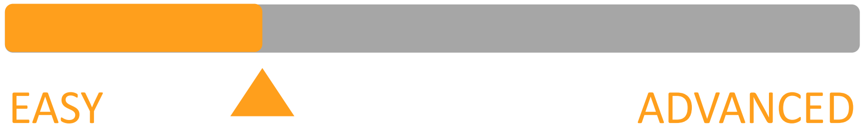
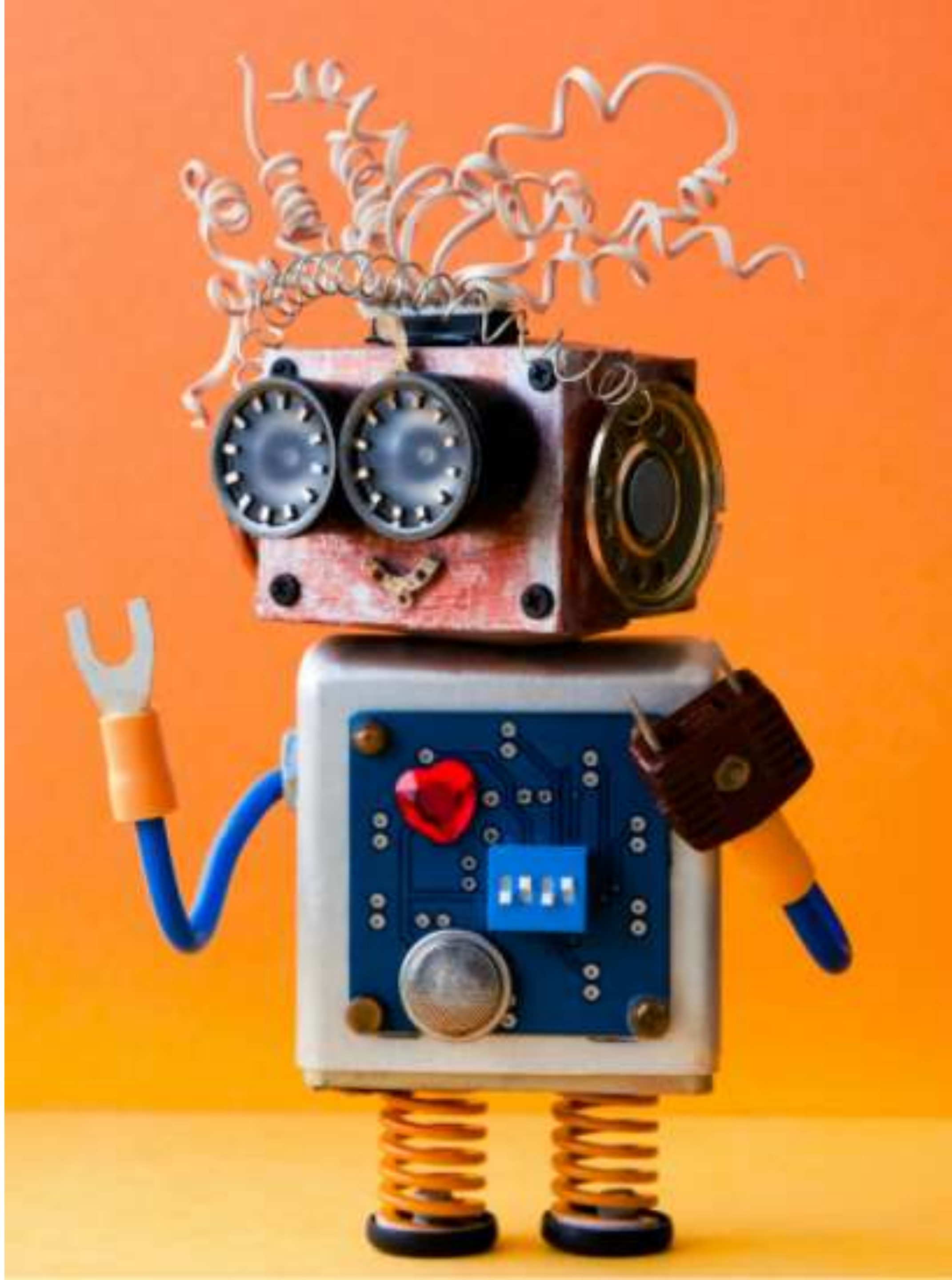
