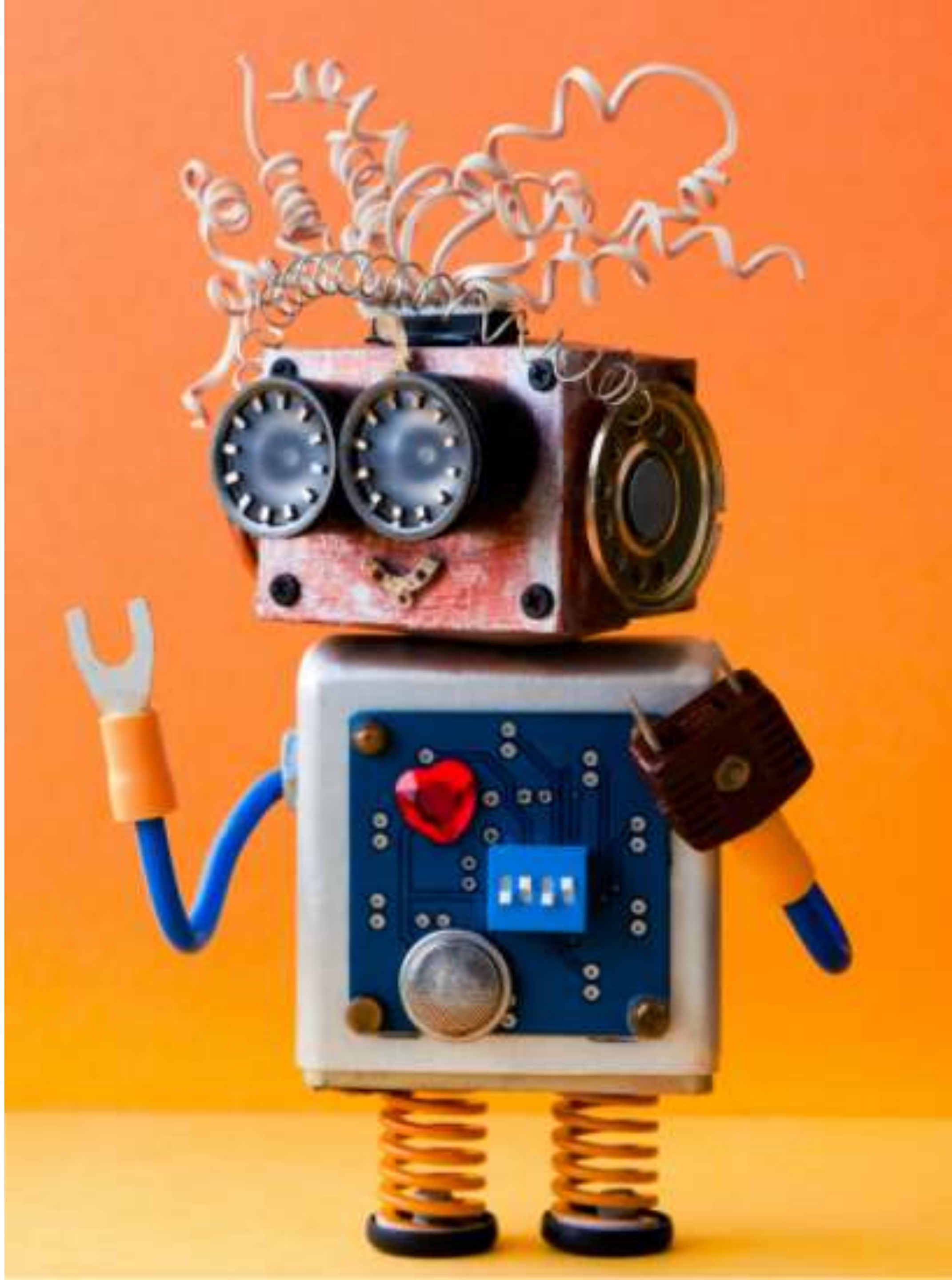
INTRODUCTION TO EXPLORATORY DATA ANALYSIS (EDA)

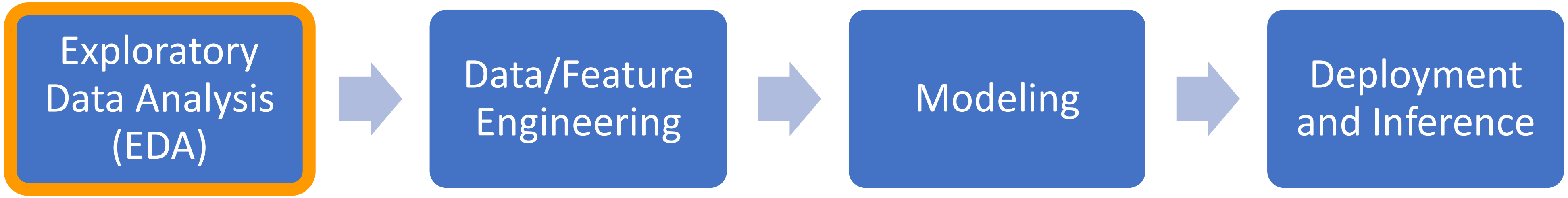
EASY

ADVANCED



EXPLORATORY DATA ANALYSIS (EDA)

- Exploratory Data Analysis (EDA) is a process used by data scientists to analyze data and gain valuable insights.
- EDA empowers data scientists to gain better understanding of the data, detect patterns, and identify outliers.
- EDA tools work by generating statistical summary (Minimum, Maximum, Mean, and Count) and perform data visualizations.
- EDA is the first step in developing any machine learning workflow.
- Once EDA is complete, data can proceed to the next step which is data engineering.



DATA VISUALIZATION

- In order to perform data visualization, there are generally two approaches: (1) Use Developer tools or (2) use Business intelligence tools.

DEVELOPER TOOLS

Jupyter Notebook

The Jupyter Notebook is a web-based interactive computing platform that allows users to author data- and code-driven narratives that combine live code, equations, narrative text, visualizations, interactive dashboards and other media.

