A SHORT COURSE IN DATA SCIENCE USING PYTHON

DR JUAN H KLOPPER

SCHOOL FOR DATA SCIENCE AND COMPUTATIONAL THINKING AT STELLENBOSCH UNIVERSITY

Our world has undergone tremendous changes in the last few decades. Knowledge has become the cornerstone of our civilisation. We have embraced the use of data to gain that knowledge. This has been brought about by our ability to generate and capture vast amounts of data. Together with an explosion of data has come the ease of access to data and the ease of extracting knowledge from data.

Extracting knowledge from data was traditionally the task of statisticians and data analysts. Formal training in statistics was not for the masses. As data became abundant in so many fields, domain experts in these fields needed to learn how to analyse data. Their unique perspective, knowledge, and experience are invaluable in making sense of all the data in their fields.

Modern computers and software have led the democratisation of the analyses, understanding data, and use of data. Expensive, closed-source software from large corporation only available to the elite have made way for free and open-source computer languages such a Python, R, and Julia. Any domain expert or interested party can now use data to contribute to the fundamental understanding of our world, solving problems today that would otherwise have taken decades more to solve.

Data Science is the umbrella term for this ability to gather, manipulate, analyse, visualise, and learn from data. There has never been a more exciting time.

This course has one aim. To explain the essence of Data Science using the most popular and powerful tools available today. The journey is full of surprises and moments of enlightenment as you join the massive and ever-growing community of Data Scientist. The skills that you will become aware of during this course will open a new world.

Our vehicle will be the most popular language in Data Science. Python is an easy to learn, yet extremely powerful computer language. We will use cloud-based computing negating the need to install any software one your own computer.

The course has been created to jump-start your Data Science skills. As such, it is very dense with information. I want you have the best and most complete start to your new abilities. One week is not enough to learn all these new skills. As with learning a new spoken language, you will need time

and experience well beyond just this week. I do, however, want to leave you with a clear path forward. A few toy examples will not satisfy you. Instead, this course aims to highlight everything that is possible. I do not want to leave you wanting. I want you to become an expert Data Scientist in your field.

There are several educational resources available for this course. First and foremost are a set of detailed video tutorial that serves as your first contact with the course. The video lectures make use of extensive notes and code. These are provided as reference documents in portable document format. There are also sets of exercise materials that you complete before each day's life session. During these sessions we work through the exercise material after which you will also be provided with a complete answer set.

The School for Data Science and Computational Thinking wants to be your partner in the future, and we hope that you stay in touch after the course.

The course comprises 14 sections, which are described below.

01 INTRODUCTION TO DATA SCIENCE

SECTION	DECSRIPTION
Defining Data Science	Data Science is the amalgamation of tools from statistics, mathematics, and computer science that provide us with the ability to learn from data in order to understand and improve our world.
The Tools of Data Science	Computer languages have revolutionised our ability to gather and analyse data. Python has emerged as the leading language in the field of Data Science. It is easy to learn, free of charge, and a very large community of researchers and users have grown an ecosystem for Python. It is possible to create games and applications using Python. Its power to work with, analyse, visualise, and interpret data is at the core of its success.
Example Data Science Project	There is a lot to cover in a course on Data Science using Python. This short project serves as a demonstration of the power and ease of use the language.

02 DATA AND DEFINITIONS

SECTION	DECSRIPTION
Data Types	There are classifications systems for the type of data that we collect. The most emphatic divides data into a numerical and a categorical type. There are also subtypes for two main types. We learn about these subtypes with examples.
Sample Space	Age in years for humans are typically 0 to 100 years. A laboratory test can only be low, normal, or high with respect

	to a reference range of normal values. Values for a variable have a range or set. When we collect data for a variable from a sample of subjects, each value comes from this range or set of elements.
Tidy Data	When data is captured in the form of a spreadsheet or even when it is extracted from a database, it is ideal to have each row represent a subject and each column a well-defined data type. This long-format of data is emphasised and used in Data Science.
Research Questions	Computational thinking and research in general require clarity in the expression of a research question. This clarity enables us to convert our curiosity into data that we can examine in a structured way using a computer language.
Trials Experiments Outcomes	There are many confusing and overlapping terms in Data Science. This section clarifies some of the commonly used terms.
Populations and Samples	In most cases we cannot collect data from all members of a population. Instead, we take a random sample from the population. The results are then inferred back to the population.
Randomisation	In order to be able to infer the results of analysis of the data from a sample, we must select subjects without bias. Randomisation helps to ensure this selection. There are numerous ways to randomise the selection of subjects from a population. The common methods are discussed.

