Pandas Foundation

Online Course by DataRockie

Pandas Foundation

- Python Language
- 30+ Video Full HD
- Beginner Friendly
- Just 2 Hours Quick and easy to learn!

Course Outline

- Introduction
- Python Review `variable` `types` `list` `dictionary` `function`
- Basic Pandas `read_csv()` `add/ drop columns` `inspect dataframe` `summarise dataframe`
- Intermediate Pandas `select columns` `filter rows` `merge dataframe` `transformation`

Your Friend (Instructor)

{

“name”: “Toy”,

“page”: “DataRockie”,

“skills”: [“Economics”, “Marketing”, Data Science”],

“hobbies”: [“Reading”, “Learning”, “Sharing”],

“Tools”: [“Google Sheets”, “SQL”, “R”,
“Python”, “Data Studio”]

}

What exactly is DataFrame?

	id	name	gpa	C	city
0	1	Toy	3.42	good	Tokyo
1	2	Joe	3.50	good	London
2	3	Ann	4.00	good	Bangkok
3	4	Mary	2.58	ok	Bangkok
4	5	David	2.98	ok	Seoul

Python library designed to work with dataframe

Like Excel Table

Columns

	id	name	gpa	C	city
0	1	Toy	3.42	good	Tokyo
1	2	Joe	3.50	good	London
2	3	Ann	4.00	good	Bangkok
3	4	Mary	2.58	ok	Bangkok
4	5	David	2.98	ok	Seoul

Rows

Value

Easy to Learn

```
import pandas as pd
import numpy as np
```

import pandas

```
0.3s
raw_data = {
    "id": [1,2,3,4,5],
    "name": ["Toy", "Joe", "Ann", "Mary", "David"]
}

df = pd.DataFrame(
    raw_data
)

df
```

create dataframe

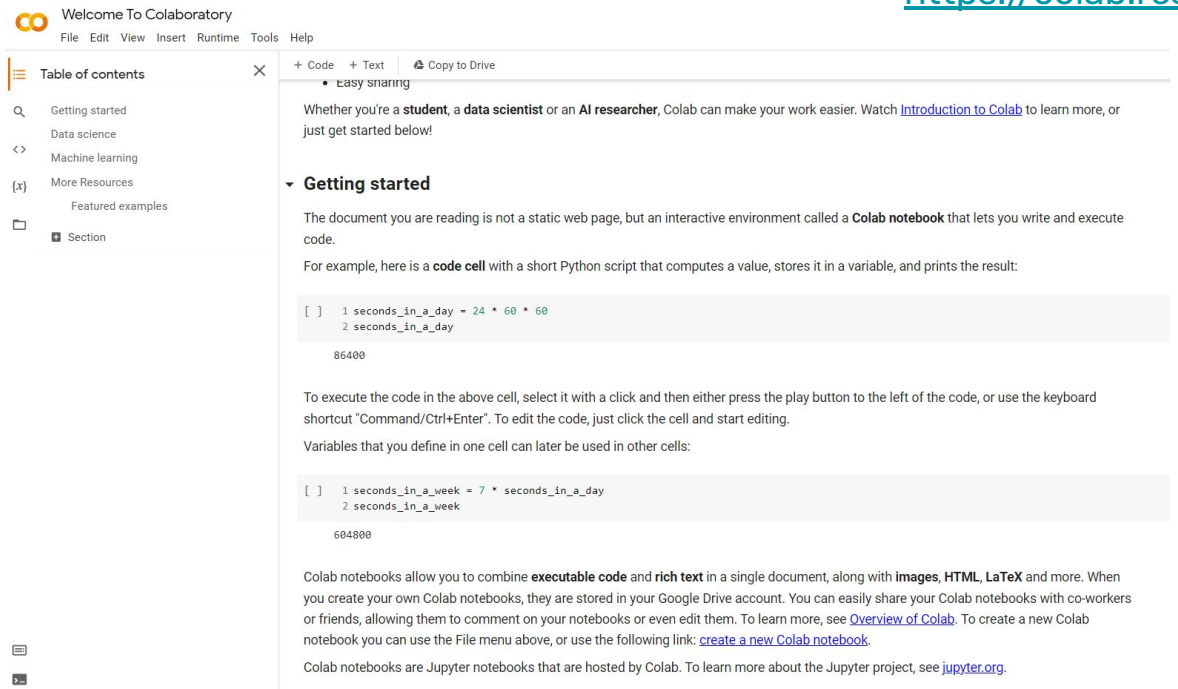
	id	name
0	1	Toy
1	2	Joe
2	3	Ann
3	4	Mary
4	5	David