AMAZON SAGEMAKER

BUSINESS INTELLIGENCE TOOLS

AMAZON QUICKSIGHT

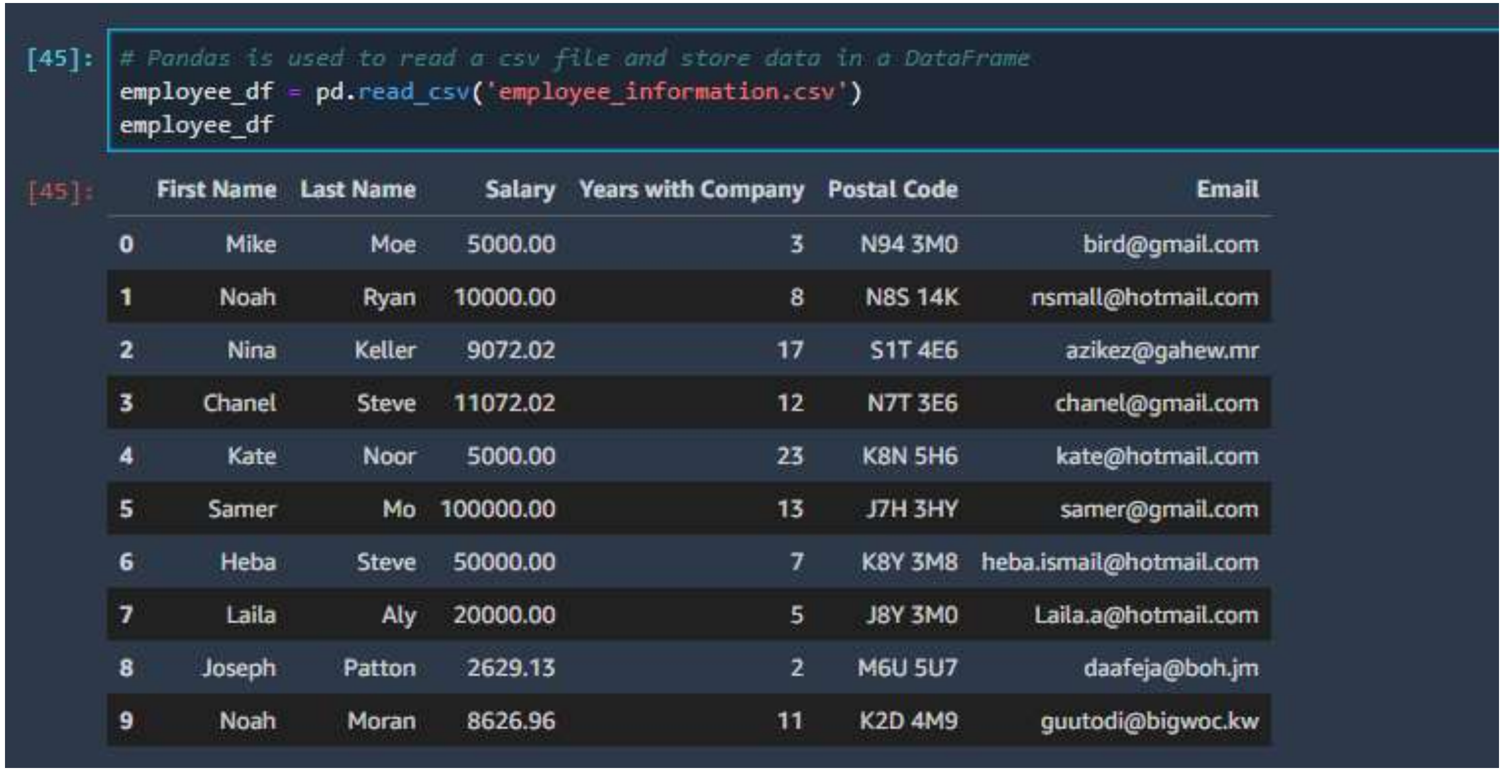
TABLEAU

POWER BI

PANDAS LIBRARY 101

- Pandas is an open source library that offers high-performance data structures and data analysis tools in python.
- Data can also be stored using pandas DataFrame.
- Think of it as using Microsoft excel in python/Jupyter environment.

THIS IS WHAT PANDAS DATAFRAME LOOK LIKE! IT'S A MULTI-DIMENSIONAL TABLE



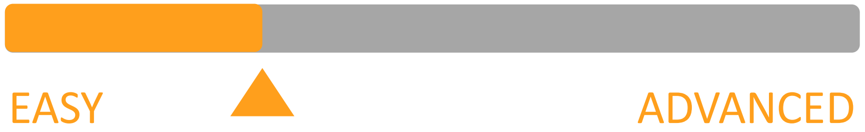
[45]:

```
# Pandas is used to read a csv file and store data in a DataFrame
employee_df = pd.read_csv('employee_information.csv')
employee_df
```

[45]:

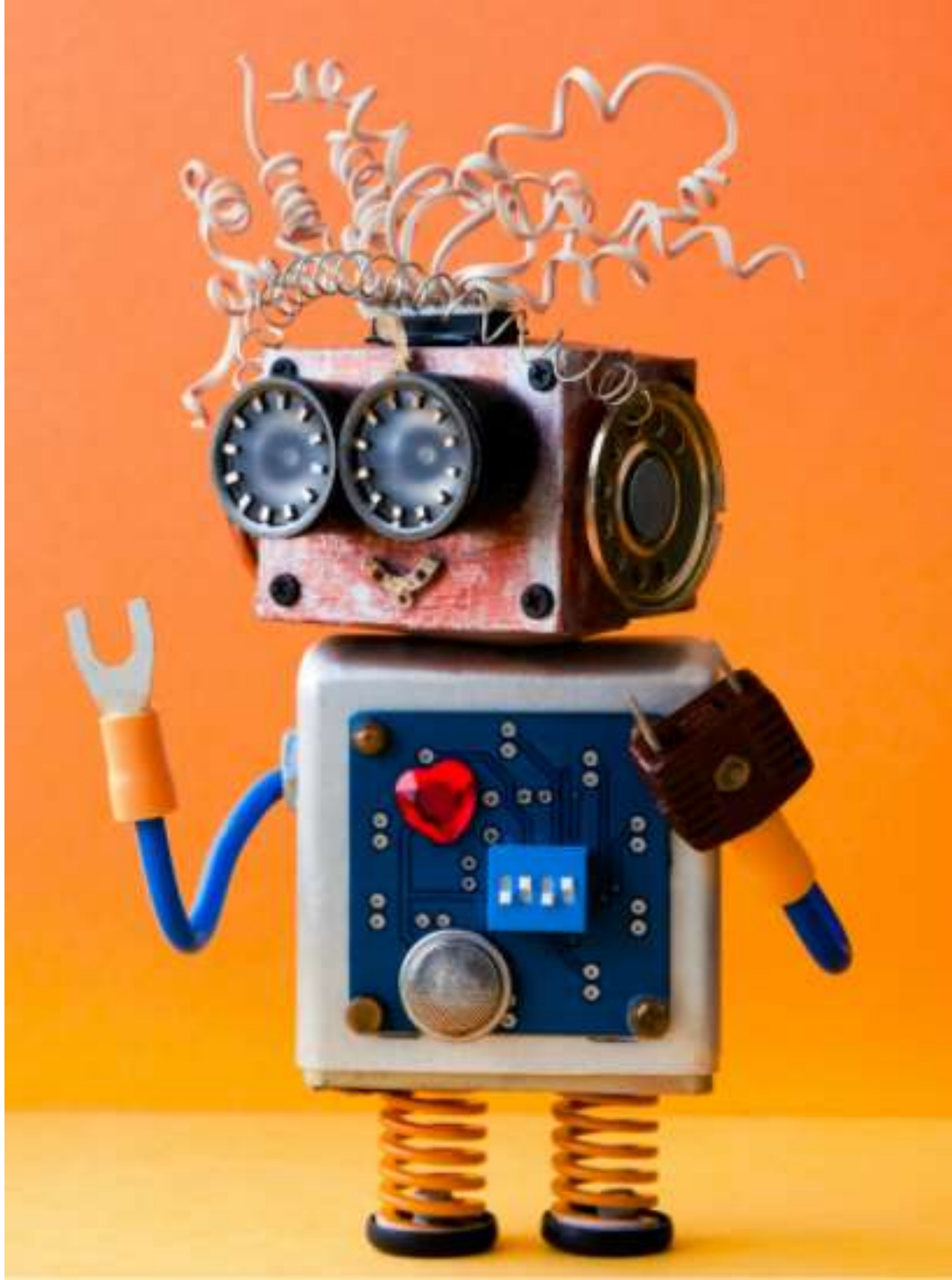
	First Name	Last Name	Salary	Years with Company	Postal Code	Email
0	Mike	Moe	5000.00	3	N94 3M0	bird@gmail.com
1	Noah	Ryan	10000.00	8	N8S 14K	nsmall@hotmail.com
2	Nina	Keller	9072.02	17	S1T 4E6	azikez@gahew.mr
3	Chanel	Steve	11072.02	12	N7T 3E6	chanel@gmail.com
4	Kate	Noor	5000.00	23	K8N 5H6	kate@hotmail.com
5	Samer	Mo	100000.00	13	J7H 3HY	samer@gmail.com
6	Heba	Steve	50000.00	7	K8Y 3M8	heba.ismail@hotmail.com
7	Laila	Aly	20000.00	5	J8Y 3M0	Laila.a@hotmail.com
8	Joseph	Patton	2629.13	2	M6U 5U7	daafeja@boh.jm
9	Noah	Moran	8626.96	11	K2D 4M9	guutodi@bigwoc.kw

