really matters. The pandas library has many techniques that make this process efficient and intuitive. I learnt how to tidy, rearrange, and restructure your data by pivoting or melting and stacking or unstacking DataFrames. These are all fundamental next steps on the road to becoming a well-rounded Data Scientist, and I got chance to apply all the concepts I learnt to real-world datasets.

- ✓ Extracting and transforming data

- ✓ Advanced indexing
- ✓ Rearranging and reshaping data
- ✓ Grouping data
- ✓ Dataset of Summer Olympic games

10. Merging DataFrames with pandas

This course was all about the act of combining, or merging, DataFrames, an essential part of any working Data Scientist's toolbox. I improved my pandas skills by learning how to organize, reshape, and aggregate multiple data sets to answer specific questions.

- ✓ Preparing data
- ✓ Concatenating data
- ✓ Merging data
- ✓ Case Study - Summer Olympics

11. Analyzing Police Activity with pandas

This course given chance to apply that knowledge by answering interesting questions about a real data-set! I explored the Stanford Open Policing Project data-set and analyzed the impact of gender on police behavior. During the course, I gained more practice cleaning messy data, creating visualizations, combining and reshaping data sets, and manipulating time series data.

- ✓ Preparing the data for analysis
- ✓ Exploring the relationship between gender and policing
- ✓ Visual exploratory data analysis
- ✓ Analyzing the effect of weather on policing

12. Intro to SQL for Data Science

This course taught syntax of SQL shared by many types of databases, such as PostgreSQL, MySQL, SQL Server, and Oracle.This course taught me everything I need to know to begin working with databases today!

- ✓ Selecting columns
- ✓ Filtering rows
- ✓ Aggregate Functions
- ✓ Sorting, grouping and joins

13. Introduction to Relational Databases in SQL

I already used SQL for querying data from databases. In this course, I experienced this firsthand by working with a real-life data set that was used to investigate questionable affiliations of universities. Column by column, table by table, I got to unlock and admire the full potential of databases. In between, I learnt how to create tables and specify their relationships as well as how to enforce data integrity. Also, I discovered other unique features of database systems, such as constraints.

- ✓ Created first database
- ✓ Enforce data consistency with attribute constraints
- ✓ Uniquely identify records with key constraints
- ✓ Glue together tables with foreign keys

14. Introduction to Data Visualization with Python

This course extended Intermediate Python for Data Science to provide a stronger foundation in data visualization in Python. The course provided a broader coverage of the Matplotlib library and an overview of Seaborn (a package for statistical graphics). Topics covered include customizing graphics, plotting two-dimensional arrays (e.g., pseudocolor plots, contour plots, images, etc.), statistical graphics (e.g., visualizing distributions & regressions), and working with time series and image data.

- ✓ Customizing plots
- ✓ Plotting 2D arrays
- ✓ Statistical plots with Seaborn
- ✓ Analyzing time series and images

15. Interactive Data Visualization with Bokeh

Bokeh is an interactive data visualization library for Python (and other languages!) that targets modern web browsers for presentation. It can create versatile, data-driven graphics, and connect the full power of the entire Python data-science stack to rich, interactive visualizations.

- ✓ Basic plotting with Bokeh
- ✓ Layouts, Interactions, and Annotations

- ✓ Building interactive apps with Bokeh
- ✓ Putting It All Together! A Case Study

16. Statistical Thinking in Python (Part 1)

After all of the hard work of acquiring data and getting them into a form I can work with, I ultimately want to make clear, succinct conclusions from them. This crucial last step of a data analysis pipeline hinges on the principles of statistical inference. In this course, I started building the foundation I need to think statistically, to speak the language of data, to understand what they are telling me. The foundations of statistical thinking took decades upon decades to build, but they can be grasped much faster today with the help of computers.

