

CSC 5741 (2021/22)

Data Mining and Warehousing

Lecture 2: Python for Data Mining and Machine Learning

Lighton Phiri

Department of Library & Information Science
University of Zambia

<http://lis.unza.zm/~lightonphiri>

Updates on Activities—March 28, 2022 (1/2)

- Trial paper reading discussion moved to next week
 - Review paper and aligned with grading rubric
- Trial Talk moved to next week
- Invited Speakers to be confirmed

Controlled Vocabularies in Digital Libraries: Challenges and Solutions for Increased Discoverability of Digital Objects

Bertha Chipangila^{1†}, Eric Liswaniso^{1†}, Andrew Mawila^{1†}, Philomena Mwanza^{1†}, Daisy Nawila^{1†}, Robert M'sendo², Mayumbo Nyirenda² and Lighton Phiri^{1*}

¹Department of Library and Information Science, University of Zambia, P.O Box 32379, Lusaka, 10101, Zambia.

²Department of Computer Science, University of Zambia, P.O Box 32379, Lusaka, 10101, Zambia.

*Corresponding author(s). E-mail(s): lighton.phiri@unza.zm;
Contributing authors: 13009428@student.unza.zm; 15058500@student.unza.zm;
15014570@student.unza.zm; 15018148@student.unza.zm; 15019551@student.unza.zm;
20171520216@student.unza.zm; mayumbo.nyirenda@cs.unza.zm;

[†]These authors contributed equally to this work.

Abstract
Digital Library Systems are widely used in the Higher Education sector, through the use of Institutional Repositories (IRs), to collect, store, manage and make available scholarly research output produced by Higher Education Institutions (HEIs). This wide application of IRs is a direct response to the increase of scholarly research output produced. In order to facilitate discoverability of digital content in IRs, accurate, consistent and comprehensive association of descriptive metadata to digital

March 28, 2022

CSC 5741 (2021/22) L03 - 2

Updates on Activities—March 28, 2022 (2/2)

A screenshot of a digital library interface showing a search result for a paper titled "A Bibliometric Approach for Detecting the Gender Gap in Computer Science". The result shows 80% coverage and a link to a PDF file. A callout box highlights the gender gap statistic.

lighton.phiri@unza.zm: (25%) Accuracy
(25%) Coverage
(10%) Depth
(20%) Presentation and Layout
(0%) Personal Reflection

> Paper readings are a nice way of identifying gaps: one of the ways of identifying gaps is contextualising research to our environment: Africa and/or Zambia... How can we be more specific? How can we take advantage of the approach or modified variations of it?
> Identifying the role of Scopus??
> Virtually no personal reflection
> Virtually no critical review/arguments presented, mostly just facts and figures
> Virtually no critical review/arguments presented, mostly just facts and figures

lighton.phiri@unza.zm * Research encompasses much more than publishing? Do you think the problem statement and indeed the title is indicative of "Gaps in Computer Science"? Perhaps this should have been "Identifying Gaps in Computer Science"?

lighton.phiri@unza.zm * Whenever a study draws comparisons with existing literature, you want to draw comparisons

> Are there any shortcomings with the proposed approach when compared with existing literature? What approach would be relatively more effective? In what instances would the proposed approach be desirable?

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
- Part III: Getting Started With Python
- Part IV: Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 4

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
- Part III: Getting Started With Python
- Part IV: Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 5

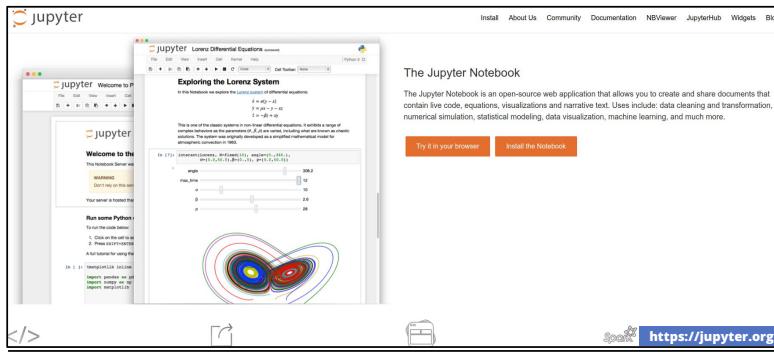
Lecture Series Outline

- Part I: Jupyter Notebooks
 - Jupyter Notebooks Interface
 - Textual Content
 - Live Code
 - Visualisations
- Part II: Google Colab
- Part III: Getting Started With Python
- Part IV: Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 6

About Jupyter Notebooks (1/2)

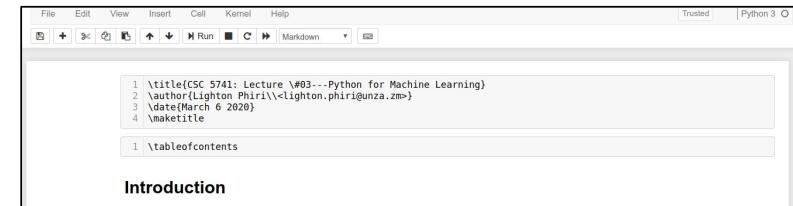


The screenshot shows the Jupyter Notebook interface. On the left, there's a sidebar with a 'jupyter' logo and various links like 'Install', 'About Us', 'Community', etc. The main area displays a notebook titled 'Exploring the Lorenz System'. It contains a plot of the Lorenz attractor and some text explaining the system. Below the plot, there are two buttons: 'Try it in your browser!' and 'Install the Notebook'. At the bottom, there's a code cell with Python code for generating the plot, and a footer with the URL <https://jupyter.org>.

March 28, 2022

CSC 5741 (2021/22) L03 - 7

About Jupyter Notebooks (2/2)



The screenshot shows a Jupyter Notebook cell. The code cell contains the following Python code:`1 \title{CSC 5741: Lecture #03---Python for Machine Learning}
2 \author{Lightfoot Phiri<lightfoot.phiri@unza.zm>}
3 \date{March 6 2020}
4 \maketitle`

Below the code cell, there's a section titled 'Introduction'.