PROJECT OVERVIEW



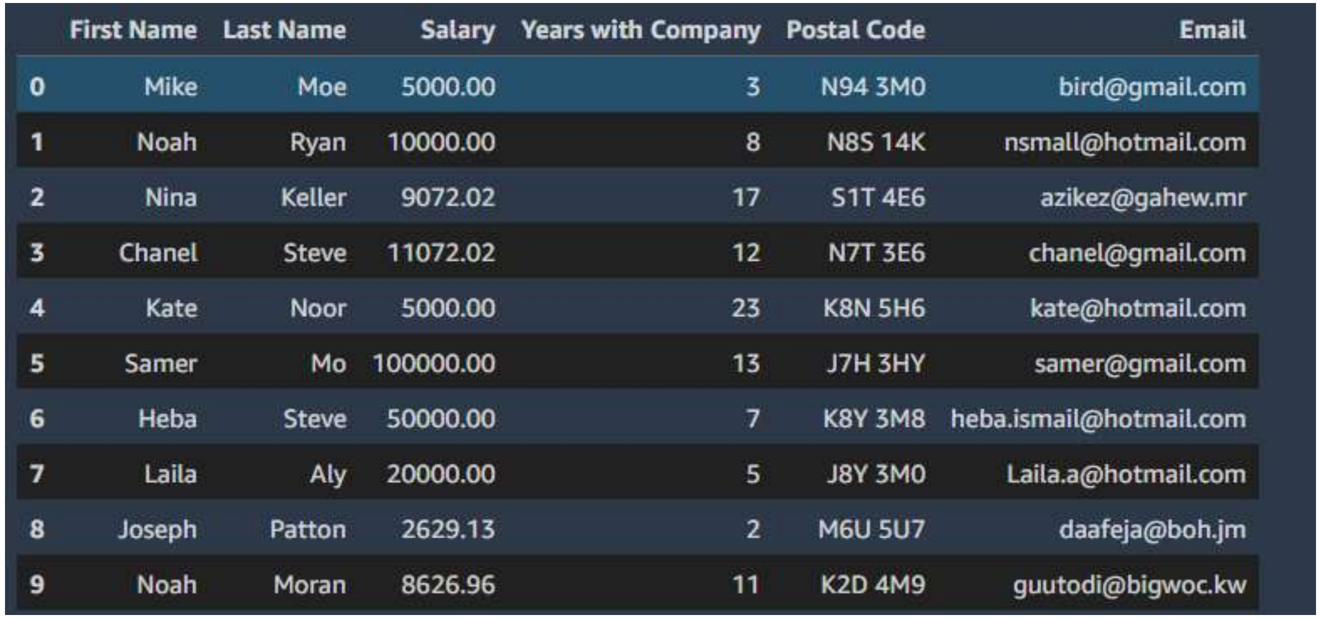
EASY

ADVANCED



PROJECT OVERVIEW

- We will analyze corporate employee information using Pandas in Jupyter Notebooks in AWS SageMaker Studio.
- We will learn how to:
  1. Define a pandas Dataframe
  2. Read CSV Data using Pandas
  3. Perform basic statistical analysis on the data
  4. Set/Reset Pandas DataFrame Index
- In the final project, you will perform basic EDA on a brand new dataset.

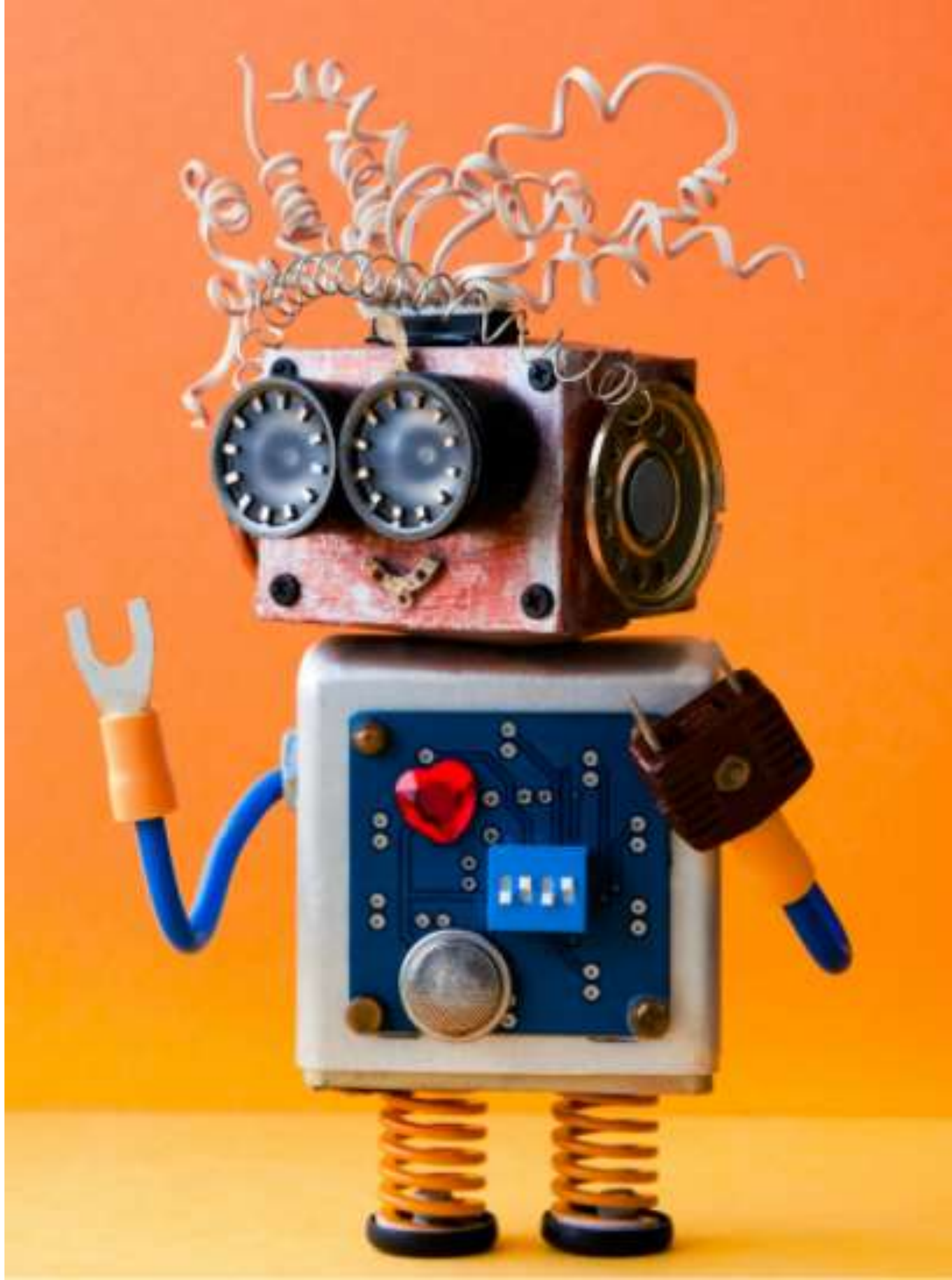


	First Name	Last Name	Salary	Years with Company	Postal Code	Email
0	Mike	Moe	5000.00	3	N94 3M0	bird@gmail.com
1	Noah	Ryan	10000.00	8	N8S 14K	nsmall@hotmail.com
2	Nina	Keller	9072.02	17	S1T 4E6	azikez@gahew.mr
3	Chanel	Steve	11072.02	12	N7T 3E6	chanel@gmail.com
4	Kate	Noor	5000.00	23	K8N 5H6	kate@hotmail.com
5	Samer	Mo	100000.00	13	J7H 3HY	samer@gmail.com
6	Heba	Steve	50000.00	7	K8Y 3M8	heba.ismail@hotmail.com
7	Laila	Aly	20000.00	5	J8Y 3M0	Laila.a@hotmail.com
8	Joseph	Patton	2629.13	2	M6U 5U7	daafeja@boh.jm
9	Noah	Moran	8626.96	11	K2D 4M9	guutodi@bigwoc.kw

