Data Analytics Course - Lesson 01

Ths. Vu Duy Khuong

Agenda

❑I. Introduction to Data Analytics (DA)

❑II. Data Analytics methodology

❑III. Requirement for Data Analytics

❑IV. Tools for Data Analytics

❑V. Basic Python for Data Analytics

I. Introduction to DA

1. What is DA?

Data Analytics is the process of working with data to collect useful information, thereby using the results of analysis to make informed decisions related to the strategic plans of the company.

I. Introduction to DA

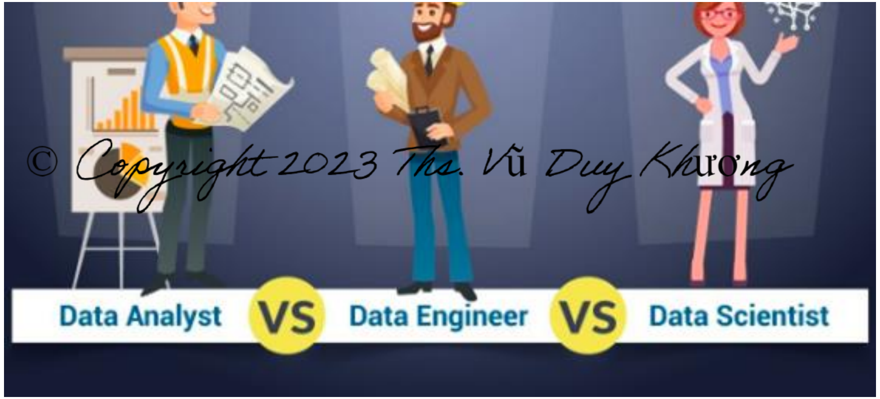
2. Some Confuse Terms

Terms	Explanation
Data Analytics	The process of examining data sets in order to find trends and draw conclusions about the information they contain.
Data Analyst	The person who make Data Analytics work.
Data Science	The field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.
Data Scientist	The person who make Data Science work.
Data Analysis	A subset of Data Analytics.

I. Introduction to DA

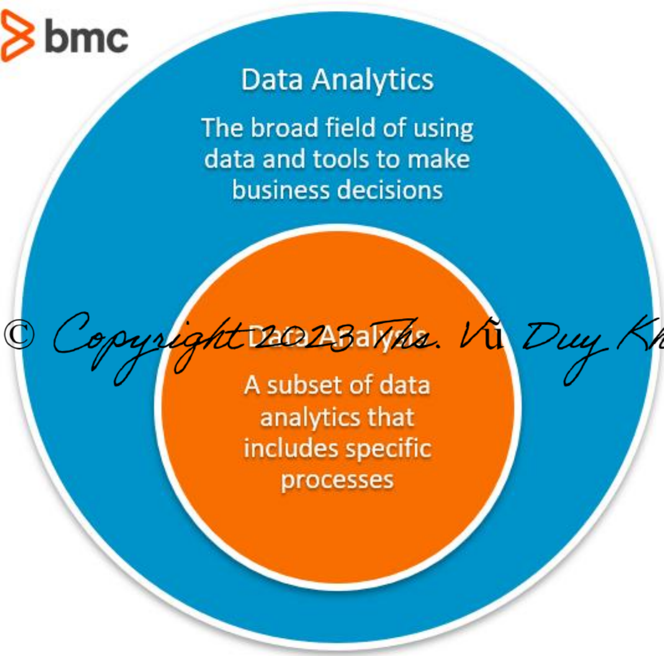
3. Compare Data Analyst, Data Science & Data Engineer

Data Analyst	Data Engineer	Data Scientist
Analyzes numeric data and uses it to help companies make better decisions.	Involves in preparing data. They develop, constructs, tests & maintain complete architecture.	Analyzes and interpret complex data. They are data wranglers who organize (big) data.



