Data Science Workflow for Beginners

START YOUR DATA SCIENCE JOURNEY INTO A SUCCESSFUL HIGH PAYING CAREER

26 Datasets TO START RIGHT AWAY!

A. J. García

Data Science Workflow for Beginners

Start your Data Science Journey into a Successful High Paying Career

Alejandro Garcia

This book is for sale at http://leanpub.com/data-science-workflow

This version was published on 2020-08-12

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Contents

Introduction

Data Science is a process which can tell wonderful stories hidden behind a bunch of numbers and information, it can be a painless process if each step in the workflow is understood and completed in the proper way.

It's an evolving field with countless possibilities since data is created at unparalleled speed these days, for a person starting in this field there are new and exciting concepts to understand and many frameworks, tools and libraries at his or her disposal to play with.

There's a lot of room for technical (programmers) and non technical people to put on different hats and collaborate together building applications and models to solve real life problems. There will be people leaning more towards any of these areas: databases, statistical analysis, model development or simply insights and data visualization.

You will soon discover **The Data Science Workflow** and later in the book I will introduce you to 26 excellent datasets which are used by the most successful data scientists to build models (classification/prediction) and to create rich data visualizations.

Once you are familiar with the whole data science workflow, it will be clear that you should start every project formulating a set of questions for which you want to find answers from the data, you will learn the different steps that needs to be completed for the whole process to end successfully with the visualization and communication of the recently discovered insights.

This is a process which sometimes requires collaboration with technical minds AKA "programmers/developers" to help you build autonomous learning models through Machine Learning and Artificial Intelligence, this opens the door to incredible collaboration opportunities.

Finally the book will take you through The Data Science Workflow while guiding you along 3 simple data visualization projects to dip your toes in the water.

Let's jump right into it

Who is this book for ?

This book is for the curious mind, it's **for a beginner who wants to discover everything it takes to work on a daily basis with data**. It is for someone who wants to get meaning and insights from large datasets. It doesn't matter if you come from a coding background or not.

If you have some questions about data science, and you would like to learn more this book will be a good introduction for you.

It also **provides invaluable data sets** for you to start playing right away with some data and get some initial feeling of the whole data science process from collection to visualization.

What this book covers ?

This book introduces you to Data Science, getting to know Data Science will let you:

1. Develop a career in an area where you can land **high paying jobs between 150k to 220k**.
2. Specialize yourself as a Big Data Engineer, Database Manager, Data Modeler, Data Scientist or Data Analyst.

This book is divided in 6 chapters: Along the chapters the book will introduce:

- Data Science.
- Machine Learning.
- Algorithms.
- Data sets.
- Data Visualization.

Chapter 1 provides an overview of Data Science identifying some of the key roles and the recommended skills to develop as a data scientist.

Chapter 2 takes you on a journey to discover from the beginning to the end the whole data science process.

Chapter 3 explore all the relations that exist between artificial intelligence and machine learning and how they are used by data scientists.

Chapter 4 put at your disposal 26 public datasets you can start using today to build successful data science projects.

Chapter 5 gives you 10 extra ordinary sites for you to dig deep and find professional datasets you can use and download for Free.

Chapter 6 lets you go through the whole data science process with 3 simple data visualization projects.

Chapter 1: Introduction to Data Science

What is Data Science ?

Data science is a field of study with mainly two key players, the data scientist and the machine learning engineer. Data science is a global field which looks to get answers to some problems from data. The problems might be business related or not.

The creation of Data happens every second of every hour of every single day, and as you can imagine only humans are generating billions of events each second which can be considered as data points, these data points can build enormous data conglomerates and many of them are collected by the tech industry.

There are many companies dedicated not only to the collection of data, but also to the analysis being the end goal of some of these companies to profit from data.

This is why the Data Science field has evolved so quickly and will continue offering plenty of opportunities for every one. The vast amount of data means data scientists can discover patterns which nobody else thought could exist, and these insights translate into business opportunities.

Some examples of the applications of this field are: data visualization, user behaviour prediction, recommendation engines, classification.

So, to come back to the definition of data science we can think of it as the discipline dealing with the collection of data, the cleaning and preparation of this data for analysis, the usage of artificial intelligence on the data to be able to get some insights and finally using some visualization techniques to help present and communicate the findings.

Data science is all about revealing insights from data so that predictions and business decisions can be made.

Who is a data scientist ?

As long as data keeps growing companies will have needs to hire people who can dedicate their efforts towards big data. These Data scientists are individuals who have a particular set of skills.

