

Intermediate Python & a Glimpse into AI applications

Class 6

Pepe Bonet Giner

19th January 2024

Index Class 6

Topic 1: Short recap

Topic 2: Final Project

Topic 3: What do I want to see in the final project?

Topic 4: Advanced Python

Topic 5: Learning Python & Documentation

Topic 6: Final Words

Topic 1: Recap

Train & Test sets. Why are they necessary?

Training Set

Gender	Parent Education	Age	Grade
Male	College	21	8
Female	High School	22	9

Test Set

Gender	Parent Education	Age	Grade
Male	None	20	6
Female	College	21	5

To train the model. Can represent up to 90 % of the data

To test how the model generalizes to never seen data

Overfitting: The model is not good on never seen data

Underfitting: The model is bad everywhere

The reason to have a separate train and test set is to avoid overfitting of the model to the data

Objective: Find the best fit of a line to the data

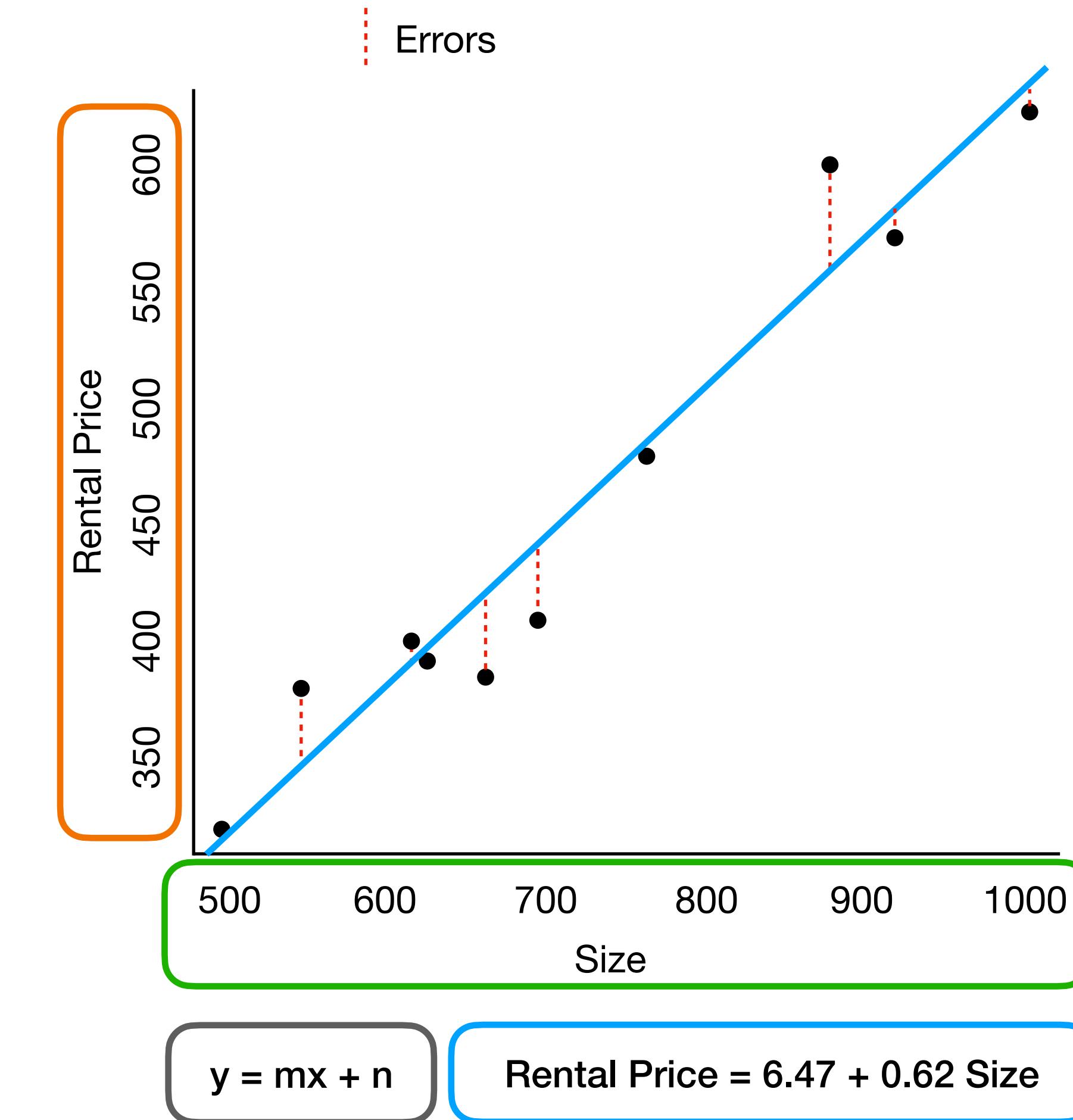
We want to obtain the line that minimizes the sum of squared errors.

