

**STK353: THE SCIENCE OF DATA ANALYTICS  
STUDY GUIDE**



## 1. INTRODUCTION

Welcome to the science of data analytics. Currently, the terminology dominating this exciting research field are data science, machine learning and artificial intelligence. The collective field of data analytics is vibrant and dynamic. This course will teach you to do Data Science in Python: You'll learn how to get your data into Python, get it into the most useful structure, transform it, visualise it and model it. This course will teach you how to use a computer as a lab, just as a chemist learns how to clean test tubes and stock a lab. Skills like cleaning data, choosing (filtering) data and making plots allow data science to happen. In this course you will find the best practices for doing each of these things with Python. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. The objective of this course is to empower you as a data scientist to put all the theory you have mastered in the past two and a half years in practice.

Many, many resources on Data Science with Python and R exist. In this course, we will use the textbook *The Data Science Handbook* by Field Cady.

## 2. ADMINISTRATIVE MATTERS

**2.1. Course changes and suspensions.** Consult the University calendar in connection with last days of course changes. An official form at the Administration can be used. Course suspensions must be official. A form is also available at Faculty Administration. The lecturers concerned must know about the suspension.

**2.2. Consultation.** We assume an online teaching model for this module in 2021. For this module, the following modes of consultation will be available to support students:

**2.2.1. Discussion boards.** Discussion board forums will be created for each teaching theme and other appropriate topics. Check out the discussion forum for answers before creating a new question.

**2.2.2. Collaborate sessions.** The practical session on Mondays consists of 6 collaborate sessions, starting on the half hour. A tutor will be available at each of these sessions to answer any questions related to the practical - whether it relates to an exercise or assignment.

**2.2.3. Individual consultation with tutors.** Tutors post their consultation timetables on Clickup. Students can make appointments with tutors prior to the consultation time. The consultation will take place on Google meet - the tutor will send a meeting link. Furthermore, the student must fill in, and email a consultation template to the tutor in order for the tutor to prepare for the meeting.


**2.2.4. Individual consultation with lecturers.** Lecturers post their consultation timetables on Clickup. Students can make appointments with lecturers prior to the consultation time of choice. The consultation will take place on Google meet - the lecturer will send a meeting link. Furthermore, the student must fill in, and email a consultation template to the lecturer in order for the lecturer to prepare for the meeting.

**2.3. Dishonesty and Plagiarism.** The dishonest missing of a test as well as dishonesty during the writing of tests and examinations will not be tolerated under any circumstances. All irregularities will be seen in a serious light and will be reported to the registrar (academic).

Plagiarism is a serious form of academic misconduct. It involves both appropriating someone else's work and passing it off as one's own work afterwards. Thus, you commit plagiarism when you present someone else's written or creative work (words, images, ideas, opinions, discoveries, artwork, music, recordings, computer-generated work, etc.) as your own. Only hand in your own original work. Indicate precisely and accurately when you have used information provided by someone else. Referencing must be done in accordance with a recognised system. Indicate whether you have downloaded information from the Internet.

## 3. STUDENT SUPPORT

The University of Pretoria supports you in various ways free of charge. For student support see below.