I. Introduction to DA

4. Compare Data Analytics and Data Analysis

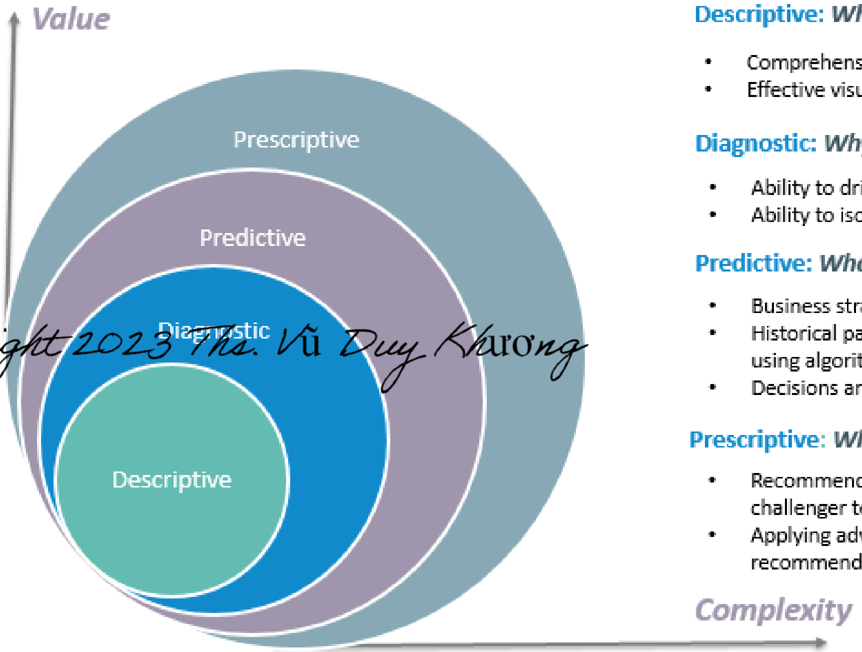

bmc
Data Analytics
The broad field of using data and tools to make business decisions
Data Analysis
A subset of data analytics that includes specific processes

Data Analytics	Data Analysis
Data analytics is the broad field of using data and tools to make business decisions.	Data analysis, a subset of data analytics, refers to specific actions.