$$\min(L_2)$$

To find the best parameters for the line that we are searching for (m and n in this case)

This search can be done through something called **gradient descent**

This method is at the core of many machine learning models



Linear regression model in Python to predict math scores

The technical aspects explained and many more are already handled by sci-kit learn when building any model



Therefore, our work consists of having the data ready and selecting the model. After that we are done in some lines of code

Load packages

Start linear model

Fit Model to data

Test your model

Performance measurement

```
from sklearn import datasets, linear_model  
from sklearn.metrics import mean_absolute_error
```

```
# Create linear regression object  
regr = linear_model.LinearRegression()  
  
# Train the model using the training sets  
regr.fit(trainX, trainY)
```

```
# Make predictions using the testing set  
y_pred = regr.predict(testX)  
  
# The mean absolute error  
print("MAE: %.2f" % mean_absolute_error(testY, y_pred))  
MAE: 4.41
```

Exercise

- Does anybody see a problem with the time at which we did the normalization step?
- The train test split should come first and then normalize each group separately. Example of data leakage. Try fixing this in your code.

Not an Exercise

- Where else could we have a problem with the time at which we did one of our steps?

Not an Exercise

- Where else could we have a problem with the time at which we did the normalization step?
- Filling missing values with the mean for instance. Be careful with Data Leakage

Topic 2: Final Project Explanation

How is it going to be graded?

- **60% Active participation and class attendance:**
 - 15 % Attendance
 - 45 % coding exercises
 - Exercises will be uploaded to the Moodle after every session.
- **40% Final project:**
 - Python code on Jupyter Notebook, working and applying the concepts learned in class on a dataset, explaining the steps taken.
 - A 2-page report summarizing the results, the value extracted from data, and possible applications of AI.
 - Upload both to the Moodle max two weeks after the last lecture.



Final work

Code in Jupyter Notebooks
and Html

Class 1. Intermediate Python & AI

Table of contents

- Variables
- Lists
- If-else
- For loops
- While Loops
- Functions

Variables

```
In [5]: a = 10
b = 5
print(a, b)
```

10 5

```
In [6]: a + b
```

Out[6]: 15



Report Explaining Results
Report to your boss / customer

Study of Police deaths in the USA from 1791 to 2022

Pepe Bonet Giner
9th January 2023

- 1.- Introduction to the dataset
- 2.- Results obtained (in line with the code)
- 3.- Possible further AI applications

Grade Split

Code in Jupyter Notebooks
and Html

70 %

Report Explaining Results
Report to your boss / customer

30 %

Max 2 Pages

1 Paragraph intro to Dataset (20%)
1 Paragraph AI Applications (20 %)
Else: Results (No tech)
MUST INCLUDE 2 FIGURES (60%)

Grade Split

Code in Jupyter Notebooks
and Html

70 %

Report Explaining Results
Report to your boss / customer

30 %

Max 2 Pages

1 Paragraph intro to Dataset (20%)

1 Paragraph AI Applications (20 %)

Else: Results (No tech)

MUST INCLUDE 2 FIGURES (60%)

FOLLOW GUIDELINES!

Final work

There is a folder in the Moodle with several datasets extracted from Kaggle. You also have a file with the links to the Kaggle website that could give you some ideas

The Kaggle logo is displayed in a large, bold, blue sans-serif font. The word "kaggle" is written in lowercase, with each letter having a distinct vertical stroke. The letters are slightly rounded at the top and bottom.

Final work

There is a folder in the Moodle with several datasets extracted from Kaggle. You also have a file with the links to the Kaggle website that could give you some ideas



You can either take one of them or search in Kaggle and use one from there. If you do so, you need to add the link to the dataset

Final work

There is a folder in the Moodle with several datasets extracted from Kaggle. You also have a file with the links to the Kaggle website that could give you some ideas



You can either take one of them or search in Kaggle and use one from there. If you do so, you need to add the link to the dataset