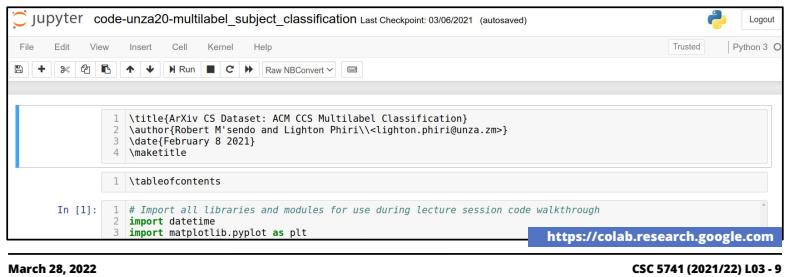
- A notebook is a web application that contains descriptive textual content, live code, equations and visualizations.
 - While we shall predominantly use the Python kernel, additional kernels for other languages like R can be installed.

March 28, 2022

CSC 5741 (2021/22) L03 - 8

About Jupyter Notebooks: Installation

- Installation instructions are available online (<https://jupyter.org/install>)



A screenshot of a Jupyter Notebook interface. The title bar says "jupyter code-unza20-multilabel_subject_classification Last Checkpoint: 03/06/2021 (autosaved)". The menu bar includes File, Edit, View, Insert, Cell, Kernel, Help. The toolbar has icons for New, Open, Save, Run, Cell, Kernel, Help, and Code. The code cell contains Python code for setting a title and importing modules. The output cell shows the rendered HTML output from the code.

```
1 \title{ArXiv CS Dataset: ACM CCS Multilabel Classification}
2 \author{Robert M'sendo and Lighton Phiri\\<lighton.phiri@unza.zm>}
3 \date{February 8 2021}
4 \maketitle

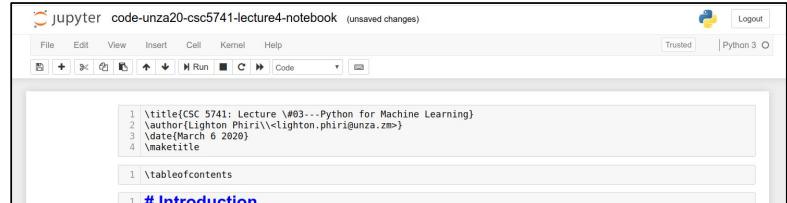
1 \tableofcontents

In [1]: 1 # Import all libraries and modules for use during lecture session code walkthrough
2 import datetime
3 import matplotlib.pyplot as plt
```

<https://colab.research.google.com/>

March 28, 2022 CSC 5741 (2021/22) L03 - 9

About Jupyter Notebooks: UI



A screenshot of a Jupyter Notebook interface. The title bar says "jupyter code-unza20-csc5741-lecture4-notebook (unsaved changes)". The menu bar includes File, Edit, View, Insert, Cell, Kernel, Help. The toolbar has icons for New, Open, Save, Run, Cell, Kernel, Help, and Code. The code cell contains Python code for setting a title and importing modules. The output cell shows the rendered HTML output from the code.

```
1 \title{CSC 5741: Lecture #03---Python for Machine Learning}
2 \author{Lighton Phiri\\<lighton.phiri@unza.zm>}
3 \date{March 6 2020}
4 \maketitle

1 \tableofcontents

1 # Introduction
```

- Text and live code is specified in cells and menubar and/or toolbar are used to execute cell contents
- Output appears immediately below the input cell

March 28, 2022

CSC 5741 (2021/22) L03 - 10

About Jupyter Notebooks: Text (1/4)

```
1 # Introduction
2
3 During these "hands-on" activities, we will explore and experiment the following:
4 1. Jupyter Notebooks---Quick walkthrough of Jupyter Notebooks
5 2. Python 3---Crash course introduction to Python 3
6 3. Core Python Modules---Quick walkthrough of some core Python modules that will be used in the course.
7
8 In all instances, you are encouraged to make references to online documentation for the various tools.
9 Additionally, you can exploit tools such as [Zéal Offline Documentation Browser](https://zealdocs.org) to
10 download and search through offline documentation. You are also encouraged to look up and explore other
11 libraries, especially as you work towards the Mini Projects.
```

- Textual content is primarily specified using the markup language “Markdown”
 - You essentially specify the structure of your text, similar to HTML

March 28, 2022

CSC 5741 (2021/22) L03 - 11

About Jupyter Notebooks: Text (2/4)

```
1 # Introduction
2
3 During these "hands-on" activities, we will explore and experiment the following:
4 1. Jupyter Notebooks---Quick walkthrough of Jupyter Notebooks
5 2. Python 3---Crash course introduction to Python 3
6 3. Core Python Modules---Quick walkthrough of some core Python modules that will be used in the course.
7
8 In all instances, you are encouraged to make references to online documentation for the various tools.
9 Additionally, you can exploit tools such as [Zéal Offline Documentation Browser](https://zealdocs.org) to
10 download and search through offline documentation. You are also encouraged to look up and explore other
11 libraries, especially as you work towards the Mini Projects.
```

- Common markup: Headings
 - # h1
 - ## h2
 - ### h3
 - ##### h4

March 28, 2022

CSC 5741 (2021/22) L03 - 12

About Jupyter Notebooks: Text (3/4)

```
1 # Introduction
2
3 During these "hands-on" activities, we will explore and experiment the following:
4 1. Jupyter Notebooks---Quick walkthrough of Jupyter Notebooks
5 2. Python 3---Crash course introduction to Python 3
6 3. Core Python Modules---Quick walkthrough of some core Python modules that will be used in the course.
7
8 In all instances, you are encouraged to make reference to online documentation for the various tools.
Additionally, you can exploit tools like [Zeal Offline Documentation Browser](https://zealdocs.org) to
download and search through offline documentation. You are also encouraged to look up and explore other
libraries, especially as you work towards the Mini Projects.
```

- Common markup: Lists—Unordered
 - * Jupyter Notebooks
 - * Python 3
 - * Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 13

About Jupyter Notebooks: Text (4/4)

```
1 # Introduction
2
3 During these "hands-on" activities, we will explore and experiment the following:
4 1. Jupyter Notebooks---Quick walkthrough of Jupyter Notebooks
5 2. Python 3---Crash course introduction to Python 3
6 3. Core Python Modules---Quick walkthrough of some core Python modules that will be used in the course.
7
8 In all instances, you are encouraged to make reference to online documentation for the various tools.
Additionally, you can exploit tools like [Zeal Offline Documentation Browser](https://zealdocs.org) to
download and search through offline documentation. You are also encouraged to look up and explore other
libraries, especially as you work towards the Mini Projects.
```