I. Introduction to DA

5. Types of DA

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

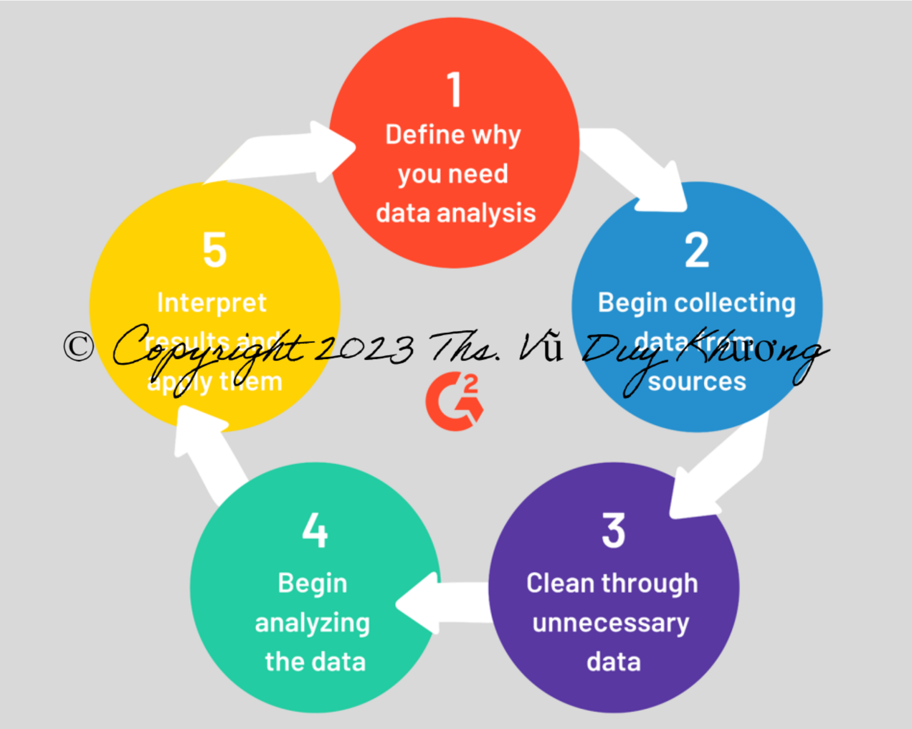
II. DA Methodology - DA Lifecycle

1. Importance of DA Lifecycle

- Data Analytics Lifecycle defines the roadmap of how data is generated, collected, processed, used, and analyzed to achieve business goals.
- It offers a systematic way to manage data for converting it into information that can be used to fulfill organizational and project goals.
- The process provides the direction and methods to extract information from the data and proceed in the right direction to accomplish business goals.
- Data professionals use the lifecycle’s circular form to proceed with data analytics in either forward or backward direction.
- Based on the newly received insights, they can decide whether to proceed with their existing research or scrap it and redo the complete analysis.

II. DA Methodology - DA Lifecycle

2. Phases of DA Lifecycle


1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

II. DA Methodology - DA Lifecycle

3. Roles of a DA

- Pre-processing and data gathering.

- Emphasis on representing data via reporting and visualization.

- Responsible for statistical analysis & data interpretation.

- Ensures data acquisition & maintenance.

- Optimize Statistical Efficiency & Quality.

III. Roles and Skill set for DA

1. Skill set of a DA

- Critical Thinking

- SQL - Structured Query Language.

- Probability and statistics

- R or Python–Statistical Programming

- Data Visualization

- Presentation & Report Skills

- Machine Learning

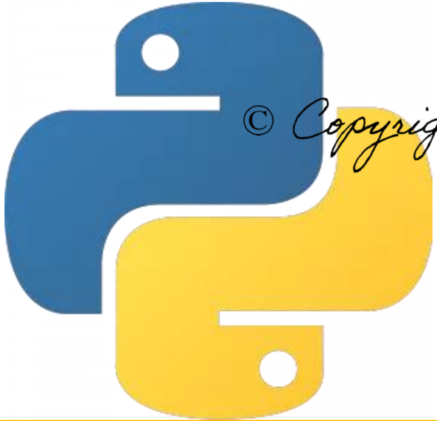
- Domain Knowledge

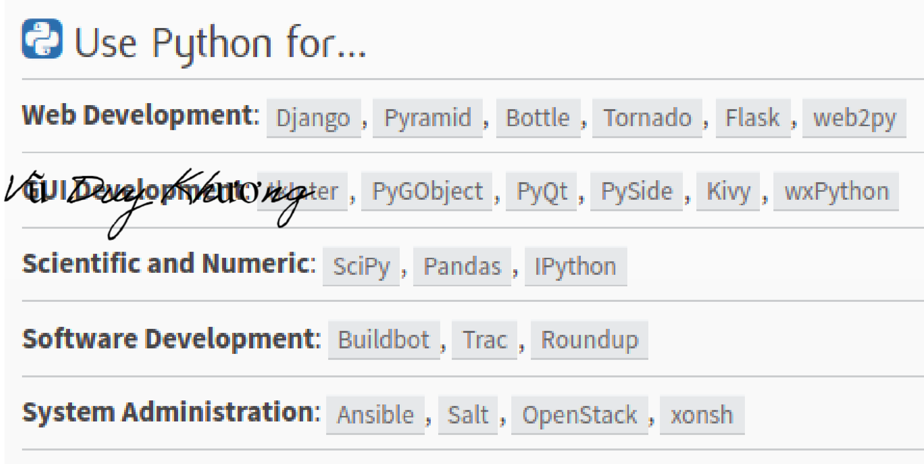
IV. Tools for DA

1. Programming Language

a, *Python*

- Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.
- Python is simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance.





Use Python for...

Web Development: Django , Pyramid , Bottle , Tornado , Flask , web2py

GUI Development: tkInter , PyGObject , PyQt , PySide , Kivy , wxPython

Scientific and Numeric: SciPy , Pandas , IPython

Software Development: Buildbot , Trac , Roundup

System Administration: Ansible , Salt , OpenStack , xonsh

IV. Tools for DA

1. Programming Language

b, R

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

IV. Tools for DA

1. Programming Language

c, SQL

- SQL is a standard language for accessing and manipulating databases.
- SQL lets you access and manipulate databases.