- ✓ Graphical exploratory data analysis
- ✓ Quantitative exploratory data analysis
- ✓ Thinking probabilistically -- Discrete variables
- ✓ Thinking probabilistically -- Continuous variables

17. Statistical Thinking in Python (Part 2)

In this course, I did just that, expanding and honing hacker stats toolbox to perform the two key tasks in statistical inference, parameter estimation and hypothesis testing. I worked with real data sets as I learn, culminating with analysis of measurements of the beaks of the Darwin's famous finches. I emerged from this course with new knowledge and lots of practice under belt, ready to attack your own inference problems out in the world.

- ✓ Parameter estimation by optimization
- ✓ Bootstrap confidence intervals
- ✓ Introduction to hypothesis testing
- ✓ Hypothesis test examples
- ✓ Putting it all together: a case study

18. Joining Data in SQL

In this course I learnt all about the power of joining tables while exploring interesting features of countries and their cities throughout the world. I master inner and outer joins, as well as self-joins, semi-joins, anti-joins and cross joins - fundamental tools in any PostgreSQL wizard's toolbox.

- ✓ Introduction to joins
- ✓ Outer joins and cross joins
- ✓ Set theory clauses
- ✓ Subqueries

19. Introduction to Shell for Data Science

The Unix command line has survived and thrived for almost fifty years because it lets people do complex things with just a few keystrokes. Sometimes called "the universal glue of programming", it helps users combine existing programs in new ways, automate repetitive tasks, and run programs on clusters and clouds that may be halfway around the world. This course introduced its key elements and showed me how to use them efficiently.

- ✓ Manipulating files and directories
- ✓ Manipulating data
- ✓ Combining tools
- ✓ Batch processing
- ✓ Creating new tools

20. Conda Essentials

This course explained how to use its core features to manage your software so that me and my colleagues can reproduce working environments reliably with minimum effort.

- ✓ Installing Packages
- ✓ Utilizing Channels
- ✓ Working with Environments
- ✓ Case Study on Using Environments

21. Supervised Learning with scikit-learn

In this course, I learnt how to use Python to perform supervised learning, an essential component of Machine Learning. I learnt how to build predictive models, how to tune their parameters and how to tell how well they will perform on unseen data, all the while using real world datasets.

- ✓ Classification
- ✓ Regression

- ✓ Fine-tuning model
- ✓ Prepossessing and pipelines

22. Machine Learning with the Experts: School Budgets

Data science isn't just for predicting ad-clicks-it's also useful for social impact! This course is a case study from a machine learning competition on DrivenData. I explored a problem related to school district budgeting. By building a model to automatically classify items in a school's budget, it makes it easier and faster for schools to compare their spending with other schools. In this course, I begun by building a baseline model that is a simple, first-pass approach.Finally, I saw how the winner was able to combine a number of expert techniques to build the most accurate model.

- ✓ Exploring the raw data
- ✓ Creating a simple first model
- ✓ Improving the model
- ✓ Learning from the experts

23. Unsupervised Learning in Python

In this course, I learnt the fundamentals of unsupervised learning and implement the essential algorithms using scikit-learn and scipy. I learnt how to cluster, transform, visualize, and extract insights from unlabeled datasets, and end the course by building a recommender system to recommend popular musical artists.

- ✓ Clustering for dataset exploration
- ✓ Visualization with hierarchical clustering and t-SNE
- ✓ Decorrelating your data and dimension reduction
- ✓ Discovering interpretable features

24. Machine Learning with Tree-Based Models in Python

In this course, I learnt how to use Python to train decision trees and tree-based models with the user-friendly scikit-learn machine learning library. I understood the advantages and shortcomings of trees and demonstrate how ensembling can alleviate these shortcomings, all while practicing on real-world datasets. Finally, I also understood how to tune the most influential hyperparameters in order to get the most out of models.

- ✓ Classification and Regression Trees

- ✓ The Bias-Variance Trade-off
- ✓ Bagging and Random Forests
- ✓ Boosting
- ✓ Model Tuning

25. Deep Learning in Python

In this course, I gained hands-on, practical knowledge of how to use deep learning with Keras 2.0, the latest version of a cutting edge library for deep learning in Python.