AMAZON SAGEMAKER STUDIO SETUP

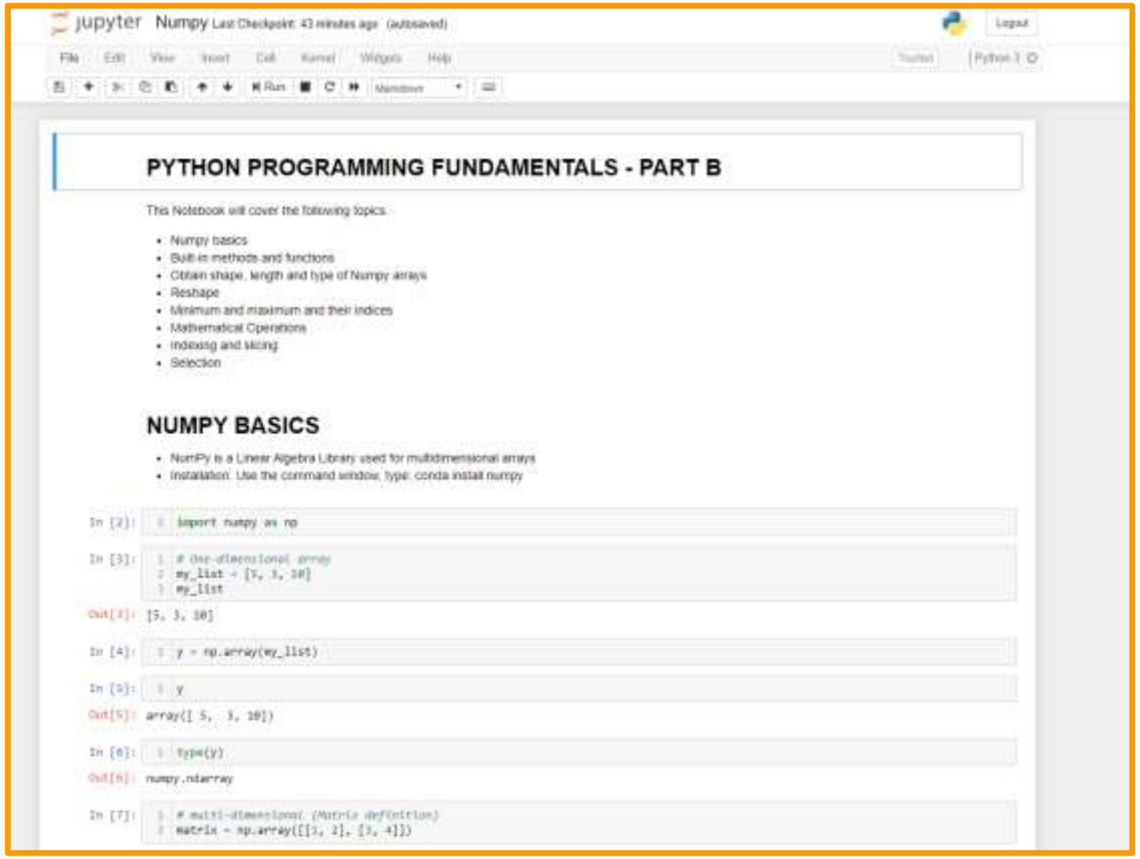
EASY

ADVANCED



JUPYTER NOTEBOOKS

- Jupyter Notebooks are open-source web application that enable developers to develop and distribute codes, text, equations, and figures in one place.
- It’s one of the top tools used by machine learning developers.
- In Jupyter notebooks, you can write in 40 programming languages such as Python, R, and Scala.
- You can Share notebooks including code results with other.
- https://jupyter.org/



PYTHON PROGRAMMING FUNDAMENTALS - PART B

This Notebook will cover the following topics:

- Numpy basics
- Built-in methods and functions
- Obtain shape, length and type of Numpy arrays
- Reshape
- Minimum and maximum and their indices
- Mathematical Operations
- Indexing and slicing
- Selection

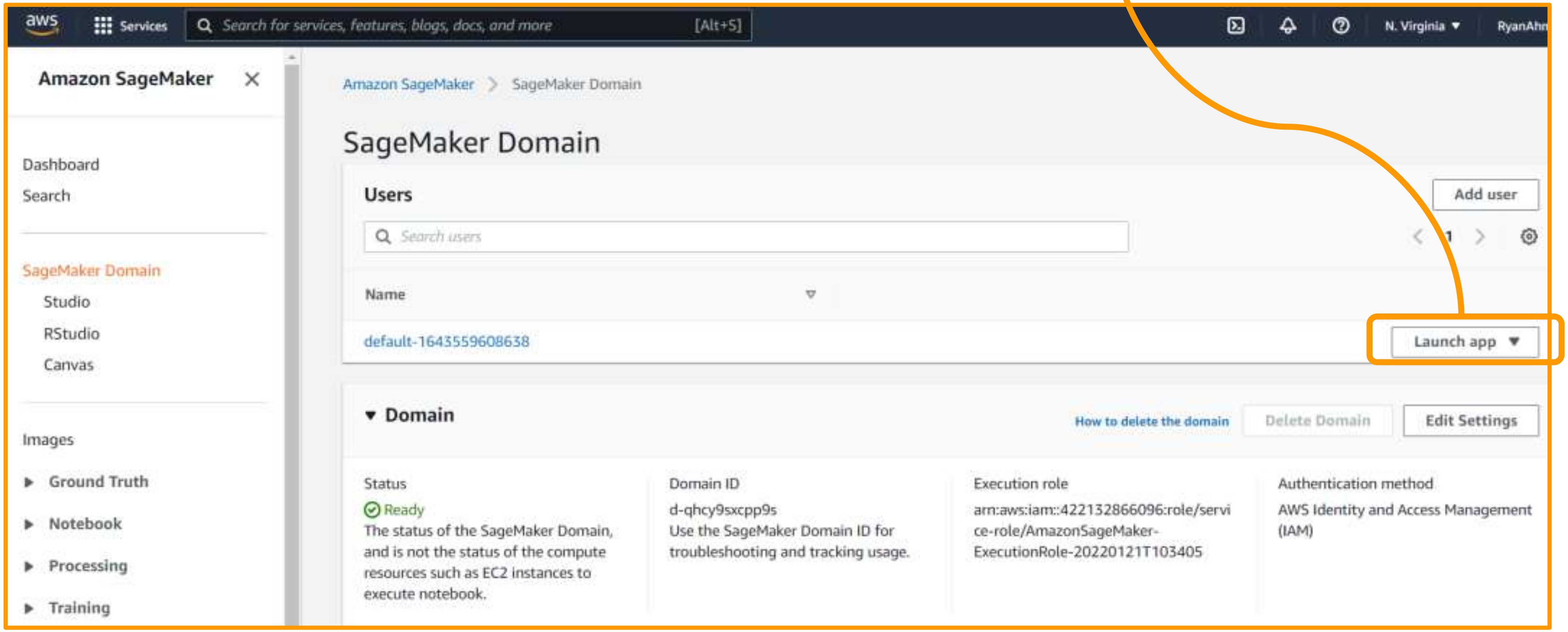
NUMPY BASICS

- NumPy is a Linear Algebra Library used for multidimensional arrays
- Installation: Use the command window, type: conda install numpy

```
In [2]:  import numpy as np

In [3]:  # One-dimensional array
         my_list = [5, 3, 10]
         my_list
Out[3]:  [5, 3, 10]

In [4]:  y = np.array(my_list)

In [5]:  y
Out[5]:  array([ 5,  3, 10])

In [6]:  type(y)
Out[6]:  numpy.ndarray

In [7]:  # multi-dimensional (Matrix definition)
         matrix = np.array([[1, 2], [3, 4]])
```

JUPYTER NOTEBOOKS IN SAGEMAKER STUDIO

LAUNCH SAGEMAKER STUDIO



SageMaker Domain

Users

Name
default-1643559608638

▼ Domain

Status
Ready
The status of the SageMaker Domain, and is not the status of the compute resources such as EC2 instances to execute notebook.