- Common markup: Lists—Unordered
 - 1. Jupyter Notebooks
 - 2. Python 3
 - 3. Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 14

About Jupyter Notebooks: Code (1/2)

```
In [3]: 1 # 1. Draw a line plot showing the trends of the Lusaka BNB between November 2016 and April 2018
2 # We will plot BNB as a function of months
3
4 # Format input data points
5 var_bnb_months = ['Nov 16','Dec 16','Jan 17','Feb 17','Mar 17','Apr 17','May 17','June 17','July 17','Aug
6 17','Sept 17','Oct 17','Nov 17','Dec 17','Jan 18','Feb 18','Mar 18','Apr 18']
6 var_bnb_values =
[5085.14,4976.67,4935.46,4918.76,5017.89,4973.03,4952.69,4958.52,4859.35,4928.37,4883.57,4869.47,4924.54,4957.
7 47,5229.14,5385.42,5574.81,5433.04]
8 plt.style.use("ggplot") # Use the visually appealing ggplot R theme
9 plt.plot(var_bnb_months,var_bnb_values,color="red") # plot BNB months vs BNB values
```

- Python code, shell commands and magics are the most common type of code
 - Python code is specified in its raw form in the cells

March 28, 2022

CSC 5741 (2021/22) L03 - 15

About Jupyter Notebooks: Code (2/2)

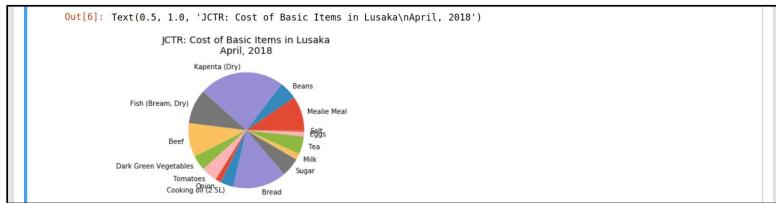
```
In [3]: 1 # 1. Draw a line plot showing the trends of the Lusaka BNB between November 2016 and April 2018
2 # We will plot BNB as a function of months
3
4 # Format input data points
5 var_bnb_months = ['Nov 16','Dec 16','Jan 17','Feb 17','Mar 17','Apr 17','May 17','June 17','July 17','Aug
6 17','Sept 17','Oct 17','Nov 17','Dec 17','Jan 18','Feb 18','Mar 18','Apr 18']
6 var_bnb_values =
[5085.14,4976.67,4935.46,4918.76,5017.89,4973.03,4952.69,4958.52,4859.35,4928.37,4883.57,4869.47,4924.54,4957.
7 47,5229.14,5385.42,5574.81,5433.04]
8 plt.style.use("ggplot") # Use the visually appealing ggplot R theme
9 plt.plot(var_bnb_months,var_bnb_values,color="red") # plot BNB months vs BNB values
```

- Python code, shell commands and magics are the most common type of code
 - Shell commands are prefixed with "!"
 - Cell magics are prefixed with "%"
 - Available magics specified with "%!smagic"

March 28, 2022

CSC 5741 (2021/22) L03 - 16

About Jupyter Notebooks: Visualisations



- Visualisations can be generated using plotting libraries like matplotlib or using HTML via the "%html" magic

March 28, 2022

CSC 5741 (2021/22) L03 - 17

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
 - Google Colab Interface
- Part III: Getting Started With Python
- Part IV: Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 18

Google Colab Interface (1/5)

- Google Colaboratory is a cloud-based alternative to Jupyter Notebook

<https://colab.research.google.com>

March 28, 2022

CSC 5741 (2021/22) L03 - 19

Google Colab Interface (2/5)

- Google Colaboratory is a cloud-based alternative to Jupyter Notebook

<https://colab.research.google.com>

March 28, 2022

CSC 5741 (2021/22) L03 - 20

Google Colab Interface (3/5)

- Google Colaboratory is a cloud-based alternative to Jupyter Notebook

A screenshot of the Google Colab interface. The top navigation bar includes File, Edit, View, Insert, Runtime, Tools, Help, and a user icon. A red box highlights the 'Code' button in the top right corner. The sidebar on the left contains sections like 'Code snippets', 'Adding form fields', 'Camera Capture', 'Cross-output communication', 'display JavaScript to execute Java...', and 'Downloading files or importing dat...'. Below the sidebar is a code cell with the following content:

```
Unsupported Cell Type. Double-Click to inspect/edit the content.
```

The code cell has a red border. At the bottom of the screen, there is footer text: 'March 28, 2022' and 'CSC 5741 (2021/22) L03 - 21'.

Google Colab Interface (4/5)

- Google Colaboratory is a cloud-based alternative to Jupyter Notebook

A screenshot of the Google Colab interface, identical to the one above but with a red box highlighting the entire sidebar area. The sidebar contains sections like 'Code snippets', 'Adding form fields', 'Camera Capture', 'Cross-output communication', 'display JavaScript to execute Java...', and 'Downloading files or importing dat...'. Below the sidebar is a code cell with the following content:

```
Unsupported Cell Type. Double-Click to inspect/edit the content.
```

The code cell has a red border. At the bottom of the screen, there is footer text: 'March 28, 2022' and 'CSC 5741 (2021/22) L03 - 22'.

Google Colab Interface (5/5)

- Google Colaboratory is a cloud-based alternative to Jupyter Notebook

A screenshot of the Google Colab interface. The top navigation bar includes File, Edit, View, Insert, Runtime, Tools, Help, and a user icon. A red box highlights the 'Code' button in the top right corner. The sidebar on the left contains sections like 'Code snippets', 'Adding form fields', 'Camera Capture', 'Cross-output communication', 'display JavaScript to execute Java...', and 'Downloading files or importing dat...'. Below the sidebar is a code cell with the following content:

```
Unsupported Cell Type. Double-Click to inspect/edit the content.
```

The code cell has a red border. At the bottom of the screen, there is footer text: 'March 28, 2022' and 'CSC 5741 (2021/22) L03 - 23'.