IV. Tools for DA

2. Libraries

a, Numpy and Scipy – Fundamental Scientific Computing

- NumPy stands for Numerical Python. The most powerful feature of NumPy is n-dimensional array. This library also contains basic linear algebra functions, Fourier transforms, advanced random number capabilities and tools for integration with other low level languages like Fortran, C and C++.
- SciPy stands for Scientific Python. It is built on NumPy. Scipy is one of the most useful library for variety of high level science and engineering modules like discrete Fourier transform, Linear Algebra, Optimization and Sparse matrices.

IV. Tools for DA

2. Libraries

b, *Pandas* *– Data Manipulation and Analysis*

- Pandas for structured data operations and manipulations. It is extensively used for data munging and preparation.
- Pandas were added relatively recently to Python and have been instrumental in boosting Python’s usage in data scientist community.

IV. Tools for DA

2. Libraries

c, Plotly - Python Open Source Graphing Library

- Plotly makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts.

IV. Tools for DA

2. Libraries

d, Scikit-learn – Machine Learning and Data Mining

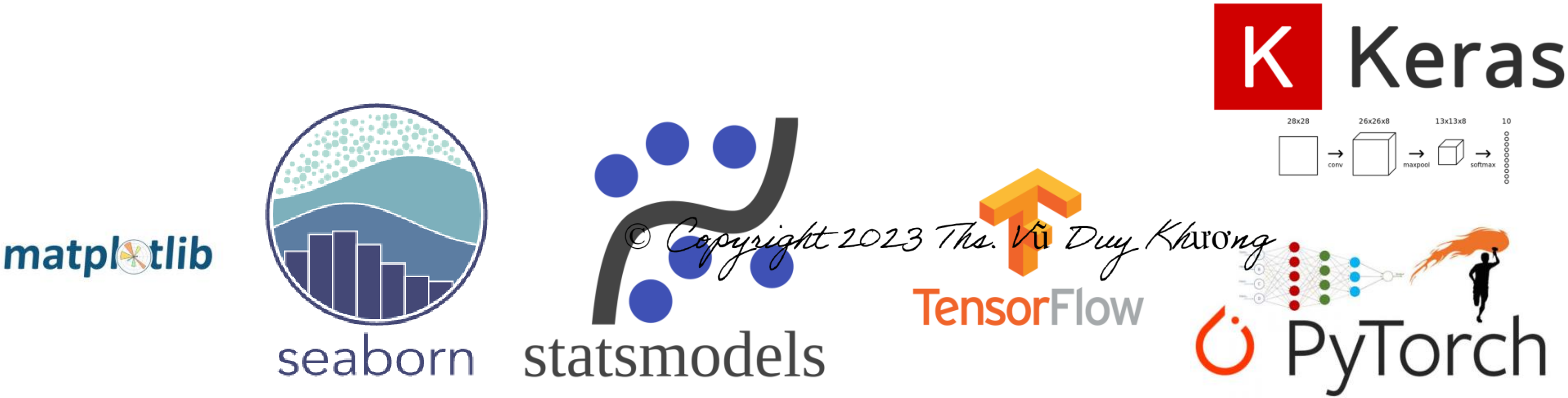
- This library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensional reduction algorithms.

IV. Tools for DA

2. Libraries

e, Others

- Matplotlib, Seaborn - for Data Visualization.
- StatsModels – for Statistical Modeling, Testing, and Analysis.
- Tensorflow, Keras, Pytorch - for building AI/ML models.



IV. Tools for DA

3. Open source tools

a, Github

- GitHub is a service that provides a web-based Git source code repository for software development projects.
- Currently, there are over 73 million developers who use github frequently.

IV. Tools for DA

3. Open source tools

b, Jupyter Notebook

- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

IV. Tools for DA

3. Open source tools

c, *Power BI*

- Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights.
- Your data may be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses.
- Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.