Basic Python (easy!)

- Variable
- Type
- List/ Dictionary
- Function/ Method
- OOP (not required, but good to know)

Google Colab

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



DataLore Notebook

[28]
```
import pandas as pd
import numpy as np
```

[31]
```
raw_data = {
    "id": [1,2,3,4,5],
    "name": ["Toy", "Joe", "Ann", "Mary", "David"]
}

df = pd.DataFrame(
    raw_data
)

df
```

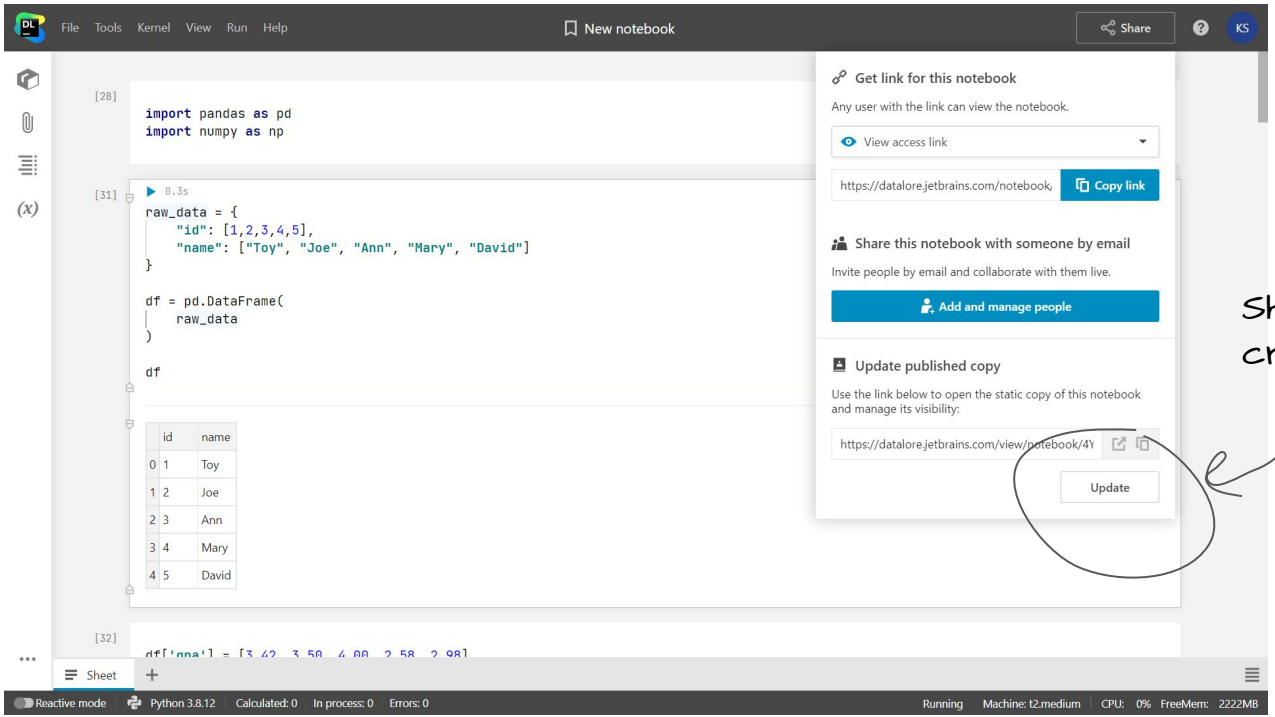
	id	name
0	1	Toy
1	2	Joe
2	3	Ann
3	4	Mary
4	5	David