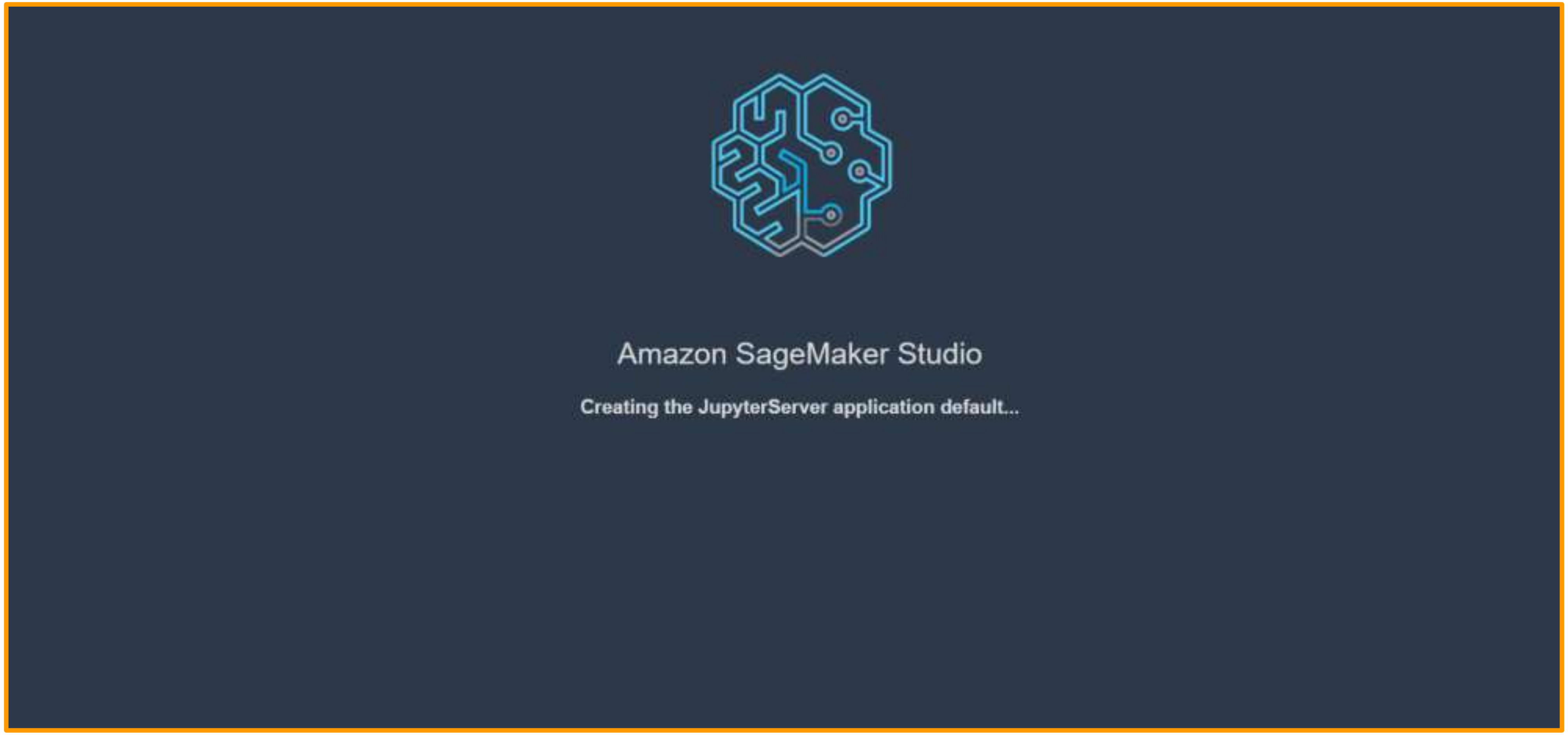
Domain ID
d-qhcy9sxcpp9s
Use the SageMaker Domain ID for troubleshooting and tracking usage.

Execution role
arn:aws:iam::422132866096:role/service-role/AmazonSageMaker-ExecutionRole-20220121T103405

Authentication method
AWS Identity and Access Management (IAM)

JUPYTER NOTEBOOKS IN SAGEMAKER STUDIO

YOU SHOULD SEE THIS SCREEN!

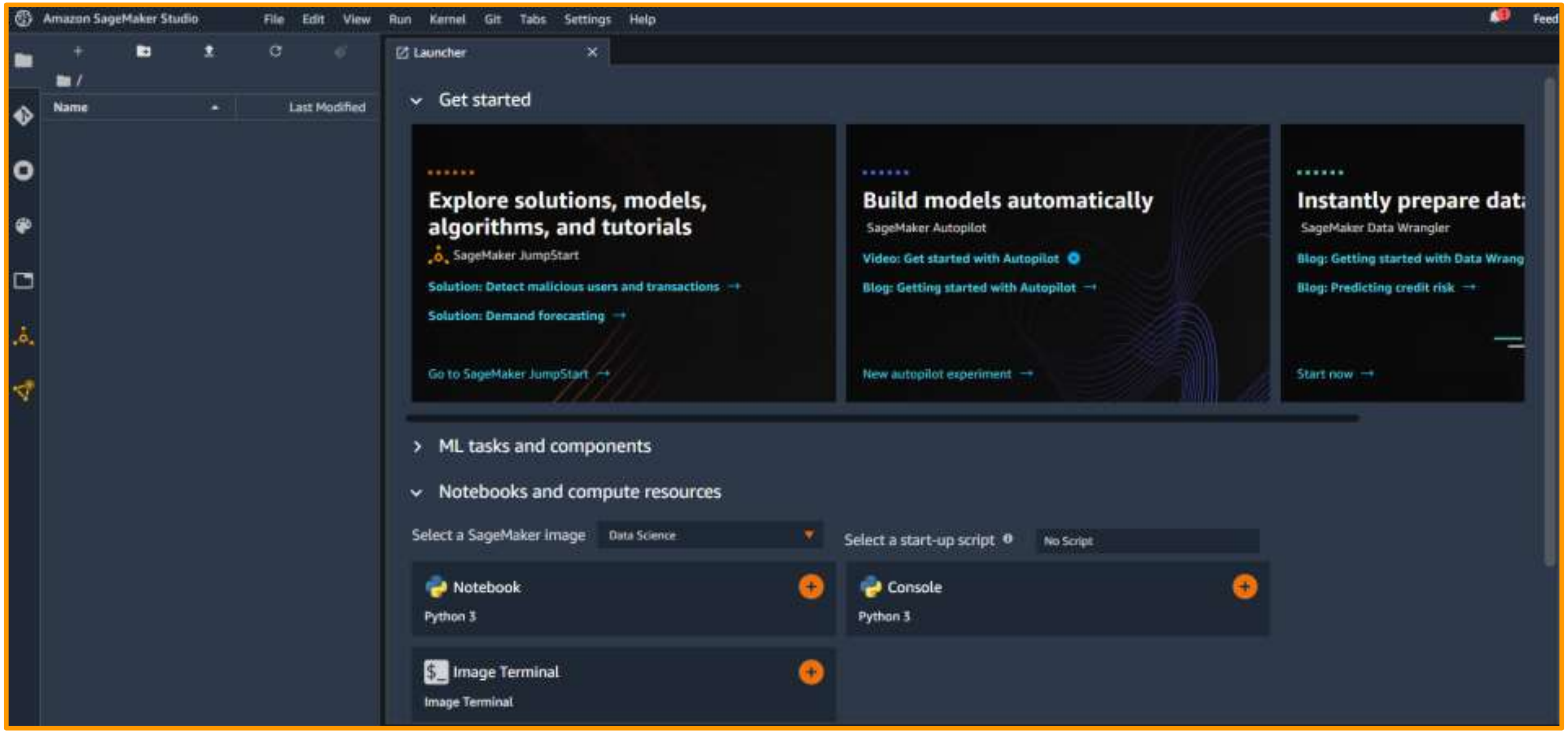


Amazon SageMaker Studio

Creating the JupyterServer application default...

JUPYTER NOTEBOOKS IN SAGEMAKER STUDIO

WELCOME TO SAGEMAKER STUDIO HOME PAGE!