- ✓ Basics of deep learning and neural networks
- ✓ Optimizing a neural network with backward propagation
- ✓ Building deep learning models with keras
- ✓ Fine-tuning keras models

26. Network Analysis in Python

This course will equipped me with the skills to analyze, visualize, and make sense of networks. I applied the concepts you learn to real-world network data using the powerful NetworkX library. With the knowledge gained in this course, I developed network thinking skills and be able to start looking at your data with a fresh perspective!

- ✓ Introduction to networks
- ✓ Important nodes
- ✓ Structures
- ✓ Bringing it all together

REASON FOR CHOOSING THIS TECHNOLOGY

1. **The increasing demand for Data Scientist:** The data scientist job is creating publicity all around the world through its demand. According to a report given by McKinsey and Company, by the end of 2020, there will be around 140,000 to 180,000 data scientist who is less than the required. The demand for a data scientist is increasing, but the supply is very less. As compared to engineers and chartered accountants India needs more than 200,000 data scientist by 2020. Thus, the candidate who wants to make their career in data science should not wait and start their career in data science, as it is a very demanding field.

2. **High ranging salary:** As per a report from the Glassdoor, in 2016 data science was the highest paid field to start a career. According to their findings, the national average salary for a Data scientist is around INR 6, 50,000 in India and the national average salary is $1, 20,931 in the United States coming to Europe. As per another report by O'Reilly a survey carried out for data science salary, the annual base salary of U.S. based survey defendants were $104,000. Robert Half's tech guide represents the salary between $109,000 and $153,750. Another review carried out by the Burtch Work for data science shows the median base salary ranges from $97,000 for Level 1 contributors for $152,000 for Level 3 contributors. Additionally, the median bonuses start at $10,000 for level 1 contributor. Thus, as compared to other jobs, these **salaries are much higher.**

3. **Data Science is an emerging field:** Due to the growing demand of data all around the world data, data science is evolving rapidly. Data scientists have a wide assortment of ranges of skills that can use the information and data to assist associations in making better business decisions. They can get inspired energizing chances to work and explore different avenues regarding data to create the appropriate answers for the organizations. In a data science field, there are numerous new exciting fields are also evolving involving Big Data, Artificial Intelligence, Machine Learning along with some recent technologies such as Blockchain, edge computing, Serverless Computing, Digital Twins, and others which engage various practices and approaches within the Data Science industry.

4. **Data Scientists give importance to the Business:** Data scientists are flourishing in almost every field of businesses ranging from IT to healthcare, from E-commerce to marketing and retail. As data is the critical asset of any company, data scientists play an essential role they serve as a trusted adviser and strategic partner to their management. They are responsible for examining the data for a treasured resource which support to enhance their niche, recognize the preferred target viewers and handle future marketing and development policies.
5. **Easy to take a job:** Data science is a booming field, and it is the most demanding job of 2019. Most of the companies are greatly looking for data scientists because the demand for a data scientist is high and the supply is very less. Not only E-commerce companies are recruiting data scientist, but, the companies from almost every field are hiring data scientists. **Many start-up companies are depending upon data science to proceed further.**

Apart from these reasons, there are other reasons as well, which compel candidates to select data science is a career option:

1. **Experience Factor:** Experience is likely a standout amongst the most well-known words found in a job description, and to be perfectly honest, organizations, for the most part, need representatives with a huge amount of it. However, data science is data science is such a moderately new field. According to Burtch Work's reports, 40% of the data scientists have around five years of experience, and 69% people have less than ten years of experience. As it is mentioned already, salaries to coordinate the wages with the experience levels. Level one individual contributor, typically has 0-3 years of experience. Level 2 contributors will have 4 to 8 years of experience, and level 3 candidates will have more than nine years of experience.

2. **The range of undergraduate majors:** Since data science is an emerging field, numerous colleges are included data science as an undergraduate degree program. Meanwhile, data scientists hail from a grouping of scholastic foundations, involving arithmetic/measurements, software engineering, building, and natural science. Additionally, a few data scientist have degrees in financial matters, sociology, business, and even medical science.