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
- Part III: Getting Started With Python
 - Introduction
 - Installation and Setup
 - Basics
 - Data Structures
 - Flow Control
 - Functions and Modules
- Part IV: Core Python Libraries

March 28, 2022

CSC 5741 (2021/22) L03 - 24

CSC 5741 Code and Datasets

lightonphiri initial commit 1ecff39 7 minutes ago 1 commit

README.md Initial commit 7 minutes ago

README.md

About 2021/22 CSC 5741: Data Mining and Warehousing

[launch binder](#)

2021/22 CSC 5741: Data Mining and Warehousing

Readme
 0 stars
 1 watching
 0 forks

Releases
 No releases published Create a new release

Packages
 No packages published Publish your first package

Learning Outcomes <https://github.com/lightonphiri/misc-unza22-csc5741>

March 28, 2022 CSC 5741 (2021/22) L03 - 25

CSC 5741 Code and Datasets

lightonphiri Added code and datasets ... db7809e 5 minutes ago 3 commits

code Added code and datasets 5 minutes ago
 slides Added slides and notes 17 minutes ago

README.md Initial commit 21 minutes ago

README.md

About 2021/22 CSC 5741: Data Mining and Warehousing

[launch binder](#)

2021/22 CSC 5741: Data Mining and Warehousing

Readme
 0 stars
 1 watching
 0 forks

Releases
 No releases published Create a new release

Packages
 No packages published Publish your first package

<https://github.com/lightonphiri/misc-unza22-csc5741>

March 28, 2022 CSC 5741 (2021/22) L03 - 26

Motivation

Where ML developers deploy their code
 % of ML developers Q4 2019 (n=2,632) | Q2 2019 (n=2,677)

Deployment Type	Q4 2019 (%)	Q2 2019 (%)
Desktop/laptop computers	56%	61%
Public cloud	32%	30%
Private cloud (cloud only available to certain users)	25%	25%
On-premise servers	24%	25%
Hardware architectures other than CPU	24%	22%

JavaScript, Python and Kotlin have grown the fastest in the past two years
 Active software developers, globally, in millions Q4 2019 (n=12,066)

Language	Q4 2019 (millions)	Q2 2019 (millions)
JavaScript	12.2M	8.4M
Python	8.2M	6.3M
Kotlin	5.8M	5.7M
C/C++	5.8M	5.8M
Swift	2.0M	2.0M
Go	1.4M	1.4M
Ruby	1.3M	1.3M
Objective-C	1.2M	1.2M
Rust	0.6M	0.6M
Lua	0.5M	0.5M

March 28, 2022 CSC 5741 (2021/22) L03 - 27

Motivation

Where ML developers deploy their code
 % of ML developers Q4 2019 (n=2,632) | Q2 2019 (n=2,677)

Deployment Type	Q4 2019 (%)	Q2 2019 (%)
Desktop/laptop computers	56%	61%
Public cloud	32%	30%
Private cloud (cloud only available to certain users)	25%	25%
On-premise servers	24%	25%
Hardware architectures other than CPU	24%	22%

JavaScript, Python and Kotlin have grown the fastest in the past two years
 Active software developers, globally, in millions Q4 2019 (n=12,066)

Language	Q4 2019 (millions)	Q2 2019 (millions)
JavaScript	12.2M	8.4M
Python	8.2M	6.3M
Kotlin	5.8M	5.7M
C/C++	5.8M	5.8M
Swift	2.0M	2.0M
Go	1.4M	1.4M
Ruby	1.3M	1.3M
Objective-C	1.2M	1.2M
Rust	0.6M	0.6M
Lua	0.5M	0.5M

March 28, 2022 CSC 5741 (2021/22) L03 - 28

Getting Started With Python (1/3)

```
Python 3.6.7 (default, Oct 22 2018, 11:32:17)
[GCC 8.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import this
The Zen of Python, by Tim Peters

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
```

March 28, 2022

CSC 5741 (2021/22) L03 - 29

Getting Started With Python (2/3)

- Python is an interpreted language
- Python is a scripting language
- Python is a general purpose language
- Python is an Object Oriented language
- [...]
- [...]
- We recommend using Python 3

March 28, 2022

CSC 5741 (2021/22) L03 - 30

Getting Started With Python (3/3)

- Python statements can be executed directly from the interpreter
- Python scripts can be executed as shell commands

March 28, 2022

CSC 5741 (2021/22) L03 - 31

Installation and Setup

- [...]

March 28, 2022

CSC 5741 (2021/22) L03 - 32

Installation and Setup (1/3)

- Download and install the latest version of Python 3
 - Installers also available on course Web page, in the resources directory
- Download and install the latest version of pip

The screenshot shows the Python 3 documentation page for "Compound Data Types". It includes code examples for list comprehensions and the enumerate function, and a brief explanation of lists. Below the content is a "Learn More" button and the URL <https://www.python.org/downloads>.

March 28, 2022

CSC 5741 (2021/22) L03 - 33

Installation and Setup (2/3)

- Any text editor will be sufficient for scripting.
 - Vim, Notepad [...]
- On IDEs
 - There are plenty of IDEs to choose from
 - In the recent past, I have worked with Wing 101 and Kate

The screenshot shows a Stack Overflow search results page for the query "What IDE to use for Python? [closed]". It displays a table of 1256 results, each listing an IDE with columns for name, type (e.g., Cross Platform, Commercial, Free), and features like Auto Code Completion, Integrated Debugging, and Bracket Matching.

March 28, 2022

CSC 5741 (2021/22) L03 - 34

Installation and Setup (3/3)

The screenshot shows the Visual Studio Code interface with the Python extension installed. It displays a sidebar with Python-related tools like "Python", "Python: Linting", and "Python: Debugging". The main workspace shows a file named "python_crash_course.ipynb" which is a Jupyter Notebook. A status bar at the bottom indicates the Python interpreter is selected.

March 28, 2022

CSC 5741 (2021/22) L03 - 35

Installation and Setup (3/3)

The screenshot shows the Visual Studio Code interface with a Jupyter Notebook open. The sidebar shows "code snippets oct2141 lecture-notebook.ipynb" and "PYTHON_CRASH_COURSE.ipynb". The main area displays the content of the notebook, including sections like "Introduction" and "Jupyter Notebooks". A status bar at the bottom indicates the Python interpreter is selected.