[32]
```
df['gpa'] = [3.42, 3.50, 4.00, 2.58, 2.98]
```

https://datalore.jetbrains.com/

Free, easy to use,
lots of functions

Share Your Work



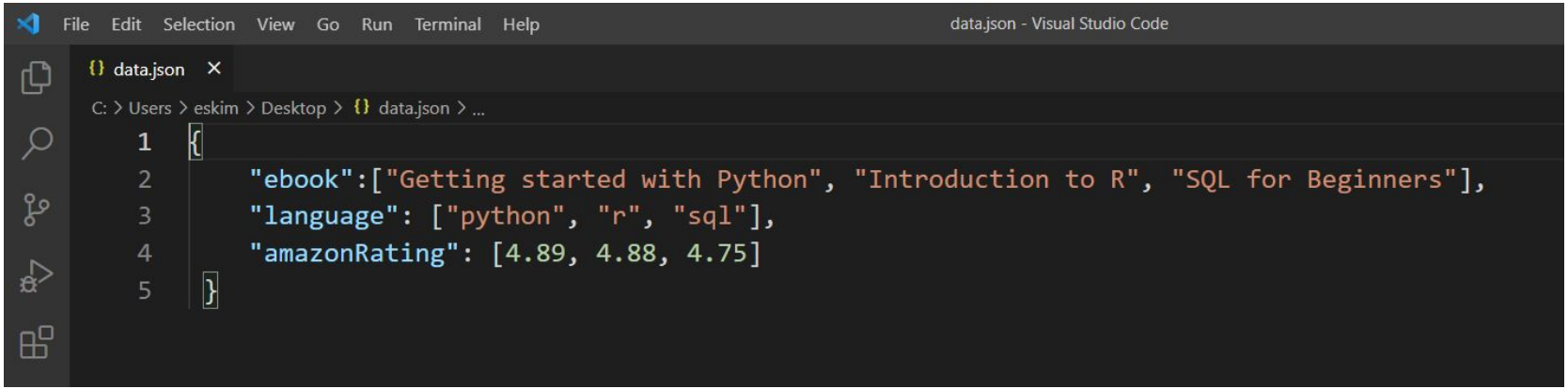
Share your work,
create portfolio

OOP

sitting(), barking(), running()

JSON

JavaScript Object Notation



Mean Imputation

	species	island	bill_length_mm	bill_depth_mm	flipper_length_...	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	nan	nan	nan	nan	nan
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

5 rows × 7 columns

Replace nan with column mean

numpy

Numerical Python https://numpy.org/

Quantum Computing	Statistical Computing	Signal Processing	Image Processing	Graphs and Networks	Astronomy Processes	Cognitive Psychology
QuTiP	Pandas	SciPy	Scikit-image	NetworkX	AstroPy	PsychoPy
PyQuil	statsmodels	PyWavelets	OpenCV	graph-tool	SunPy	
Qiskit	Xarray	python-control	Mahotas	igraph	SpacePy	
	Seaborn			PyGSP		

Bioinformatics	Bayesian Inference	Mathematical Analysis	Chemistry	Geoscience	Geographic Processing	Architecture & Engineering
BioPython	PyStan	SciPy	Cantera	Pangeo	Shapely	COMPAS
Scikit-Bio	PyMC3	SymPy	MDAnalysis	Simpeg	GeoPandas	City Energy Analyst
PyEnsembl	ArviZ	cvxpy	RDKit	ObsPy	Folium	Sverchok
ETE	emcee	FEniCS		Fatiando a Terra		

IF else using numpy

```
import numpy as np

np.where( score > 80, True, False )
```