3. **Work opportunities:** After becoming a data scientist, you can work anywhere your heart wants. While 43% of these experts deal with the West coast, and 28% are in the Northeast, they are being utilized in almost every region in the country and abroad. Well, the salary will be highest in the U.S. are on the West Coast. The innovation business utilizes the most data scientist; however, they additionally work in different enterprises extending from medicinal services/pharma to promoting and monetary administrations to counseling firms to retail and CPG ventures. Data scientists even work for gaming projects and 1% work for the administration.

4. **Lack of competition and ease of job hunting:** As data science domain is a comparatively new field, not only there is a shortage of data scientist. Also, there is a lack of competition in this field. A beginner data scientist, and an expert level will have an experience gap of decidedly less years. Thus, this increases the opportunity for career growth. The demand for data scientists is very high, and there is a scarcity of experts in the market. Therefore, it is relatively easy to search for jobs in data science field.

Data science is the field which is not just helping organizations to perceive their markets and then making better business decisions, but it is also assisting organizations to draw closer to their customers to bring them effective services. Data scientists are the superheroes, who gather, clean and organize the data with their excellent aptitude. Aspirants who want to make their career as a Data Scientist have to make a lot of personal promise to work hard and deliver outcomes that matters a lot.

LEARNING OUTCOMES

Through this course , I learned the following things :

1. The basics of data analysis in Python. Expanded my skill set by learning scientific computing with numpy , i.e. using large data sets and working upon them to get something out of it.

2. Expanded my data science skills by creating visualization using matplotlib and manipulating data frames with Pandas in which we can visualize large datasets and represent them in pictorial manner to get a clearer image of it.

3. To write my own functions in Python, as well as key concepts like scoping and error handling, Iterators and List comprehensions which could come in very handy in future as creating our own functions helps us do our tasks more efficiently and conveniently.

4. To import data into python from various sources, such as excel, SQL, SAS and right from the web.This could help me in various ways like working on data sets present in various other types of files like .txt , .csv, relational databases, excel spreadsheets etc.

5. Improved python data importing skills and learned to work with web and API data as we know that web is a rich source of data and we can extract various types of insights and findings from that.

6. To clean the data in python and how to use industry standard pandas library to import, build and manipulate Data Frames.As sometimes to deal with unstructured data, we have to perform various string matching ,string manipulation techniques to deal with missing or duplicate data.

7. To tidy, rearrange, abd restructure the data using versatile DataFrames by performing various kind of operations on them like indexing, slicing ,filtering etc.

8. I explored the Stanford open Project data-set and analyzed the impact of gender on police behavior using pandas and performing various kinds of operations on it.

9. I mastered the basics of SQL. Querying tables in relational databases such as MySQL, Oracle, SQL Server, and PostgreSQL.Here in this, I learned applying various functions on relational datasets and relational data bases

10. To create efficient databases and ways of storing data.Through this we can create multiple data sets and work accordingly on them.

11. Handling more complex data visualization techniques using Matplotlib and seaborn.

12. Creating versatile and interactive data visualization using Bokeh which can help me creating dynamic visuals of the datasets.

13. The concepts of Statistical thinking and how to perform two key tasks in statistical inference : parameter estimation and hypothesis testing.

14. How to join two, three tables together into one, combine using set theory, and work with subqueries in PostgreSQL.

15. The basic UNIX commands line which helps to combine existing programs in new ways, automate repetitive tasks, and run programs.

16. The basics of Conda and how can I manage software using Conda and it's functions.

17. How to build and tune predictive models and evaluate how well they perform on unseen data.

18. How to build a model to automatically classify items in a school budget, and how to cluster, transform, visualize, and extract insights from unlabeled datasets using scikit-learn and scipy.

19. I learnt how to use tree-based models and ensembles for regression and classification using scikit-learn.

20. To Analyze, visualize, and make sense of networks using the NetworkX library.Through this , I came to learn about how a Data Scientist deals with network and its property.

Bibliography

- ✓ DataCamp
- ✓ Quora
- ✓ Glassdoor
- ✓ Wikipedia
- ✓ Bible for DataScientist
- ✓ Python for Beginners
- ✓ SQL with Python