03 PYTHON

SECTION	DECSRIPTION
Computer Languages	There are many computer languages with most geared towards a specific task. Python is a general-purpose language. It easy of use and clarity in its syntax has made it the leading language in Data Science.
Tools	In order to use a computer language, we need to have a program in which we type the actual code. Many such tools are available. Here we introduce Google Colaboratory. It is part of Google Drive and available to anyone with a Google email account. It is as simple to use as Google Docs and is free.
Arithmetic	Python is easiest to introduce using simple arithmetic. Python can serve as a giant calculator.
Conditionals	Conditionals test a question. Is 2 greater than 4? A simple concept but of extreme importance in Data Science where we extract and manipulate data based on these questions. A conditional can only return a True or a False value. We use this include and exclude data from our analysis.
Functions	A computer language such as Python contain many keywords. They make up the syntax of a language. Many keywords are functions. They take input value that we

	provide and return a useful value after taking an appropriate action.
Python Data Types	Objects in Python are of a certain type. Numbers might be integers (whole numbers) or decimal values. We can group numbers together into list objects. Here we explore the basic data types in Python.
Math Package	Packages are code that we import into a running session of Python. These packages contain extra functions and functionality that expand the use of Python. We use the Math package to expand on the mathematical operations that we can perform in Python.
Computer Variables and Assignment	Objects in Python can be stored in computer memory for reuse. This is done by providing an appropriate name for an object and then assigning the object to that name. The name, referred to as a computer variable, signifies the area in memory that contain the object. In Python, as in most other languages, the equal symbol serves as the assignment operator.
Collection	Numbers, words, and other objects can be combined into collection. The three main collection types in Python are lists, tuples, and dictionaries. Here we explore examples of each and learn when and how to use each.
Loops	There are various loop operators in Python. They allow us to control the flow of execution of our code. This is very useful when we need to iterate over many instances of our analysis while certain conditions are met.
If Elif Else	These keywords are used in conjunction with loops to control the flow of execution of our code and analysis.
List Comprehension	List comprehension is a useful and fast way to generate data based on calculations. It can save a lot of code writing and time.
Numpy	One of the fundamental packages in Python is Numerical Python or numpy for short. It adds a host of functions and functionality to Python that are geared towards the analysis of data.

04 IMPORTING AND MANIPULATING TABULAR DATA

SECTION	DESCRIPTION
Pandas	The pandas package is one of the main reasons for the success of Python in Data Science. It allows us to create, import, manipulate, analyse, and plot data.
Importing Data	It is most common to have data saved as a spreadsheet file. The best format in Data Science is the comma separated values (CSV) file. Microsoft Excel, Google Sheets, and database application can export data as CSV files. They strip away all the extraneous additions that these applications add to data. These include formatting and colouring. We only need the actual values when analysing data. Pandas makes it easy to import data files.

Extracting Data	In many cases we only require a subset of our data for analysis. Here we learn how to use pandas to manipulate and extract our data.
Filtering Data	We can narrow our search by filtering out any unnecessary information.
Updating and Changing Data	Pandas allows for updating of data values and the addition and removal of data. These are required tasks as we explore the information contained in data.
Sorting Data	Many tests require sorting of data whether it be alphabetical or numerical. It is also a useful task when visualising data.
Missing Data	It can be rare to find a data set that contains values for all subjects and variables. Dealing with missing data has a direct impact on data analyses.
Dates and Times	There are many date and time formats. Computer hardware and software applications can be set to default formats, often in competing formats on the same computer. Data for dates and times can also be entered in many formats. Dealing with dates and times in a data set is challenging. Here we learn how to use Pandas to standardise formats for analyses.

05 SUMMARISING DATA

SECTION	DESCRIPTION
Counting	Frequency is the number of times that a sample space from a variable appears in a data set. We can also divide by the sample size to give is a relative frequency. Counting the often a very useful summary of data
Measure of Central Tendency	It is not possible to stare at rows and rows of numbers and learn anything from the exercise. Instead, we calculate single values that are representative of the whole. These values include the mean, the median, and the mode.
Measures of Dispersion	We also gain knowledge from a set of numbers if we know how spread they are. These measures include ranges, variances, standard deviations, and percentiles.

06 DATA VISUALISATION

SECTION	DESCRIPTION
Python Data Visualisation Ecosystem	Python has many packages that allow us to visualise data. This brings an even richer understanding of the knowledge hidden in the data. The plotly library generates interactive plots, ideal for Data Science.
Bar Plots	Bar plots are visual representations of frequency and relative frequency. They are the preferred visual representation of categorical or discrete data.
Histograms	Histograms similarly visualise frequencies and relative frequencies. Unlike bar charts, the are used for continuous numerical data.

Box-Whisker Plots	Box-and-whisker charts give an indication of the distribution of numerical data values by incorporating the median and quartiles of the data.
Scatter Plots	Scatter plots compare pairs of numerical variables for each subject. They allow us to visualise correlation between numerical variables and can help in visualising linear models such as linear regression models.
Time Series Plots	Time series plots allow is to visualise change in a variable over time.