Merge DataFrames

left

key	name	age
1	toy	25
2	joe	28
3	jane	30
4	anna	22

right

key	city	zip
1	Bangkok	1001
2	London	2504
3	Seoul	2094
4	Tokyo	9802

result

key	name	age	city	zip
1	toy	25	Bangkok	1001
2	joe	28	London	2504
3	jane	30	Seoul	2094
4	anna	22	Tokyo	9802

```
result = pd.merge(left, right, on="key")
```

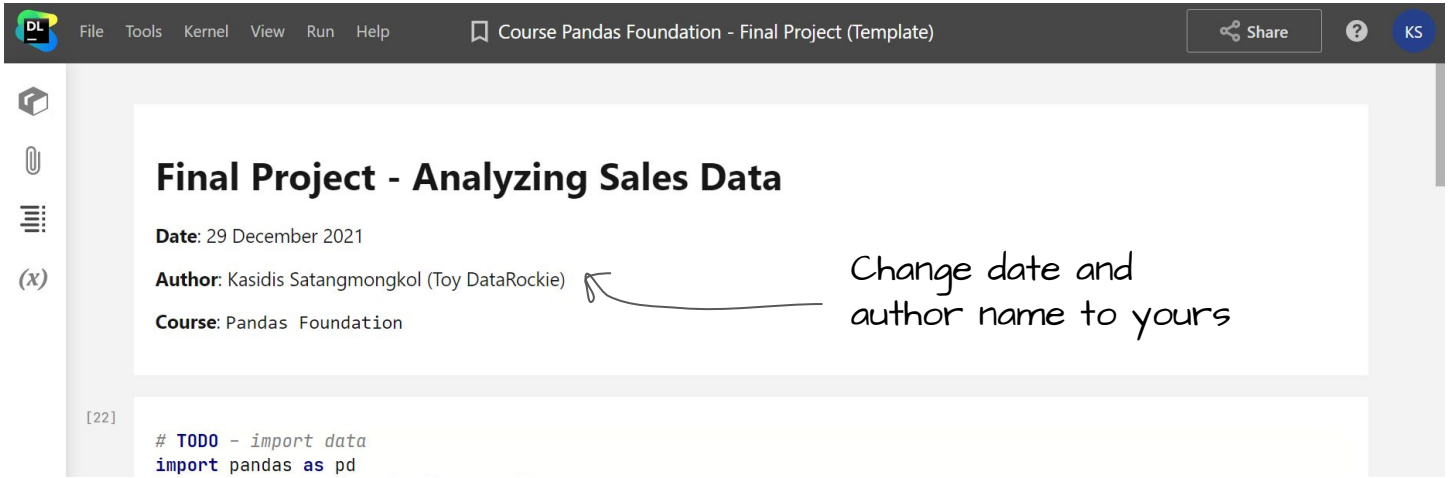
https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html

Course Review

- Python Language
- Pandas Basics
- Pandas Intermediate (Data Analyst)
- Next - Final Project

Final Project

1. Go to my datalore template
   https://datalore.jetbrains.com/view/notebook/3h5IZr0l3bZAKLkLB3omTf
2. Upload csv file [sample-store.csv] CSV
3. Write pandas to answer all questions in this notebook
4. Share your work online



Final Project

```
# TODO 01 - how many columns, rows in this dataset
```

```
# TODO 02 - is there any missing values?, if there is, which colunm? how many nan values?
```

```
# TODO 03 - your friend ask for `California` data, filter it and export csv for him
```

```
# TODO 04 - your friend ask for all order data in `California` and `Texas` in 2017 (look at Order Date), send him csv file
```

```
# TODO 05 - how much total sales, average sales, and standard deviation of sales your company make in 2017
```

```
# TODO 06 - which Segment has the highest profit in 2018
```

```
# TODO 07 - which top 5 States have the least total sales between 15 April 2019 - 31 December 2019
```

```
# TODO 08 - what is the proportion of total sales (%) in West + Central in 2019 e.g. 25%
```

```
# TODO 09 - find top 10 popular products in terms of number of orders vs. total sales during 2019-2020
```

```
# TODO 10 - plot at least 2 plots, any plot you think interesting :)
```

Pandas Foundation

Thank you and happy learning!