March 28, 2022

CSC 5741 (2021/22) L03 - 36

Basics

- No need to specify data types on variable declaration
- Indentation is important

March 28, 2022

CSC 5741 (2021/22) L03 - 37

Identifiers (1/2)

- Python is case-sensitive, meaning uppercase and lowercase are considered as different
 - age is different from AGE
 - favourite_course is different from Favourite_Course
- Variable names, like other identifiers, follow rules
 - can use letters, numbers or underscores
 - can't use other punctuation
 - can't start with a number
 - can't use Python keywords (reserved words)
- The assignment operator in Python is the equals sign =
 - >>> age = 19

March 28, 2022

CSC 5741 (2021/22) L03 - 38

Identifiers (2/2)

- Python keywords (reserved words) can't be used when naming identifiers
- >>> import keyword
- >>> keyword.kwlist
- ['False', 'None', 'True', 'and', 'as', 'assert', 'break', 'class', 'continue', 'def', 'del', 'elif', 'else', 'except', 'finally', 'for', 'from', 'global', 'if', 'import', 'in', 'is', 'lambda', 'nonlocal', 'not', 'or', 'pass', 'raise', 'return', 'try', 'while', 'with', 'yield']

March 28, 2022

CSC 5741 (2021/22) L03 - 39

Comments (1/2)

- Comments are useful in explaining your code, and are ignored by the Python interpreter
- Single line comments are simply indicated with a hash # character
- Everything to the right of the hash is ignored
 - >>> course_code = "csc5741" # creates a variable course_code

March 28, 2022

CSC 5741 (2021/22) L03 - 40

Comments (2/2)

- Multiple line comments are specified between sets of three quotes, ''' or """

```
'''Author: Mwangala Sikota  
Course: CSC 5741  
Lecture #04'''
```

```
"""Author: Mumbi Mumbi  
Course: CSC 5741  
Lecture #04 """
```

March 28, 2022

CSC 5741 (2021/22) L03 - 41

Data Types (1/3)

- Variables don't require explicit type declaration in Python, as in other programming languages
 - >>> x = 5
- There are a few basic data types in Python

Integers	int
Float	float
String	str
Boolean	bool

March 28, 2022

CSC 5741 (2021/22) L03 - 42

Data Types (2/3)

- Integer, whole numbers
 - >>> i = 23
- Float, floating point numbers
 - full stop indicates decimal point
 - >>> d = 2.345
- String, piece of text
 - enclosed in single ("") or double quotes ("""")
 - >>> x = 'CSC 5741'
 - >>> y = "CSC 5741"

March 28, 2022

CSC 5741 (2021/22) L03 - 43

Data Types (3/3)

- Boolean, true or false
 - values True and False, start with capital letter
 - 0, "", [], (), {}, None are considered False, everything else is True
 - >>> weekday = True

March 28, 2022

CSC 5741 (2021/22) L03 - 44

Functions (1/3)

- Functions are used to perform simple operations, sometimes on values
- Functions are called with round brackets ()
 - function_name()
- Functions can be passed certain values, which are referred to as parameters (or arguments) separated by commas
 - function_name(parameter)
 - function_name(parameter1, parameter2, ...)

March 28, 2022

CSC 5741 (2021/22) L03 - 45

Functions (2/3)

- Python has many built-in functions, here are some:
 - print() function prints information to the screen
 - input() function gets information from the user
 - type() function returns data type of variable or value
 - >>> x = 3
 - >>> type(x)
 - <class 'int'>

March 28, 2022

CSC 5741 (2021/22) L03 - 46

Functions (3/3)

```
def csc5741(x, y='Y', z='Z'):
    print(x + ' ' + y)
    return 0

csc5741('Xxxx', 'Yyyyy')
```

- All arguments are named
- Naming useful for optional arguments
- Return is optional

March 28, 2022

CSC 5741 (2021/22) L03 - 47

Data Structures

- Tuple
 - var = (1, 2, 3, 4, 5)
- List
 - var = [1, 2, 3, 4, 5]
- Dictionary
 - var = {"one":1, "two":2, "three":3, "four":4, "five":5}
- Set
 - var = {1, 2, 3, 4, 5}

March 28, 2022

CSC 5741 (2021/22) L03 - 48

Loops

```
for i in [1,2,3]:  
    print(i)  
  
while i < 5:  
    i += 1  
    print(i)
```

- No curly braces or "end for"
- Structure is derived from level of indentation
- One statement per line
- No semicolons required

March 28, 2022

CSC 5741 (2021/22) L03 - 49

Modules (1/2)

- Modules facilitate extensibility and reusability
- Modules are collections of functions adding functionality to Python
- Modules can be imported using import keyword
 - Once modules are imported, their functions can be accessed by using the module name
 - The help() function displays what is contained in a module

```
from math import sqrt  
import math
```

March 28, 2022

CSC 5741 (2021/22) L03 - 50

Modules (2/2)

- Single functions can be imported using the from statement
 - >>> from math import sqrt
- When using the from statement functions can be accessed without the module name
 - >>> sqrt(16)
- Everything from the module can be imported using an asterisk with the from statement
 - >>> from math import *

March 28, 2022

CSC 5741 (2021/22) L03 - 51

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
- Part III: Getting Started With Python
- Part IV: Core Python Libraries
 - Datasets
 - matplotlib
 - pandas
 - Scikit-learn

March 28, 2022

CSC 5741 (2021/22) L03 - 52

Datasets

- See “2021/22 CSC 5741” Astria and/or Google Drive folder (https://drive.google.com/drive/folders/1JUcWNxB_pYsncEx3do9ZrpLODX-KHsgL)

March 28, 2022

CSC 5741 (2021/22) L03 - 53

Matplotlib (1/7)

- The matplotlib library is best installed using pip, as with all libraries or using apt-get, if on Mac or Linux
 - pip3 install matplotlib
 - sudo apt-get install python-matplotlib
- Test installation by importing a library module