07 RANDOMNESS AND SAMPLING

SECTION	DESCRIPTION
Randomness	The value for a variable in a subject can be viewed as random. The numpy package allows us to explore the topics and types of randomness that are essential in Data Science.
Probabilities	The likelihood or probability of a given value or statistic forms the core of expressing results in Data Science and in statistics.
Random Variables	A random variable is a function that assigns a value (that we can capture in a spreadsheet) to a random outcome. By understanding randomness and random variables and probabilities we can make sense of the knowledge in data.
Distributions	There are patterns to random variables when we consider their frequency in a data set. These patterns are termed distributions. Sampling distributions form the core of many analyses.

08 HYPOTHESIS TESTING

SECTION	DESCRIPTION
Sampling based on Proportions	Given a set of data for a categorical variable we determine how likely it is to have found the frequency of each sample space element.
Differences in Means	Hear we look at how likely it is to find the difference in means for a variable given two groups.
Hypothesis Testing	Hypothesis testing is the bedrock of the scientific method. We start with a research question stated in such a way that it is amenable to the collection of data of specific types that can be analysed to provide an answer to our question
Stating Hypotheses	Here we learn to state the two hypothesis that make up the method of hypothesis testing: the null hypothesis and alternative hypothesis.
Estimating Differences	Here, we look at an example problem to illustrate hypothesis testing.
One-tailed Hypothesis	While the two-tailed alternative hypothesis is used in most cases, we also learn about one-tailed hypothesis.

09 COMPARISONS FOR A NUMERICAL VARIABLE

SECTION	DESCRIPTION
Simulating the Difference in Means for a Numerical Variable Between Two Groups	Under the null hypothesis there is no difference in the means for a variable between two groups (a test statistic). Given a set of data values for a variable in two groups, we build a sampling distribution of the test statistic under the null hypothesis to understand the likelihood of the original test statistic
Comparing with Student's *t* Test	In this section we compare the results from our sampling distribution to a parametric *t* test.
Unequal variances	In this section we explore the effect of a difference in variance in our variable and see an alternative test for this situation.

10 UNCERTAINTY

SECTION	DESCRIPTION
Inference and Uncertainty	We aim to infer our results onto the population. To do so, we need to calculate possible values for a parameter given a statistic. This expresses the uncertainty in our results.
Bootstrapping	Bootstrap resampling is a technique that resamples from our sample data with replacement and allows us to build a distribution of possible values.
Confidence Levels and Intervals	The uncertainty in our results is expressed as lower and upper limits for the true population parameter given our statistic. We set a level for these limits, which is typically at 95%.

11 LINEAR MODELING

SECTION	DESCRIPTION
Correlation	Correlation is a measure of change in one numerical variable given a change in another numerical variable. Correlation is visualised using a scatter plot, where each marker shows the value of each of the two variables for each subject.
The *F* Distribution	The F distribution is one of the most important and useful sampling distributions in data analysis. We can use it to consider the difference in means between more than two groups amongst many other tests.
Simple linear regression	We can create a model that will calculate a value for a numerical variable given another numerical variable.
Multivariable linear regression	In this section we use more than one numerical variable to serve as independent variable in our model.
Analysis of variance	ANOVA is a simple technique that is used in this section to compare the means between more than two groups.

12 MACHINE LEARNING

SECTION	DESCRIPTION
Defining Machine Learning	Machine learning is a group term for many mathematically based algorithms that aim to learn from data. This learning can be used to predict an outcome or to generate new data.

Types	Machine learning can be classified into supervised and nonsupervised types. Each of these can be either classification or regression problems. There are also other algorithms such as reinforcement learning that is used by machine to play games and perform tasks.
Tools for machine learning	Python has the most used machine learning packages including scikit-learn and TensorFlow. The latter is a deep neural network architecture that I teach in a separate course.
Design Matrices	Linear regression can be classified as a very simple machine learning algorithm. In this section we learn about the patsy package. It converts data for easy use in linear models.
Statsmodels and scikit-learn	This section introduces these two important modelling packages in Python.

13 k NEAREST NEIGHBOURS ALGORITHMS

SECTION	DECSRIPTION
Distance	The k nearest neighbour algorithm uses the concept of distance between data point.
***k* Nearest Neighbour Classifier**	In this section we investigate an example of using the *k* nearest neighbour algorithm to predict the value of a categorical variable.
***k* Nearest Neighbour Regressor**	In this section, we predict a continuous numerical variable.

14 RANDOM FOREST ALGORITHMS

SECTION	DESCRIPTION
Decision trees	Ensemble machine learning techniques such as random forests make use of decision tree building blocks. Decision trees are relatable flow charts.
Random Forests	Random forests combine hundreds of trees to improve our models.
TensorFlow Decision Forests	At the cutting edge of available models in machine learning is Google's Decision Forests. In this section we use it in an example.