Title	Motivation	Contact
Faculty Student Advisors	<ul style="list-style-type: none"> <li>• Academic support</li> <li>• Goal setting &amp; motivation</li> <li>• Adjustment to university life</li> <li>• Test/Exam preparation</li> <li>• Stress management</li> <li>• Career exploration</li> </ul>	Individual consultations and workshops 
<b>FLY@UP</b>	<ul style="list-style-type: none"> <li>• Think carefully before dropping modules (after the closing date for amendments or cancellation of modules).</li> <li>• Make responsible choices with your time and work consistently.</li> <li>• Aim for a good semester mark. Do not rely on the examination to pass</li> </ul>	<a href="http://www.up.ac.za/fly@up">www.up.ac.za/fly@up</a> email:fly@up.ac.za 
e-learning support	<ul style="list-style-type: none"> <li>• Report a problem you experience to the Student Help Desk.</li> <li>• Approach the assistants at the help desks (adjacent to the Student Computer Laboratories in IT Building, NW2, CBT, etc.).</li> <li>• Visit the open labs in the Informatorium Building to report problems at the offices of the Student Help Desk.</li> </ul>	<ul style="list-style-type: none"> <li>• Call 012 420 3837</li> <li>• Email: <a href="mailto:studenthelp@up.ac.za">studenthelp@up.ac.za</a></li> </ul>
Night safety: Green Route	<ul style="list-style-type: none"> <li>• From 18:00 until 06:00 Security Officers are available to escort you (on foot) to and from your residence or campus anywhere east of the Hatfield campus through to the LC de Villiers terrain.</li> <li>• Departure point is at the ABSA ATM next to the Merensky Library.</li> <li>• Phone the Operational Management Centre if you need a Security Officer to accompany you from your residence to campus.</li> </ul>	Green Route <ul style="list-style-type: none"> <li>• Call 012 420 2310/2760</li> </ul>

## 4. MODULE INFORMATION

**4.1. Prerequisites.** If you registered before 2020: STK 210, STK 220 or WST 211, WST 221  
**If you registered in 2020/2021:** STK 210, STK 220, WST 212 or WST 211, WST 221, WST 212

## 4.2. Lecturers.

Name	Responsibilities
Dr Gao Maribe	Course coordinator and Lecturer
Ms Gandhi Jafta	Teaching assistant

For communication on courses content please make use of the click up discussion board, for administration related issues please email the lecturer directly on [g.maribe@up.ac.za](mailto:g.maribe@up.ac.za)

## 4.3. Online class schedule.

Day	Time	Type
Monday	15:30 – 18:20	Online practical consultation
Monday	13:30 – 14:20	<b>Lecture 1</b>
Wednesday	15:30 – 17:20	<b>Lecture 2</b>
Thursday	11:30 – 12:20	Pre-recorded video / Collaborate online lecture

## 4.4. Study Material.

4.4.1. *Handbook.* The prescribed book for this course is:

Title:	The Data Science Handbook
Authors:	Field Cady
Publisher:	MIT Press

4.4.2. *Software.* STK 353 is a coding module, meaning that all tests and assignments will involve coding to a lesser or larger extend (but mostly larger extend). In order to optimise the learning experience, we recommend that you install the Anaconda platform on you laptop or computer. Installation documentation can be found here: <https://docs.anaconda.com/anaconda/install/>.

**4.5. Module credits.** This module carries a weighting of 25 credits, indicating that a student should spend an average of 250 hours to master the required skills. This means that you should devote an average of 20 hours of study time per week to this module. The scheduled contact time is approximately 5 hours per week, which means that at least another 12 hours per week of own study time should be devoted to the module. Very important to note is that the allocated two hours for practical are not enough to complete the assignments. Practical assignments are made available well in advance in order for you to start working on the assignment in a timely manner.

**4.6. Continous Assessment.** The assessment model for STK 353 is formative continuous assessment. This means that no formal exam nor formal semester tests will take place. A total of 12 formative continuous assessment opportunities are available during the duration of the course. The continous assessment will take on three forms:

4.6.1. *Gradescope Quizzes.* A gradescope quiz will be released bi-weekly (unless otherwise stated). A total of 5 best out of 6 quizzes will contribute towards the continuous assessment mark. For this reason there will be no sick or special quiz due to unforeseen circumstances.

4.6.2. *Gradescope Practical Assignments.* Practical assignments to evaluate your understanding of a combination of learning outcomes. Note that only individual assignments forms part of this course. You'll be given clear instructions on how to submit your practical coding assignments. Note that regulations on plagiarism apply to practical assignments. A total of 5 best out of 6 assignments will contribute towards the continuous assessment mark and there will be no special or sick assignment. The release date (when the assignment will be made available) and submission deadline for each assignment will be posted on a weekly schedule.