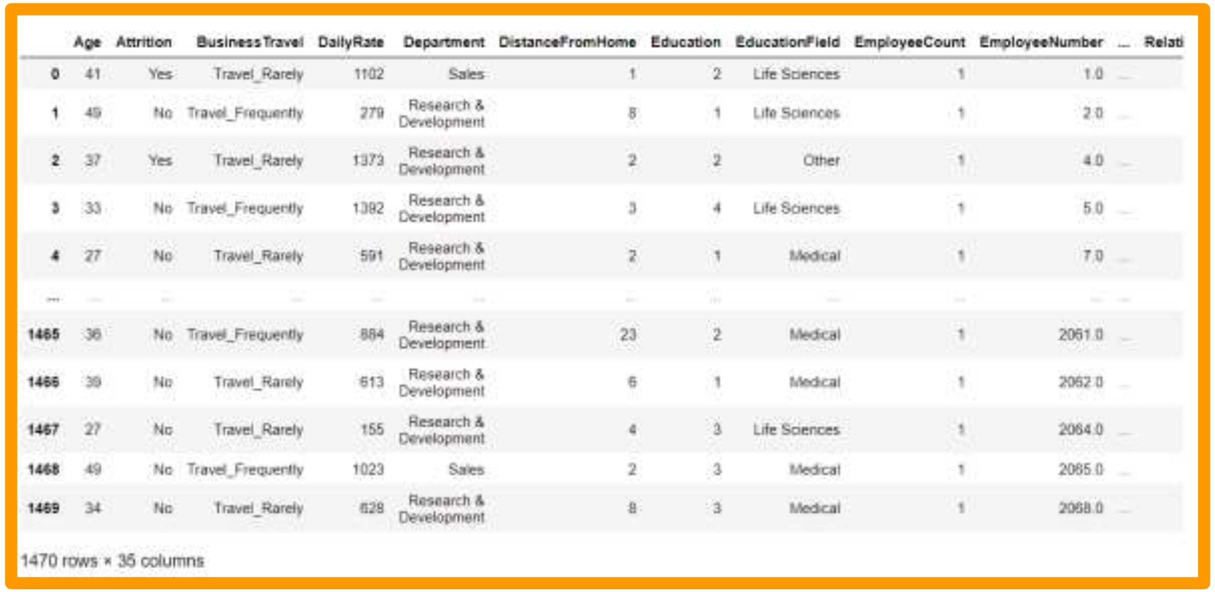
PROJECT OVERVIEW AND KEY LEARNING OUTCOMES