IV. Tools for DA

3. Open source tools

d, Others

- Visual Studio Code
- Pycharm
- RStudio
- Excel

IV. Tools for DA

4. Virtual Environment

a, Anaconda - the birthplace of Python data science

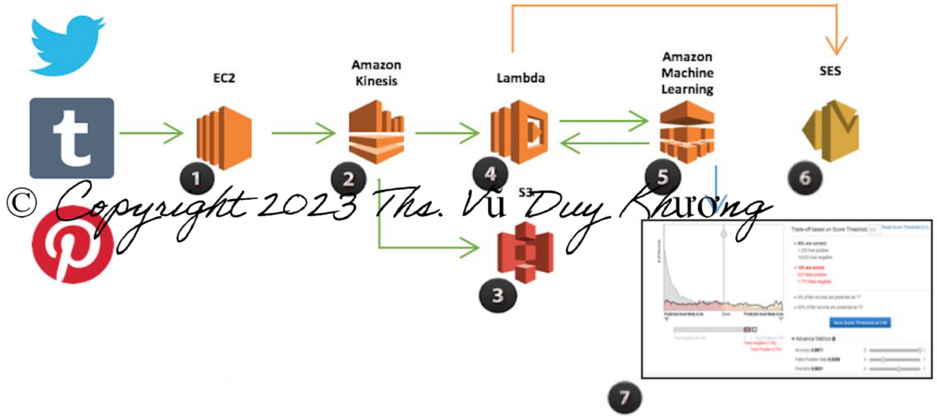
- Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

IV. Tools for DA

5. Commercial Tools

a, AWS - Amazon Web Services

- AWS is a comprehensive, evolving cloud computing platform provided by Amazon that includes a mixture of infrastructure as a service (IaaS), platform as a service (PaaS) and packaged software as a service (SaaS) offerings. AWS services can offer an organization tools such as compute power, database storage and content delivery services.

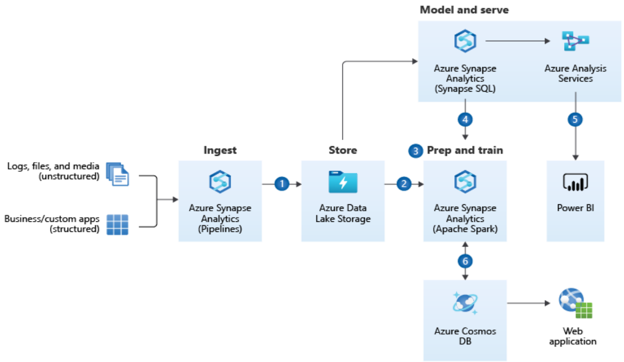


IV. Tools for DA

5. Commercial Tools

b, Azure

- Microsoft Azure is a public cloud platform with more than 200 products and services accessible over the public internet.
- Like other public cloud vendors, Azure manages and maintains hardware, infrastructure, and resources that can be accessed for free or pay-per-use, on-demand basis.

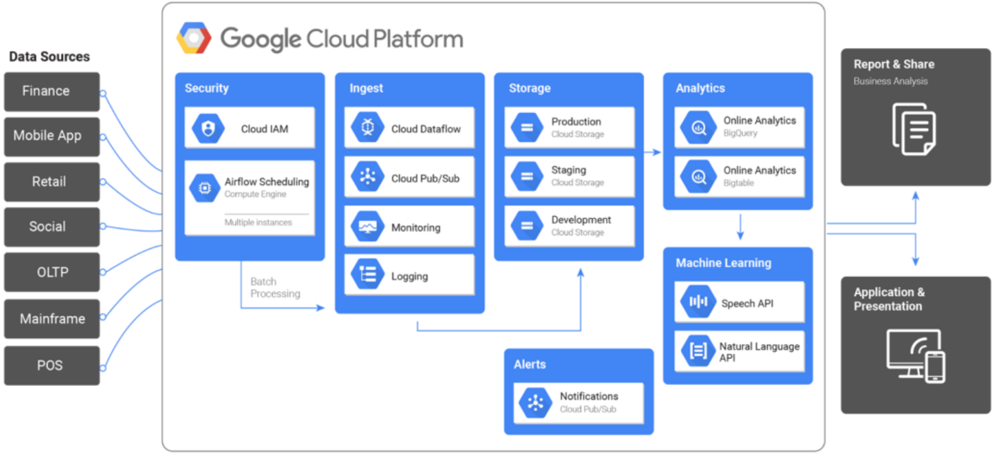

Model and serve
Azure Synapse Analytics (Synapse SQL)
Azure Analysis Services
4
5
Logs, files, and media (unstructured)
Business/custom apps (structured)
Ingest
Azure Synapse Analytics (Pipelines)
1
Store
Azure Data Lake Storage
2
3 Prep and train
Azure Synapse Analytics (Apache Spark)
Power BI
6
Azure Cosmos DB
Web application