If you want to be the ultimate data scientist, then you should aim to learn in which areas of data science you are good and which ones you could develop a bit more. All data scientists should have some of the following skills:

- Data Collection.
- Data Extraction.
- Data Processing and cleaning.
- Statistics.
- Data Mining.
- Understanding of Artificial Intelligence.
- Machine learning as a subset of AI.
- Programming Skills.
- Database Skills.
- Dashboards and Data Visualization.

Data collection & extraction:

Data might come from different sources, data scientists need to know how to work with different file formats so they can receive the data and have it ready to begin the process. It's useful to also know about databases SQL and NoSQL since many times as a data scientist you would have to make queries in order to extract information from tables in databases.

Sometimes the information you are looking for would come from an API request, so basic understanding of APIs and formats like Json is also recommended.

Data Processing:

As soon as data is collected a good data scientist should take a look and try to identify errors in the data, the type of data he or she is dealing with, the format and getting a feeling of whether the data makes sense or not.

Remember data needs to be processed in a way that is error free and can be used in the next steps to identify trends and be able to get predictions or useful insights from it.

Statistics: Some key concepts you should be familiar with are probability, probability distributions, statistical significance, regression, statistical features, sampling and bayesian statistics.

Data Mining:

Some people define data mining as a process where the data scientist would try to connect the variables in the data to find new meanings and extract useful information.

Probably more an hypothesis based process related to reasoning, thinking and logical soft skills.

Understanding of Artificial Intelligence:

One of the simplest AI definitions might be the use of algorithms to perform certain actions which are time consuming for humans, so we use computers to transform data, learn from it and understand it in a way that allows us to find patterns and generate new insights.

Machine Learning:

ML is a subset of AI and data scientists need to know the basics, here is where they start working with ML engineers to implement the AI algorithms that would allow the processing and classification of massive data sets in order to make predictions out of the data.

Programming Skills:

Python or R seem to be the main programming languages used for data science, they are important because of the great libraries available so the ML engineers can build what is called a “Model”.

The model simply refers to the collection of code and algorithms which will take data from the data scientist to learn from it, so the model can be trained and in a subsequent step work with new data classifying it, making predictions or generating new insights.

Database Skills:

As mentioned before usually data is stored in the cloud or hosted on servers, in order to extract the right data, queries need to be built, these queries are instructions for the database engine to return the specified data.

An understanding of how the data is structured on the database, tables and rows, relations between data and query commands is necessary.

Dashboards and Visualization:

There are many tools which can help with this final step, ranging from Excel to Power BI, Javascript or Python libraries. The goal is to be able to present the data in a clear and easy to understand way so the insights can be understood thus allowing the business to make its decisions.

Structured Vs Unstructured data

Data can be of different types, the most common ones being text and numbers. You could also see machine learning algorithms that work on images and even sounds.

Structured Data:
Usually data collected from databases or even excel files and csv files would be structured since every column would be labeled and the data will be organized in that way.

Unstructured Data:
Unstructured data is the one not following a particular structure, think of it for example text information you are trying to extract from tweets, blog posts, emails or product reviews.

This category is relevant for those data scientists working with *Natural Language Processing* and that kind of application.

Chapter 2: The Workflow of Data Science

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Chapter 3: Data Science and Machine Learning

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

What is Machine Learning ?

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Relation between Data Science and Machine Learning

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Machine learning model

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Supervised learning, unsupervised learning and deep learning

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Algorithms used in Machine Learning

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Chapter 4: 26 Datasets to Build Successful Data Science Projects

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

World Development Indicators

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Education Statistics

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

World Bank Projects & Operations

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Airline Safety

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Weather in the US

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Adderall Drug Study

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Government Surveillance Planes Analysis

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Zika Virus Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

FBI Firearm Background Check Data

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Political Ads on Facebook

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Hate News Story Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Voting Machines in 2016 Election

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

n-grams from Google Books

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Common Crawl Corpus

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Landsat 8 Images

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Annual Survey of School System Finances

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

World Bank Open Data

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

International Monetary Fund Data

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Financial Times Market Data

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Google’s Open Images Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Stanford Dogs Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Indoor Scene Recognition Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Large Movie Review Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Berkeley DeepDrive Dataset

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Peking University/Baidu - Autonomous Driving

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Learning Analytics Data Set

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Chapter 5: 10 Amazing sites with Free Datasets

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Chapter 6: Data Visualization Projects

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Keep exploring and learning about Data Science

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.

Source Code Download

This content is not available in the sample book. The book can be purchased on Leanpub at http://leanpub.com/data-science-workflow.