## 5. STUDY OVERVIEW

The course consists of two modules: Basics – Module A and Modelling – Module B. Module A provides a solid foundation in the most important Python techniques needed to master a typical data science project. Module B introduces statistical models in order to make predictions and decisions based on the data. You will be familiar with some of these models and other will be new. Module A and B combined forms the workflow of a typical data science project (as illustrated in Figure 5). The workflow illustrates the flow of tasks – from where the data scientist receive the data, to where the results are communicated to the client or end user. All the phases in the data science workflow are addressed in this module.

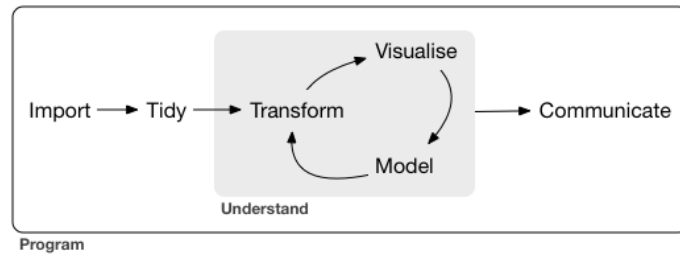


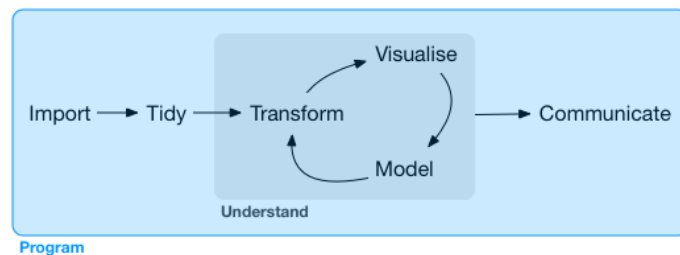
FIGURE 1. Data Science Workflow (Taken from [1])

## 6. MODULE A - BASICS

In preparation for the module, the student needs to have a basic understanding of Python. The first practical consists of an introductory exercise in Python. The mastering of this exercise is vital to ensure successful continuation of this practical course.

### 6.1. Theme 1: Program. (2 weeks)

Programming is a cross-cutting skill needed for all data science work: you must use a computer to do data science; you cannot do it in your head, or with pencil and paper. This theme focuses on the fundamental programming skills which is necessary in every step of the data science workflow.



Text book:	Chapters 3, 20
Quiz 1 :	Basic Syntax

6.1.1. *Arrays, Matrices and Dictionaries.* Outcomes: At the end of this study unit, the student should be able to understand and implement

- lists
- arrays
- matrices
- dictionaries
- lists and dictionary comprehensions

### 6.1.2. *Algorithms.*

Text book: Chapters 3, 20

Outcomes: At the end of this study unit, the student should be able to:

- implement loops and functional programming
- understand and implement the basic structure of an algorithm

Assignment 1: Coding

## 6.2. **Theme 2: Explore.** (2 weeks)

The goal of this theme is to master the basic tools of data exploration.

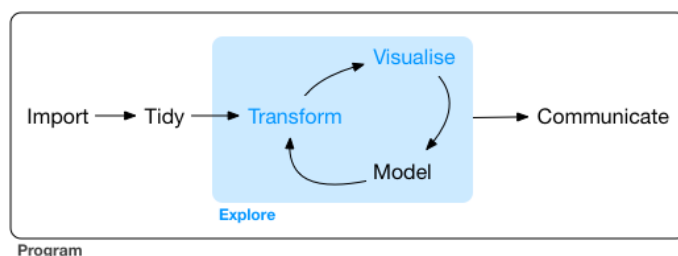


FIGURE 2. Explore theme in the Data Science Workflow

### 6.2.1. *Visualisation.*

Text book: Chapters 9

Outcomes: At the end of this study unit, the student should be able to:

- Implement the basic structure of a matplotlib plot
- Identify common plot types
- Create presentation quality statistical graphs, including appropriate use of scale, color, labels, reference markers

### 6.2.2. *Data Transformation.*

Textbook: Chapter 3  
Quiz 2 : Pandas

Outcomes: At the end of this study unit, the student should know the key verbs to:

- select important variables,
- filter out key observations,
- create new variables, and
- compute summaries.