March 28, 2022

CSC 5741 (2021/22) L03 - 54

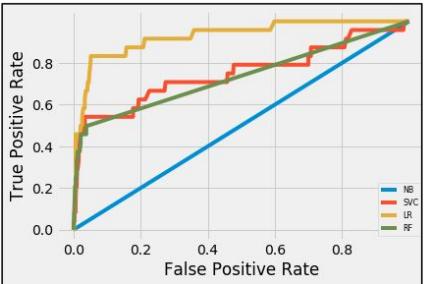
```
lightonphiri@lightonphiri-Lenovo-Ideapad-320-15IKB:~$ pip3 install matplotlib
Requirement already satisfied: matplotlib<3.0.0,>=2.0.0 in /usr/lib/python3/dist-packages
Requirement already satisfied: python-dateutil<2.3.1,>=2.1 in /local/lib/python3.6/site-packages
Requirement already satisfied: cycler>=0.10 in ./local/lib/python3.6/site-packages/cycler.py
Requirement already satisfied: pytz!=2019.3,>=2018.9 in /local/lib/python3.6/site-packages
Requirement already satisfied: six>=1.5 in ./local/lib/python3.6/site-packages
Requirement already satisfied: setuptools in /usr/lib/python3/dist-packages (from matplotlib)
lightonphiri@lightonphiri-Lenovo-Ideapad-320-15IKB:~$ 
```



```
lightonphiri@lightonphiri-Lenovo-Ideapad-320-15IKB:~$ python3
[GCC 8.2.1] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import matplotlib
>>> import matplotlib.pyplot as plt
>>> plt.show()
>>> 
```

Matplotlib (2/7)

- Basic elements of a plot
 - Plot title
 - Axis labels
 - Legend



March 28, 2022

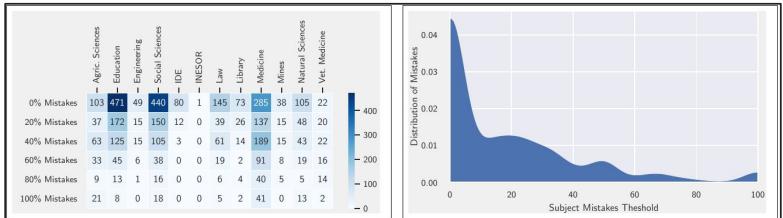
CSC 5741 (2021/22) L03 - 55

Matplotlib (2/7)

- Basic elements of a plot
 - Plot title, Axis labels and Legend

March 28, 2022

CSC 5741 (2021/22) L03 - 56



Matplotlib (3/7)

- **Creating plots is a four-step process**

- 1) Import matplotlib
- 2) Draw the plot
- 3) Specify plot aesthetics
- 4) Render plot

March 28, 2022

CSC 5741 (2021/22) L03 - 57

Matplotlib (4/7)

- **Creating plots is a four-step process**

- 1) Import matplotlib
import matplotlib.pyplot as plt
- 2) Draw the plot
- 3) Specify plot aesthetics
- 4) Render plot

March 28, 2022

CSC 5741 (2021/22) L03 - 58

Matplotlib (5/7)

- **Creating plots is a four-step process**

- 1) Import matplotlib
import matplotlib.pyplot as plt
- 2) Draw the plot
plt.plot([...])
plt.hist([...])
- 3) Specify plot aesthetics
- 4) Render plot

- **Illustration**

- Simple plots
- Plots using pandas
dataframe

March 28, 2022

CSC 5741 (2021/22) L03 - 59

Matplotlib (6/7)

- **Creating plots is a four-step process**

- 1) Import matplotlib
import matplotlib.pyplot as plt
- 2) Draw the plot
plt.plot([...])
plt.hist([...])
- 3) Specify plot aesthetics
plt.xlabel("...")
plt.ylabel("...")
- 4) Render plot

- **Illustration**

- Simple plots
- Plots using pandas
dataframe

March 28, 2022

CSC 5741 (2021/22) L03 - 60

Matplotlib (7/7)

- Creating plots is a four-step process
 - 1) Import matplotlib
 - import matplotlib.pyplot as plt
- Illustration
 - o Simple plots
 - o Plots using pandas dataframe
- 2) Draw the plot
plt.plot([...])
plt.hist([...])
- 3) Specify plot aesthetics
plt.xlabel("[...]"); plt.ylabel("[...]")
 plt.legend()
- 4) Render plot
plt.show()

March 28, 2022

CSC 5741 (2021/22) L03 - 61

Matplotlib—Exercises

- See “2021/22 CSC 5741” Astria and/or Google Drive folder (https://drive.google.com/drive/folders/1JUcWNxB_pYsncEx3do9ZrpLODX-KHsgL)

March 28, 2022

CSC 5741 (2021/22) L03 - 62

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
- Part III: Getting Started With Python
- Part IV: Core Python Libraries
 - o Datasets
 - o matplotlib
 - o pandas
 - o Scikit-learn

March 28, 2022

CSC 5741 (2021/22) L03 - 63

Pandas (1/9)

- Why use pandas instead a spreadsheet for data analysis
 - o Efficiency as data scales
 - o Very user-friendly
 - o Dataframe similar to spreadsheet

March 28, 2022

CSC 5741 (2021/22) L03 - 64

Pandas (2/9)

- Why use pandas instead a spreadsheet for data analysis
 - Efficiency as data scales
 - Very user-friendly
 - Dataframe similar to spreadsheet

March 28, 2022

CSC 5741 (2021/22) L03 - 55

Pandas (3/9)

- Pandas DataFrame
 - Two dimensional labeled data structure
 - DataFrame can be viewed as a representation of a Spreadsheet worksheet