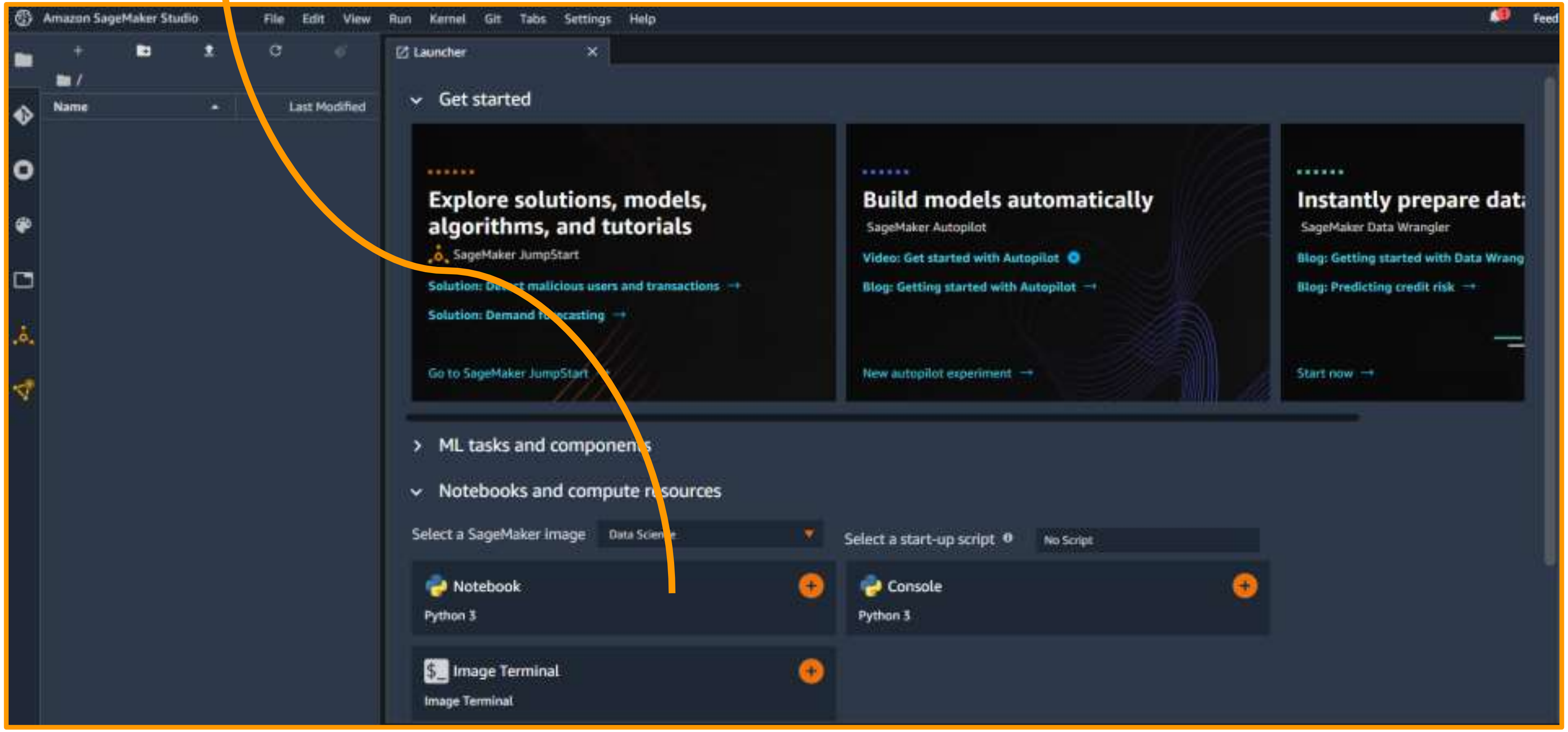

Amazon SageMaker Studio
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Feed
/
Name
Last Modified
Launcher
Get started
Explore solutions, models, algorithms, and tutorials
SageMaker JumpStart
Solution: Detect malicious users and transactions →
Solution: Demand forecasting →
Go to SageMaker JumpStart →
Build models automatically
SageMaker Autopilot
Video: Get started with Autopilot
Blog: Getting started with Autopilot →
New autopilot experiment →
Instantly prepare dat
SageMaker Data Wrangler
Blog: Getting started with Data Wrang
Blog: Predicting credit risk →
Start now →
ML tasks and components
Notebooks and compute resources
Notebook
Python 3
Console
Python 3
Image Terminal
Image Terminal