PROJECT OVERVIEW

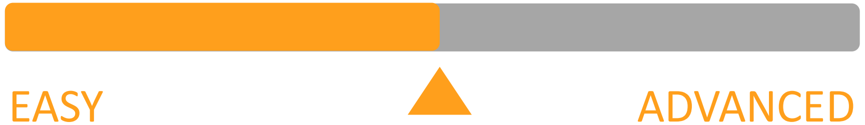
- We will analyze human resources information using Pandas in AWS SageMaker Studio.
- We will learn how to:
    1. Perform statistical analysis on real world datasets.
    2. Deal with missing data using pandas
    3. Change pandas DataFrame datatypes
    4. Define a function and apply it to a Pandas DataFrame column
    5. Pandas Operations and filtering
    6. Calculate and display correlation matrix
    7. Use seaborn library to show heatmap



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	Relati
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1.0	...	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2.0	...	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4.0	...	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5.0	...	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
1465	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	2061.0	...	
1466	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	2062.0	...	
1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	2064.0	...	
1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	2065.0	...	
1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	2068.0	...	

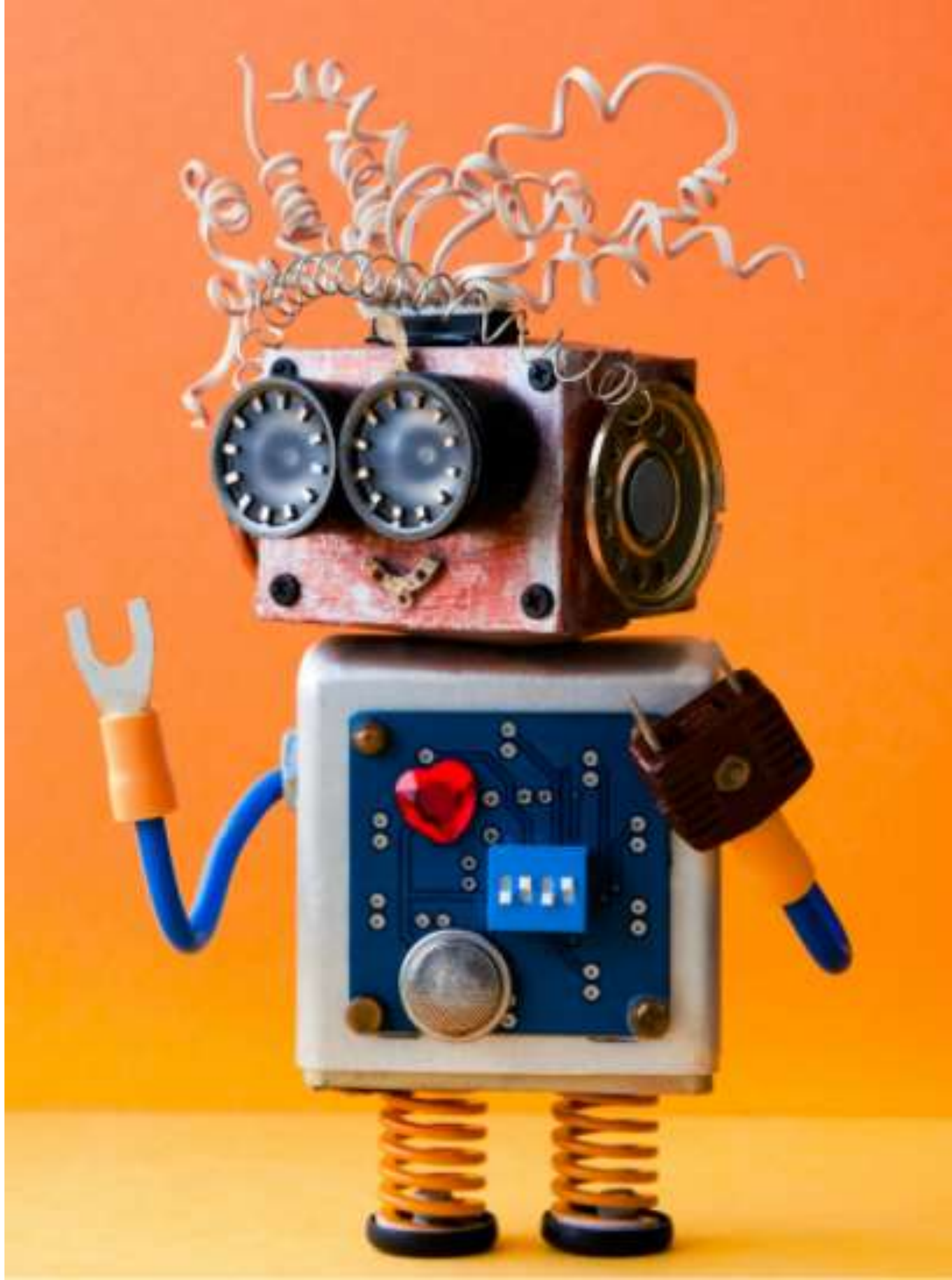
1470 rows × 35 columns

PROJECT DEMO

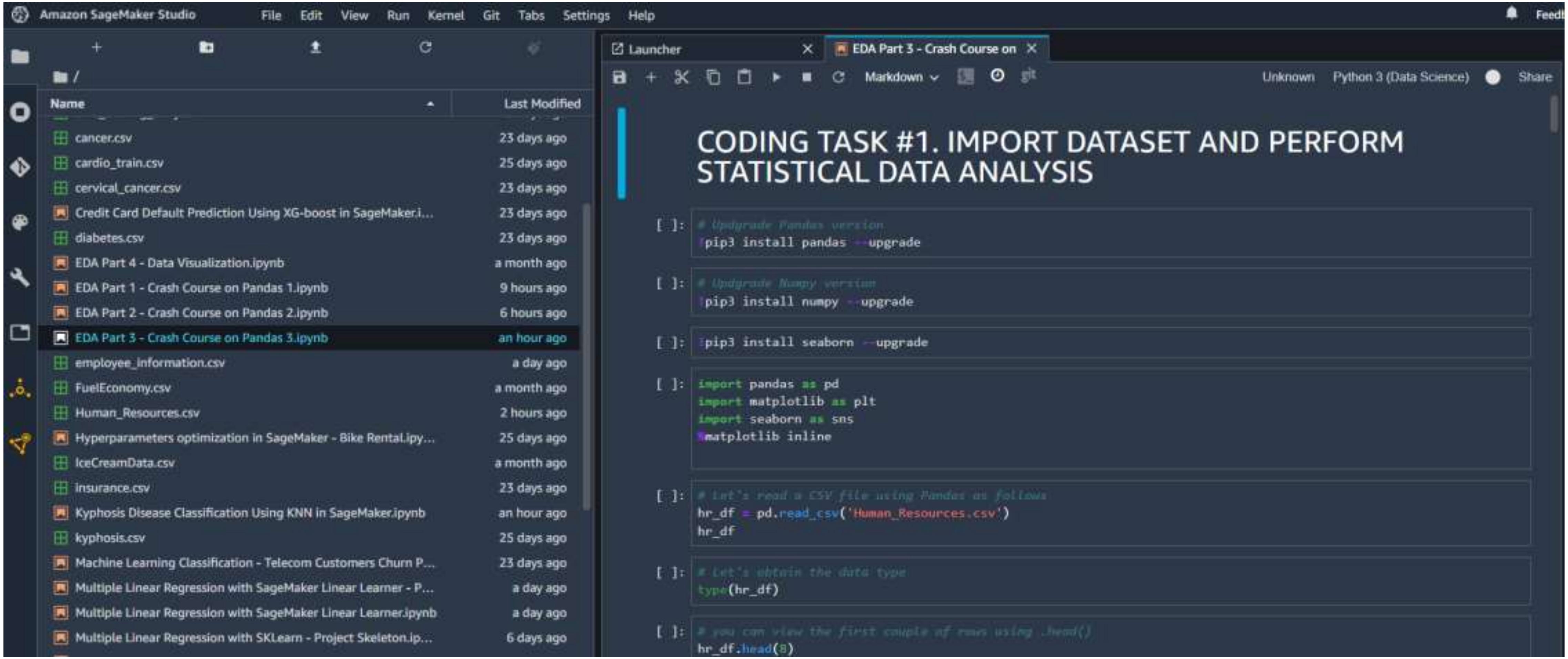


EASY

ADVANCED



PROJECT DEMO

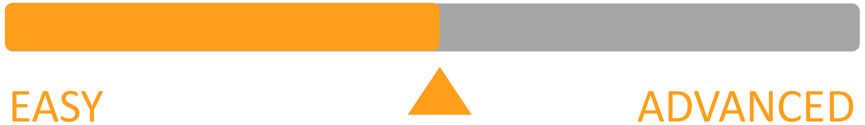


Name	Last Modified
cancer.csv	23 days ago
cardio_train.csv	25 days ago
cervical_cancer.csv	23 days ago
Credit Card Default Prediction Using XG-boost in SageMaker.i...	23 days ago
diabetes.csv	23 days ago
EDA Part 4 - Data Visualization.ipynb	a month ago
EDA Part 1 - Crash Course on Pandas 1.ipynb	9 hours ago
EDA Part 2 - Crash Course on Pandas 2.ipynb	6 hours ago
EDA Part 3 - Crash Course on Pandas 3.ipynb	an hour ago
employee_information.csv	a day ago
FuelEconomy.csv	a month ago
Human_Resources.csv	2 hours ago
Hyperparameters optimization in SageMaker - Bike Rental.ipy...	25 days ago
IceCreamData.csv	a month ago
insurance.csv	23 days ago
Kyphosis Disease Classification Using KNN in SageMaker.ipynb	an hour ago
kyphosis.csv	25 days ago
Machine Learning Classification - Telecom Customers Churn P...	23 days ago
Multiple Linear Regression with SageMaker Linear Learner - P...	a day ago
Multiple Linear Regression with SageMaker Linear Learner.ipynb	a day ago
Multiple Linear Regression with SKLearn - Project Skeleton.ip...	6 days ago