	StudentID	Gender	Minor	LastName	... PassedTest3	
0	2017013156@student.unza.zm	M	Geography	Anayawa	...	NO
1	2017012891@student.unza.zm	M	Civic	Banda	...	NO
2	2017012962@student.unza.zm	M	Languages	Banda	...	NO
3	2017012963@student.unza.zm	M	Civic	Bwalya	...	NO
4	2017008514@student.unza.zm	M	History	Bwalya	...	NO
5	2017010497@student.unza.zm	M	Civic	Chafuka	...	NO
6	2017012923@student.unza.zm	M	Civic	Chatabela	...	YES
7	2017012983@student.unza.zm	M	History	Chakulya	...	NO
8	2017012964@student.unza.zm	M	Mathematics	Chitete	...	NO
9	2017008345@student.unza.zm	M	Mathematics	Chileshe	...	YES
10	2017012961@student.unza.zm	M	Mathematics	Chilumba	...	NO
11	2017012965@student.unza.zm	F	Languages	Chisha	...	NO
12	2017012930@student.unza.zm	F	History	Gondwe	...	NO
13	2017012990@student.unza.zm	M	Mathematics	Hamamba	...	NO
14	2017012912@student.unza.zm	M	Civic	Imakando	...	NO
15	2017012971@student.unza.zm	M	Mathematics	Jere	...	NO
16	2017012980@student.unza.zm	M	Civic	Kabaso	...	NO
17	2017012932@student.unza.zm	F	Civic	Kabwe	...	YES
18	2017012973@student.unza.zm	M	Geography	Kafwale	...	NO
19	2017001431@student.unza.zm	M	Mathematics	Kamanga	...	YES

March 28, 2022

CSC 5741 (2021/22) L03 - 66

Pandas (4/9)

- Pandas series
 - One dimensional labeled array that can hold any data type.
 - Similar to column in Spreadsheet applications

	StudentID	Gender	Minor	LastName	... PassedTest3	
0	2017013156@student.unza.zm	M	Geography	Anayawa	...	NO
1	2017012891@student.unza.zm	M	Civic	Banda	...	NO
2	2017012962@student.unza.zm	M	Languages	Banda	...	NO
3	2017012963@student.unza.zm	M	Civic	Bwalya	...	NO
4	2017008514@student.unza.zm	M	History	Bwalya	...	NO
5	2017010497@student.unza.zm	M	Civic	Chafuka	...	NO
6	2017012923@student.unza.zm	M	Civic	Chatabela	...	YES
7	2017012983@student.unza.zm	M	History	Chakulya	...	NO
8	2017012964@student.unza.zm	M	Mathematics	Chitete	...	NO
9	2017008345@student.unza.zm	M	Mathematics	Chileshe	...	YES
10	2017012961@student.unza.zm	M	Mathematics	Chilumba	...	NO
11	2017012965@student.unza.zm	F	Languages	Chisha	...	NO
12	2017012930@student.unza.zm	F	History	Gondwe	...	NO
13	2017012990@student.unza.zm	M	Mathematics	Hamamba	...	NO
14	2017012912@student.unza.zm	M	Civic	Imakando	...	NO
15	2017012971@student.unza.zm	M	Mathematics	Jere	...	NO
16	2017012980@student.unza.zm	M	Civic	Kabaso	...	NO
17	2017012932@student.unza.zm	F	Civic	Kabwe	...	YES
18	2017012973@student.unza.zm	M	Geography	Kafwale	...	NO
19	2017001431@student.unza.zm	M	Mathematics	Kamanga	...	YES

March 28, 2022

CSC 5741 (2021/22) L03 - 67

Pandas (5/9)

- Columns
 - Ellipse indicate more columns. Structure of data frame indicated on last line of output

	StudentID	Gender	Minor	LastName	... PassedTest3	
0	2017013156@student.unza.zm	M	Geography	Anayawa	...	NO
1	2017012891@student.unza.zm	M	Civic	Banda	...	NO
2	2017012962@student.unza.zm	M	Languages	Banda	...	NO
3	2017012963@student.unza.zm	M	Civic	Bwalya	...	NO
4	2017008514@student.unza.zm	M	History	Bwalya	...	NO
5	2017010497@student.unza.zm	M	Civic	Chafuka	...	NO
6	2017012923@student.unza.zm	M	Civic	Chatabela	...	YES
7	2017012983@student.unza.zm	M	History	Chakulya	...	NO
8	2017012964@student.unza.zm	M	Mathematics	Chitete	...	NO
9	2017008345@student.unza.zm	M	Mathematics	Chileshe	...	YES
10	2017012961@student.unza.zm	M	Mathematics	Chilumba	...	NO
11	2017012965@student.unza.zm	F	Languages	Chisha	...	NO
12	2017012930@student.unza.zm	F	History	Gondwe	...	NO
13	2017012990@student.unza.zm	M	Mathematics	Hamamba	...	NO
14	2017012912@student.unza.zm	M	Civic	Imakando	...	NO
15	2017012971@student.unza.zm	M	Mathematics	Jere	...	NO
16	2017012980@student.unza.zm	M	Civic	Kabaso	...	NO
17	2017012932@student.unza.zm	F	Civic	Kabwe	...	YES
18	2017012973@student.unza.zm	M	Geography	Kafwale	...	NO
19	2017001431@student.unza.zm	M	Mathematics	Kamanga	...	YES

March 28, 2022

CSC 5741 (2021/22) L03 - 68

Pandas (6/9)

• Index

- Automatically generated, but can be changed
- Uniquely identifies rows in the DataFrame

	StudentID	Gender	Minor	LastName	... PassedTest3	
0	2017013156@student.unza.zm	M	Geography	Anayava	...	NO
1	2017012891@student.unza.zm	M	Civic	Banda	...	NO
2	2017012892@student.unza.zm	M	Languages	Banda	...	NO
3	2017012893@student.unza.zm	M	Civic	Bwalya	...	NO
4	2017080514@student.unza.zm	M	History	Bwalya	...	NO
5	2017010497@student.unza.zm	M	Civic	Chafuka	...	NO
6	2017012923@student.unza.zm	M	Civic	Chabela	...	YES
7	2017012938@student.unza.zm	M	History	Chakulya	...	NO
8	2017012939@student.unza.zm	M	Mathematics	Chabula	...	NO
9	2017008343@student.unza.zm	M	Mathematics	Chileshe	...	YES
10	2017012966@student.unza.zm	F	Languages	Chilumba	...	NO
11	2017012967@student.unza.zm	F	History	Chisha	...	NO
12	2017012939@student.unza.zm	F	History	Gondwe	...	NO
13	2017012940@student.unza.zm	M	Mathematics	Hamamunda	...	NO
14	2017012941@student.unza.zm	F	Civic	Imakando	...	NO
15	2017012971@student.unza.zm	M	Mathematics	Jere	...	NO
16	2017012980@student.unza.zm	M	Civic	Kabaso	...	NO
17	2017012932@student.unza.zm	F	Civic	Kabwe	...	YES
18	2017012973@student.unza.zm	M	Geography	Kafwale	...	NO
19	2017004431@student.unza.zm	M	Mathematics	Kamsanga	...	YES