Select a SageMaker image: Data Science

Select a start-up script: No Script

JUPYTER NOTEBOOKS IN SAGEMAKER STUDIO

CLICK ON NOTEBOOK (PYTHON 3) OR UPLOAD TO CREATE A BLANK NEW JUPYTER NOTEBOOK



Get started

Explore solutions, models, algorithms, and tutorials

SageMaker JumpStart

Solution: Detect malicious users and transactions →

Solution: Demand forecasting →

Go to SageMaker JumpStart →

Build models automatically

SageMaker Autopilot

Video: Get started with Autopilot

Blog: Getting started with Autopilot →

New autopilot experiment →

Instantly prepare dat

SageMaker Data Wrangler

Blog: Getting started with Data Wrang

Blog: Predicting credit risk →

Start now →

ML tasks and components

Notebooks and compute resources

Select a SageMaker image: Data Science

Select a start-up script: No Script

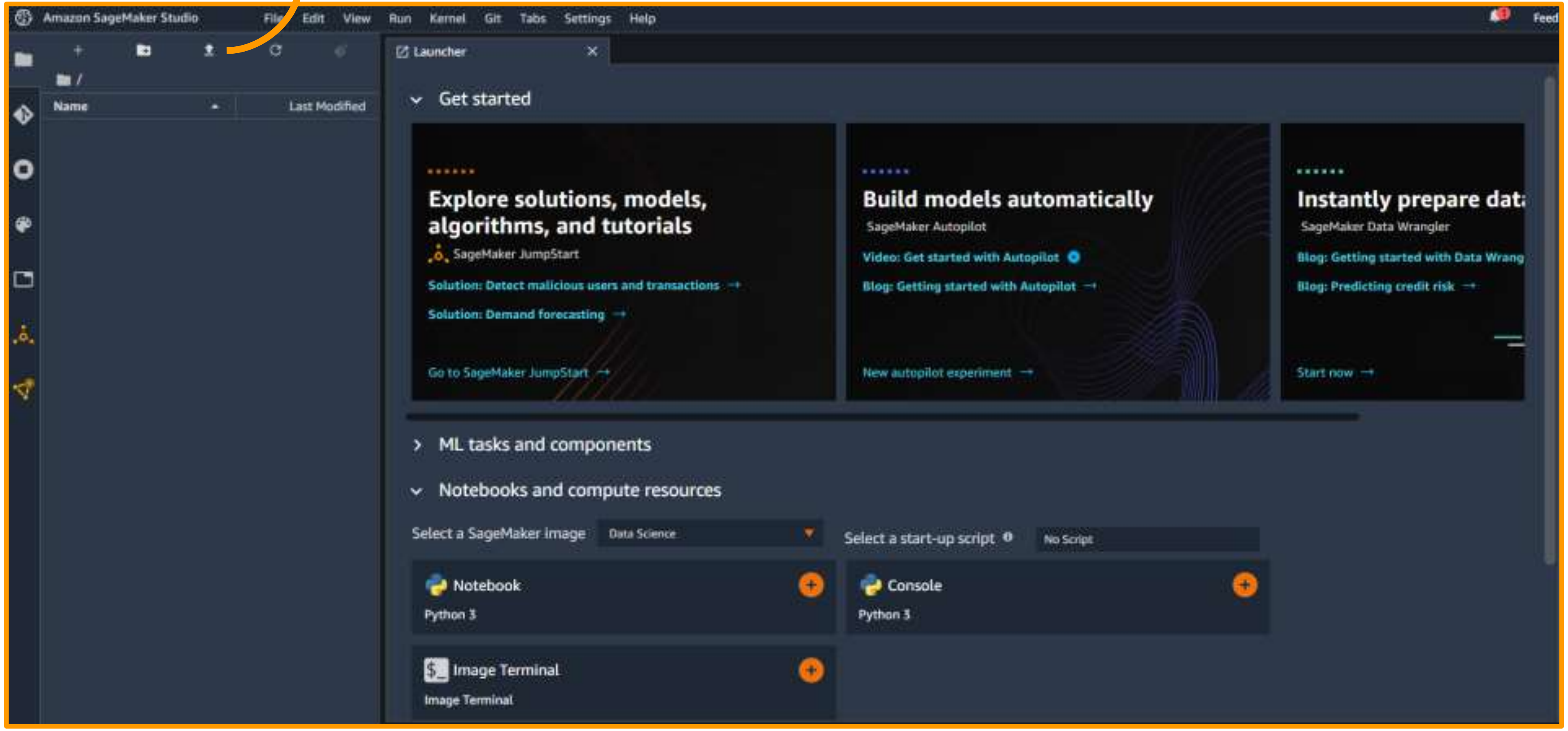
Notebook
Python 3

Console
Python 3

Image Terminal
Image Terminal

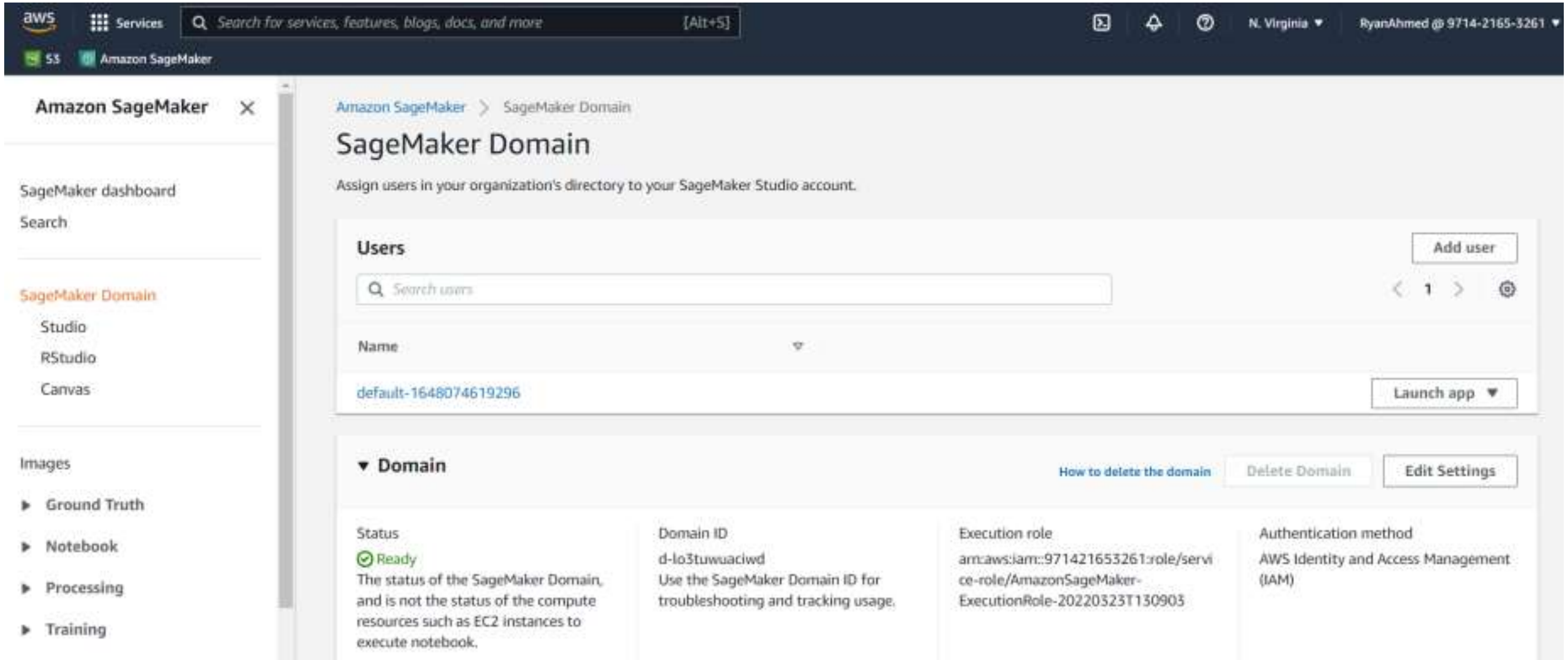
JUPYTER NOTEBOOKS IN SAGEMAKER STUDIO

ALTERNATIVELY, YOU CAN CLICK ON THE UPLOAD BUTTON TO LOAD YOUR OWN NOTEBOOK


Amazon SageMaker Studio
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Feed
Name
Last Modified
Launcher
Get started
Explore solutions, models, algorithms, and tutorials
SageMaker JumpStart
Solution: Detect malicious users and transactions →
Solution: Demand forecasting →
Go to SageMaker JumpStart →
Build models automatically
SageMaker Autopilot
Video: Get started with Autopilot
Blog: Getting started with Autopilot →
New autopilot experiment →
Instantly prepare dat
SageMaker Data Wrangler
Blog: Getting started with Data Wrang
Blog: Predicting credit risk →
Start now →
ML tasks and components
Notebooks and compute resources
Select a SageMaker image
Data Science
Select a start-up script
No Script
Notebook
Python 3
Console
Python 3
Image Terminal
Image Terminal

AWS SAGEMAKER SETUP

ALTERNATIVELY, WE CAN USE SAGEMAKER STUDIO. CLICK ON STUDIO AND CLICK ON LAUNCH APP>STUDIO



SageMaker Domain

Assign users in your organization's directory to your SageMaker Studio account.

Users

Name
default-1648074619296

Domain

Status
Ready
The status of the SageMaker Domain, and is not the status of the compute resources such as EC2 instances to execute notebook.

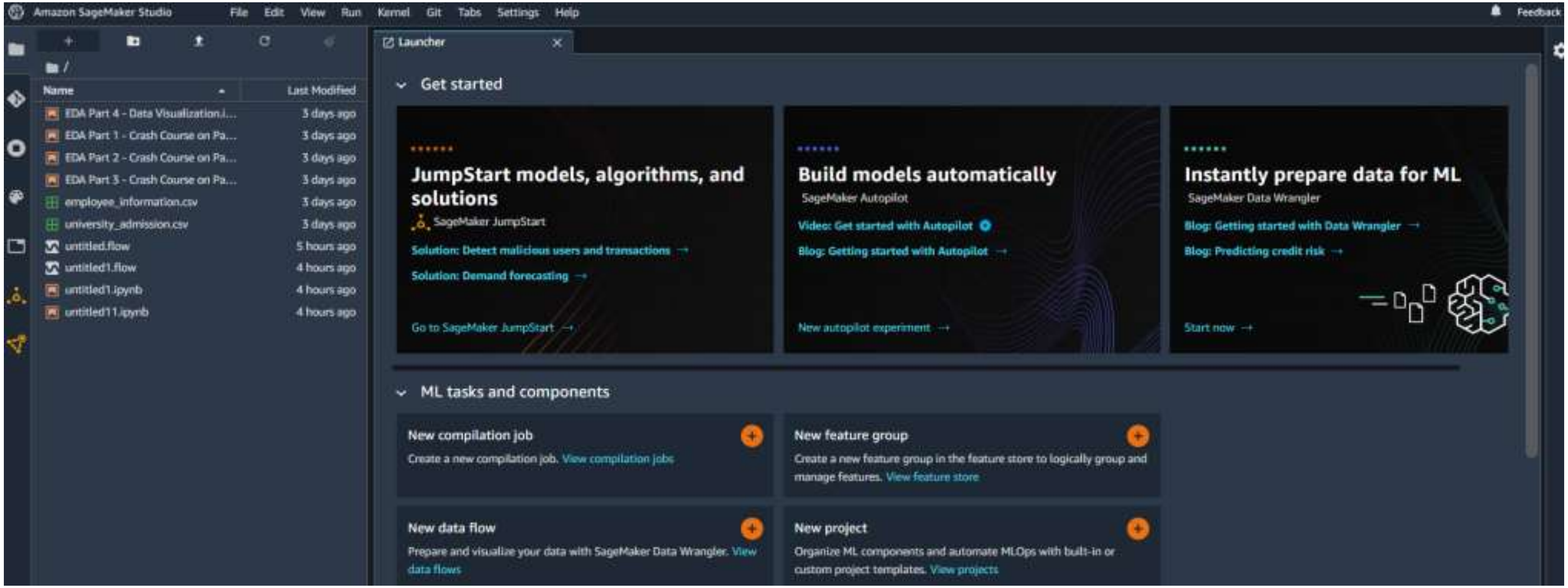
Domain ID
d-lo3tuwuaciwd
Use the SageMaker Domain ID for troubleshooting and tracking usage.

Execution role
arn:aws:iam::971421653261:role/service-role/AmazonSageMaker-ExecutionRole-20220323T130903

Authentication method
AWS Identity and Access Management (IAM)

AWS SAGEMAKER SETUP

CLICK ON UPLOAD AND SELECT THE DATASET AND JUPYTER NOTEBOOK



Name	Last Modified
EDA Part 4 - Data Visualization.i...	3 days ago
EDA Part 1 - Crash Course on Pa...	3 days ago
EDA Part 2 - Crash Course on Pa...	3 days ago
EDA Part 3 - Crash Course on Pa...	3 days ago
employee_information.csv	3 days ago
university_admission.csv	3 days ago
untitled.flow	5 hours ago
untitled1.flow	4 hours ago
untitled1.ipynb	4 hours ago
untitled11.ipynb	4 hours ago

Get started

JumpStart models, algorithms, and solutions

SageMaker JumpStart

Solution: Detect malicious users and transactions →

Solution: Demand forecasting →

Go to SageMaker JumpStart →

Build models automatically

SageMaker Autopilot

Video: Get started with Autopilot

Blog: Getting started with Autopilot →

New autopilot experiment →

Instantly prepare data for ML

SageMaker Data Wrangler

Blog: Getting started with Data Wrangler →

Blog: Predicting credit risk →

Start now →

ML tasks and components

New compilation job

Create a new compilation job. View compilation jobs

New feature group

Create a new feature group in the feature store to logically group and manage features. View feature store