CODING TASK #1. IMPORT DATASET AND PERFORM STATISTICAL DATA ANALYSIS

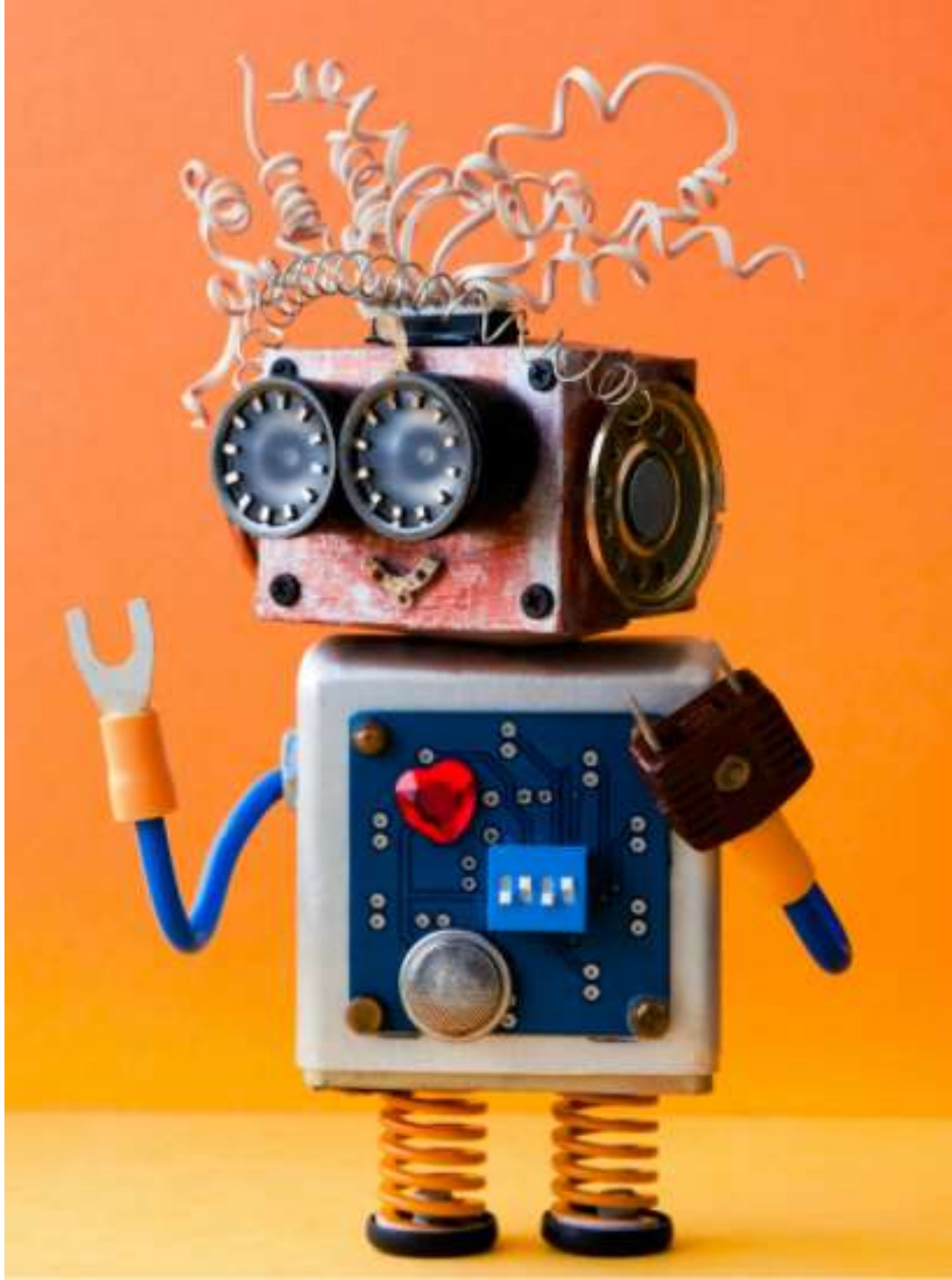
```
[ ]: # Updgrade Pandas version
     !pip3 install pandas --upgrade

[ ]: # Updgrade Numpy version
     !pip3 install numpy --upgrade

[ ]: !pip3 install seaborn --upgrade

[ ]: import pandas as pd
     import matplotlib as plt
     import seaborn as sns
     %matplotlib inline

[ ]: # Let's read a CSV file using Pandas as follows
     hr_df = pd.read_csv('Human_Resources.csv')
     hr_df

[ ]: # Let's obtain the data type
     type(hr_df)

[ ]: # you can view the first couple of rows using .head()
     hr_df.head(8)
```

FINAL CAPSTONE END-OF-DAY PROJECT



EASY

ADVANCED



FINAL PROJECT

- In this project, we will perform basic Exploratory Data Analysis (EDA) on the Kyphosis disease Dataset.
- Kyphosis is an abnormally excessive convex curvature of the spine.
- Dataset contains 81 rows and 4 columns representing data on children who have had corrective spinal surgery.
- **INPUTS:** 1. Age: in months, 2. Number: the number of vertebrae involved, 3. Start: the number of the first (topmost) vertebra operated on.
- **OUTPUTS:** Kyphosis which represents a factor with levels absent present indicating if a kyphosis (a type of deformation) was present after the operation.
- Using the “kyphosis.csv" included in the course package, write a python script to perform the following tasks:
  1. Import the “kyphosis.csv" file using Pandas
  2. Perform basic Exploratory Data Analysis (EDA) on the data
  3. List the average, minimum and maximum age (in years) considered in this study using 2 methods
  4. Plot the correlation matrix
  5. Convert the age column datatype from int64 to float64
  6. Define a function that converts age from months to years
  7. Apply the function to the “Age” column and add the results into a new column entitled “Age in Years”
  8. What are the features of the oldest and youngest child in this study?