March 28, 2022

CSC 5741 (2021/22) L03 - 9

Pandas (7/9)

• Data

	StudentID	Gender	Minor	LastName	... PassedTest3	
0	2017012891@student.unza.zm	M	Civic	Banda	...	NO
1	2017012912@student.unza.zm	M	Languages	Banda	...	NO
2	2017012913@student.unza.zm	M	Civic	Bwalya	...	NO
3	2017080514@student.unza.zm	M	History	Bwalya	...	NO
4	2017010497@student.unza.zm	M	Civic	Chafuka	...	NO
5	2017012923@student.unza.zm	M	Civic	Chabela	...	YES
6	2017012938@student.unza.zm	M	History	Chakulya	...	NO
7	2017012939@student.unza.zm	M	Mathematics	Chabula	...	NO
8	2017008343@student.unza.zm	M	Mathematics	Chileshe	...	YES
9	2017012966@student.unza.zm	M	Mathematics	Chilumba	...	NO
10	2017012967@student.unza.zm	F	Languages	Chisha	...	NO
11	2017012968@student.unza.zm	F	History	Gondwe	...	NO
12	2017012939@student.unza.zm	F	History	Gondwe	...	NO
13	2017012940@student.unza.zm	M	Mathematics	Hamamunda	...	NO
14	2017012941@student.unza.zm	F	Civic	Imakando	...	NO
15	2017012971@student.unza.zm	M	Mathematics	Jere	...	NO
16	2017012980@student.unza.zm	M	Civic	Kabaso	...	NO
17	2017012932@student.unza.zm	F	Civic	Kabwe	...	YES
18	2017012973@student.unza.zm	M	Geography	Kafwale	...	NO
19	2017004431@student.unza.zm	M	Mathematics	Kamsanga	...	NO

March 28, 2022

CSC 5741 (2021/22) L03 - 70

Pandas (8/9)

• Some common operations

- Reading data files
 - `df.read_csv([...])`
 - `df.read_html([...])`
 - `df.read_json([...])`
 - `df.read_*`
- Inspecting dataframes
 - `df.head([...])`
 - `df.tail([...])`
 - `df.columns`
 - `df['...']`

March 28, 2022

CSC 5741 (2021/22) L03 - 1

Pandas (9/9)

• Some common operations

- Converting to different file formats
 - `df.to_csv([...])`
 - `df.to_excel([...])`
 - `df.to_sql([...])`
 - `df.to_*`
- Renaming columns
 - `df.rename(columns={...})`
- Aggregating data
 - `df.groupby(['...']).mean()`
 - `df.groupby(['...']).max()`

March 28, 2022

CSC 5741 (2021/22) L03 - 72

Pandas—Exercise

- See Jupyter Notebook “2021/22 CSC 5741: Lecture #04 Notebook—Python for Machine Learning” (<http://bit.ly/2Q2T2Lw>)

March 28, 2022

CSC 5741 (2021/22) L03 - 73

Lecture Series Outline

- Part I: Jupyter Notebooks
- Part II: Google Colab
- Part III: Getting Started With Python
- Part IV: Core Python Libraries
 - Datasets
 - matplotlib
 - pandas
 - Scikit-learn

March 28, 2022

CSC 5741 (2021/22) L03 - 74

Scikit-learn

- Scikit-learn
 - Ensure that the module is installed by using the import statement

```
lightonphirl@lightonphirl-Lenovo-ideapad-320-15IKB:~$ pip3 install sklearn
Collecting sklearn
  Downloading https://files.pythonhosted.org/packages/1e/7a/dbb3be0ce9bd5c0b7e3d0skLearn-0.0.tar.gz
    Collecting scikit-learn (from sklearn)
      Downloading https://files.pythonhosted.org/packages/5e/82/c0de5839d613b82bdd00scikit_learn-0.26.3-cp36-cp36m-manylinux1_x86_64.whl (5.4MB)
        0% |████████████████████████████████| 20KB 55KB/s eta 0:01:38
```

```
lightonphirl@lightonphirl-Lenovo-ideapad-320-15IKB:~$ python3
Python 3.6.7 (default, Oct 22 2016, 11:32:17)
[GCC 8.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import sklearn
>>> dir(sklearn)
['__SKLEARN_SETUP__', '__all__', '__builtins__', '__cached__', '__check_build__', '__doc__', '__file__', '__package__', '__path__', '__spec__', '__version__', '__config__', 'base', 'clone', 'externals', 'get_config', 'logger', 'logging', 're', 'set_config', 'setup_module', 'show_margins']
>>> __
```

March 28, 2022

CSC 5741 (2021/22) L03 - 75

scikit-learn—Exercises

- See Jupyter Notebook “2021/22 CSC 5741: Lecture #04 Notebook—Python for Machine Learning” (<http://bit.ly/2Q2T2Lw>)

March 28, 2022

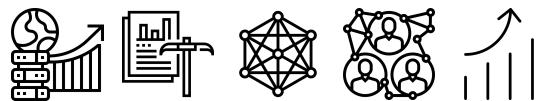
CSC 5741 (2021/22) L03 - 76

Bibliography

- [1] A Byte of Python
<https://python.swaroopch.com>
- [2] Python 3.4 Programming Tutorials
<https://www.youtube.com/playlist?list=PL6gx4Cwl9DGAcbMi1sH6oAMk4JHw91mC>
- [3] Python for Beginners | Python.org
<https://www.python.org/about/gettingstarted>
- [4] Pyplot tutorial – Matplotlib 3.0.3 documentation
<https://matplotlib.org/tutorials/introductory/pyplot.html>
- [5] 10 Minutes to pandas – pandas 0.22.0 documentation
<https://pandas.pydata.org/pandas-docs/version/0.22/10min.html>

March 28, 2022

CSC 5741 (2021/22) L03 - 7



CSC 5741 (2021/22)

Data Mining and Warehousing

Lecture 2: Python for Data Mining and Machine Learning

Lighton Phiri
Department of Library & Information Science
University of Zambia
<http://lis.unza.zm/~lightonphiri>