IV. Tools for DA

5. Commercial Tools

c, GCP - Google Cloud Platform

- The Google Cloud Platform is a platform that delivers over 90 information technology services (aka products), which businesses, IT professionals, and developers can leverage to work more efficiently, gain more flexibility, and/or enable a strategic advantage.



V. Basic of Python for DA

1. Variables and Data Types

- Variable Assignment

```
>>> x=5
>>> x
5
```

- Calculations With Variables

```
>>> x+2 # Sum of two variables
7
>>> x-2 # Subtraction of two variables
3
>>> x*2 # Multiplication of two variables
10
>>> x**2 # Exponentiation of a variable
25
>>> x%2 # Remainder of a variable
1
>>> x/float(2) # Division of a variable
2.5
```

V. Basic of Python for DA

1. Variables and Data Types

- Types and Type Conversion

str()	'5', '3.45', 'True'	Variables to strings
int()	5, 3, 1	Variables to integers
float()	5.0, 1.0	Variables to floats
bool()	True, True, True	Variables to booleans

V. Basic of Python for DA

2. Strings

- Define **string** variable:

```
>>> my_string = 'thisStringIsAwesome'
>>> my_string
'thisStringIsAwesome'
```

- String Operations

```
>>> my_string * 2
'thisStringIsAwesomethisStringIsAwesome'
>>> my_string + 'Innit'
'thisStringIsAwesomeInnit'
>>> 'm' in my_string
True
>>> my_string[3]
>>> my_string[4:9]
```

V. Basic of Python for DA

2. Strings

- String methods

>>> my_string.upper()	String to uppercase
>>> my_string.lower()	String to lowercase
>>> my_string.count('w')	Count String elements
>>> my_string.replace('e', 'i')	Replace String elements
>>> my_string.strip()	Strip whitespaces

V. Basic of Python for DA

3. Lists

- Define **list** variable:

```
>>> a = 'is'
>>> b = 'nice'
>>> my_list = ['my', 'list', a, b]
>>> my_list2 = [[4,5,6,7], [3,4,5,6]]
```

V. Basic of Python for DA

3. Lists

- Selecting List Elements

Subset	
>>> my_list[1]	Select item at index 1
>>> my_list[-3]	Select 3rd last item
Slice	
>>> my_list[1:3]	Select items at index 1 and 2
>>> my_list[1:]	Select items after index 0
>>> my_list[:3]	Select items before index 3
>>> my_list[:]	Copy my_list
Subset Lists of Lists	my_list[list][itemOfList]
>>> my_list2[1][0]	
>>> my_list2[1][:2]	

** Index starts at 0*

V. Basic of Python for DA

3. Lists

- List Operations

```
>>> my_list + my_list
['my', 'list', 'is', 'nice', 'my', 'list', 'is', 'nice']
>>> my_list * 2
['my', 'list', 'is', 'nice', 'my', 'list', 'is', 'nice']
>>> my_list2 > 4
True
```

** Index starts at 0*

V. Basic of Python for DA

3. Lists

- List Methods

>>> my_list.index(a)	Get the index of an item
>>> my_list.count(a)	Count an item
>>> my_list.append('!')	Append an item at a time
>>> my_list.remove('!')	Remove an item
>>> del(my_list[0:1])	Remove an item
>>> my_list.reverse()	Reverse the list
>>> my_list.extend('!')	Append an item
>>> my_list.pop(-1)	Remove an item
>>> my_list.insert(0,'!')	Insert an item
>>> my_list.sort()	Sort the list

** Index starts at 0*

V. Basic of Python for DA

4. Libraries

- Import libraries

```
>>> import numpy
>>> import numpy as np
```

- Selective import

```
>>> from math import pi
>>> from matplotlib import pyplot as plt
```

VI. Reference

Book:

https://automatetheboringstuff.com/

Q & A