New data flow

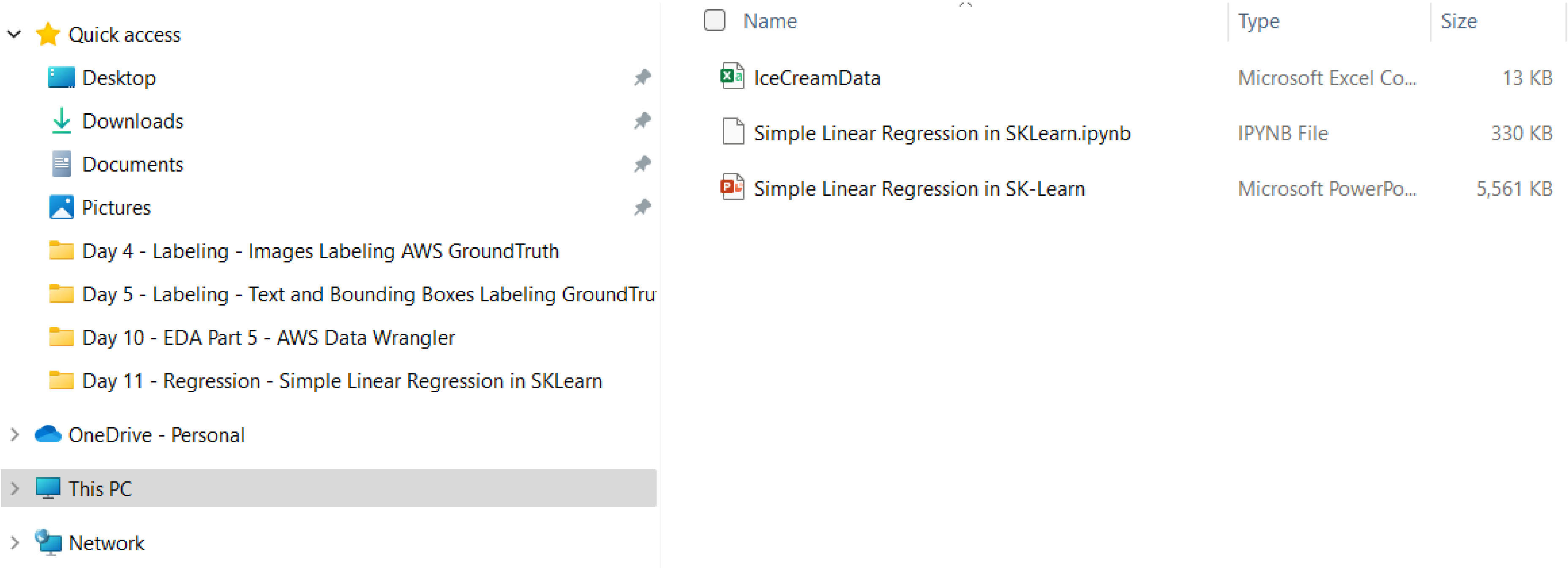
Prepare and visualize your data with SageMaker Data Wrangler. View data flows

New project

Organize ML components and automate MLOps with built-in or custom project templates. View projects

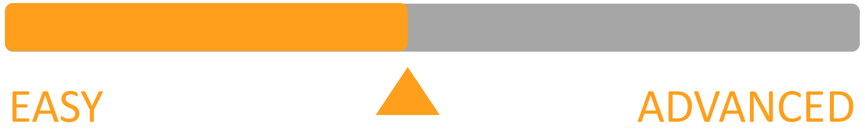
AWS SAGEMAKER SETUP

SELECT THE NOTEBOOK AND ICECREAMDATA


Quick access
Desktop
Downloads
Documents
Pictures
Day 4 - Labeling - Images Labeling AWS GroundTruth
Day 5 - Labeling - Text and Bounding Boxes Labeling GroundTru
Day 10 - EDA Part 5 - AWS Data Wrangler
Day 11 - Regression - Simple Linear Regression in SKLearn
OneDrive - Personal
This PC
Network

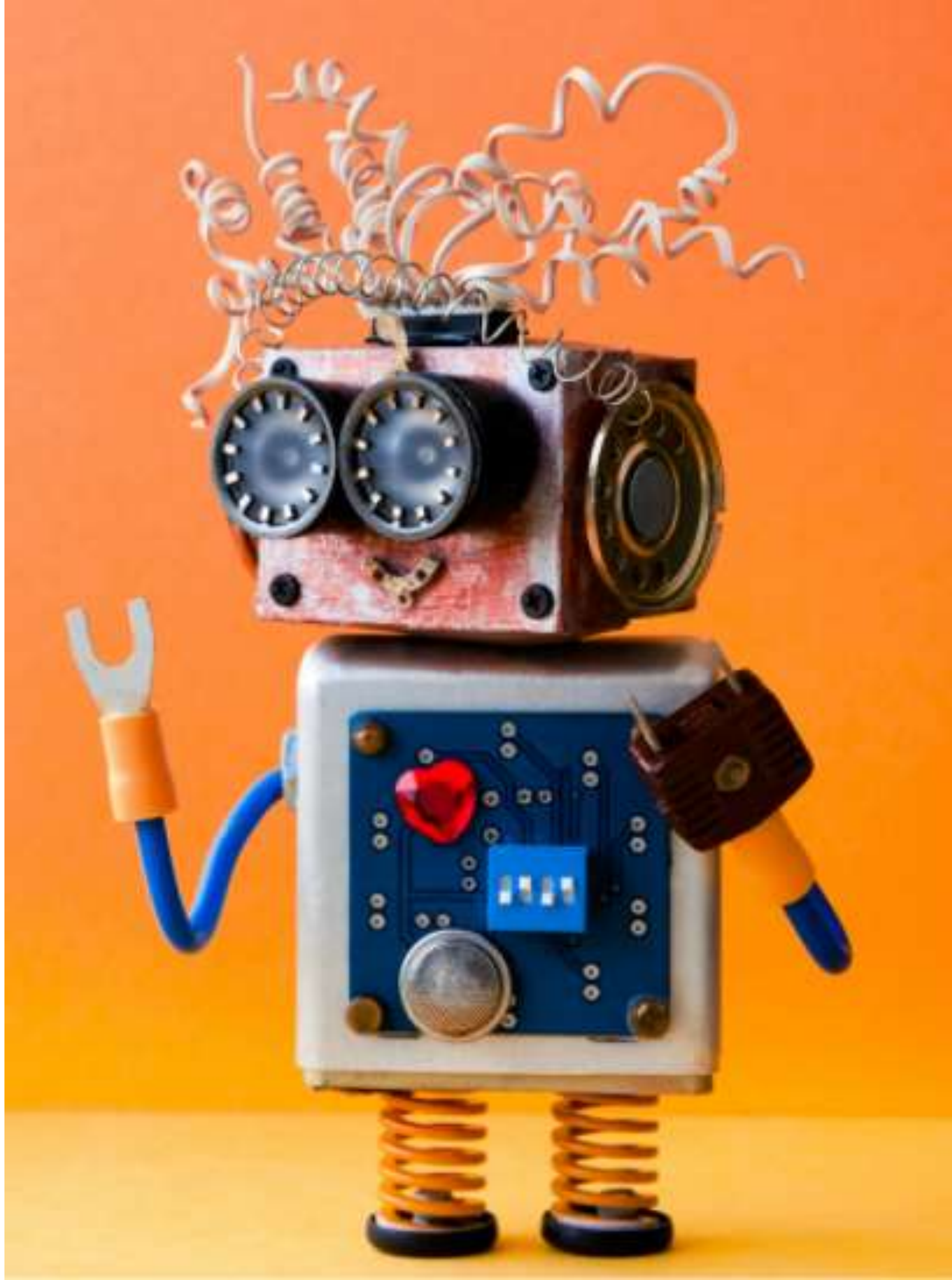
Name	Type	Size
IceCreamData	Microsoft Excel Co...	13 KB
Simple Linear Regression in SKLearn.ipynb	IPYNB File	330 KB
Simple Linear Regression in SK-Learn	Microsoft PowerPo...	5,561 KB

FINAL END-OF-DAY CAPSTONE PROJECT



EASY

ADVANCED



FINAL PROJECT

- In this project, we will perform basic Exploratory Data Analysis (EDA) on the University Admissions Dataset
- Columns definitions are as listed below:

GRE Scores (out of 340)
TOEFL Scores (out of 120)
University Rating (out of 5)
Statement of Purpose (SOP)
Letter of Recommendation (LOR) Strength (out of 5)
Undergraduate GPA (out of 10)
Research Experience (either 0 or 1)
Chance of admission (ranging from 0 to 1)

- Using the "university_admision.csv" included in the course package, write a python script to perform the following tasks:

1. Import the "university_admission.csv" file using Pandas
2. Display the first and last 8 rows in the DataFrame
3. Obtain the shape of the DataFrame
4. Calculate the average, min and max values for the LOR and SOP Columns
5. Use the GRE Score as the pandas dataframe index