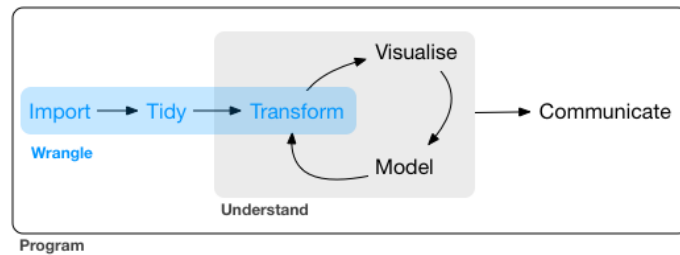
6.2.3. *Exploratory Data Analysis*. Outcomes: At the end of this study unit, the student should be able to combine visualisation and transformation in order to answer interesting questions about the data. The student should also be able to point out common visualisation mistakes and suggest alternative techniques. Communication skills of EDA should include:

- Describe a statistical graph using common vocabulary
- Read and think critically about a graph
- Create a visualisation that conveys key points of an analysis

Assignment 2: Explore
-----------------------

### 6.3. Theme 3: Wrangle. (1 week)

The goal of this theme is to master the basic tools of data wrangling.



#### 6.3.1. Data Import & Tidy Data.

Textbook:	Chapters 3–5
Quiz 3 :	Import & tidy

Outcomes: At the end of this study unit, the student should be able to:

- get data from disk and into Python
- understand underlying principles of tidy data, and how to get your data into a tidy form.

## 7. MODULE B - APPLICATION

### 7.1. Theme 4: Sampling. (2 weeks)

#### 7.1.1. Sampling games.

Textbook:	Chapter 18
Quiz 4:	Probability

Outcomes: At the end of this study unit, the student should be able to:

- Be able to frame thought experiments (or games) as probability problems
- Test and evaluate probability problems using simulation techniques
- Interpret simulation results
- Utilise the Python object oriented programming (OOP) paradigm

Assignment 3: Advanced algorithms
-----------------------------------

### 7.1.2. *Sampling Distributions.*

Textbook: Chapter 18

Outcomes: At the end of this study unit, the student should be able to:

- Understand simple methods on how to simulate/sample from simple continuous and discrete probability distributions
- Implement sampling techniques in Python
- Understand Monte Carlo Integration and simulation
- Perform a simulation study to discover probability laws
- Know how to use samples to make statements about the population

Assignment 4: Monte Carlo Simulation

### 7.2. **Theme 5: Text Mining.** (2 weeks)

This study theme deals with natural language as a data source. The four outcomes are:

- Clean text
- Locate/extract feature
- Derive features
- Analyse text

Text book: Chapter 16  
Quiz 5: Text mining 1

Assignment 5: Natural language processing

### 7.3. **Theme 6: Machine Learning.** (3 weeks)

The purpose of this theme is to introduce the student to machine learning and understand the difference between Statistics and Machine learning. Two major fields of machine learning is supervised and unsupervised learning.

#### 7.3.1. *Study Unit: Introduction to Machine Learning.*

Textbook: Chapters 6

Study unit learning objectives:

- Machine learning: What, why & when
- Supervised learning
- Unsupervised learning
- Understand experimental design in machine learning: splitting of dataset into training and test set, performance metrics, cross-validation.
- Know how to interpret and communicate results



### 7.3.2. Study Unit: Supervised learning.

Textbook: Chapters 8
----------------------

Study unit learning objectives:

- Linear regression implemented in Python
- Logistic Regression as a classification method
- Naive Bayes as a classification method
- Evaluation and interpretation of model results

Assignment 6: Supervised learning
-----------------------------------

### 7.3.3. Study Unit: Unsupervised learning.

Textbook: Chapters 10
Quiz 6: Unsupervised learning

Study unit learning objectives:

- k-means clustering
- Topic modelling
- Word embeddings

## REFERENCES

- [1] Hadley Wickham and Garrett Grolemund. 2017 *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1st ed.). O'Reilly Media, Inc..