I also uploaded my code for the 5 previous lectures

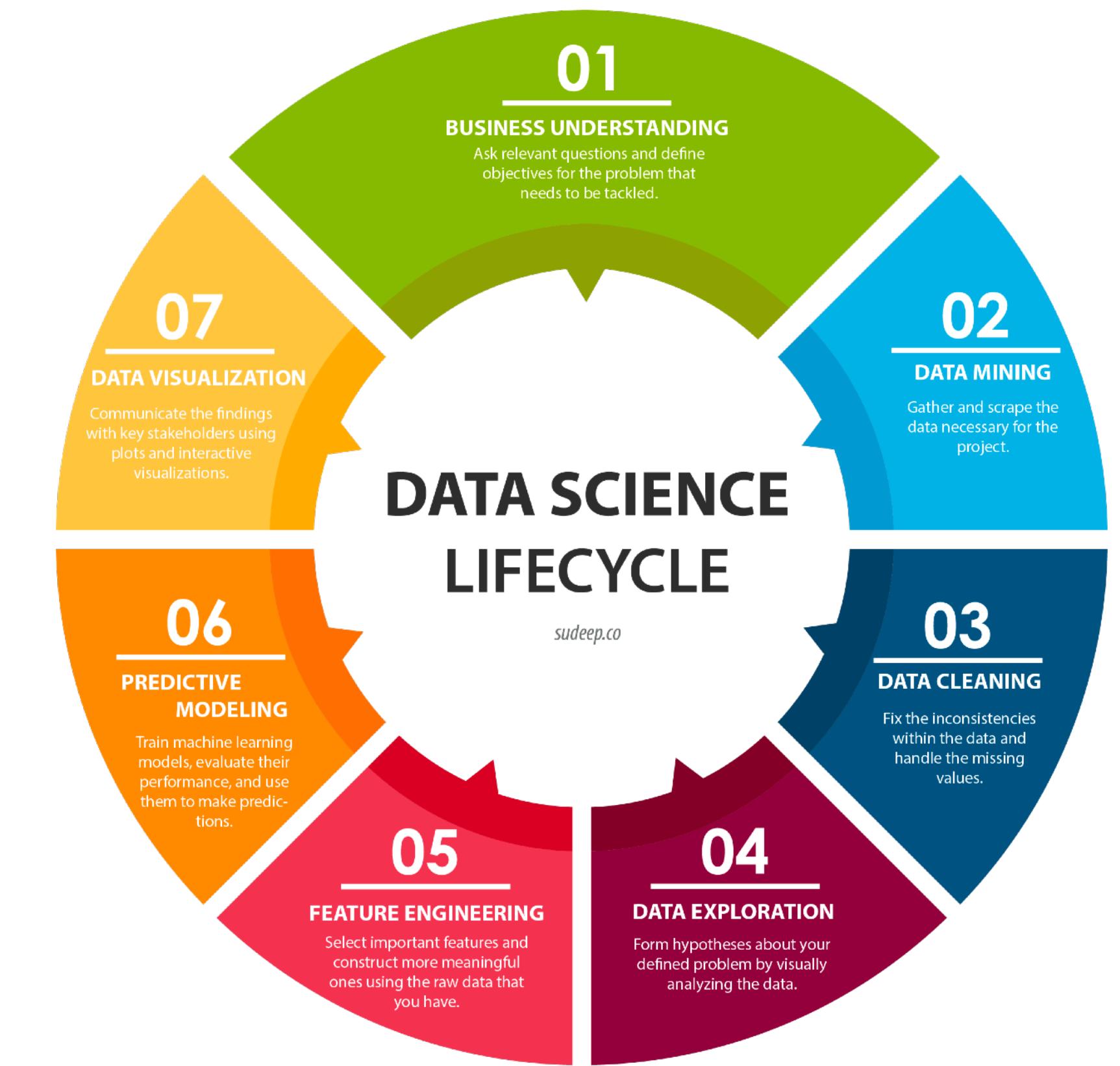
Topic 3: What do I want to see in the final project?

Review of important topics

- Lists, dictionaries, loops, if-elif-else, functions
- Pandas
 - Operations with rows and Columns
 - Groupby + apply + lambda

Review of important topics

- Lists, dictionaries, loops, if-elif-else, functions
- Pandas
 - Operations with rows and Columns
 - Groupby + apply + lambda



Review of important topics - Data Cleaning



- Step 1: Remove irrelevant data
- Step 2: Deduplicate your data
- Step 3: Fix structural errors
- Step 4: Deal with missing data
- Step 5: Deal with or Filter out data outliers
- Step 6: Validate your data

Review of important topics - EDA

- Categorical EDA
- Numerical EDA
 - Univariate analysis
 - Bivariate analysis
(Categorical + Numerical)
 - Multivariate analysis



Review of important topics - Matplotlib & Seaborn

Load Libraries

```
import matplotlib.pyplot as plt  
import seaborn as sns
```

Fully customizable

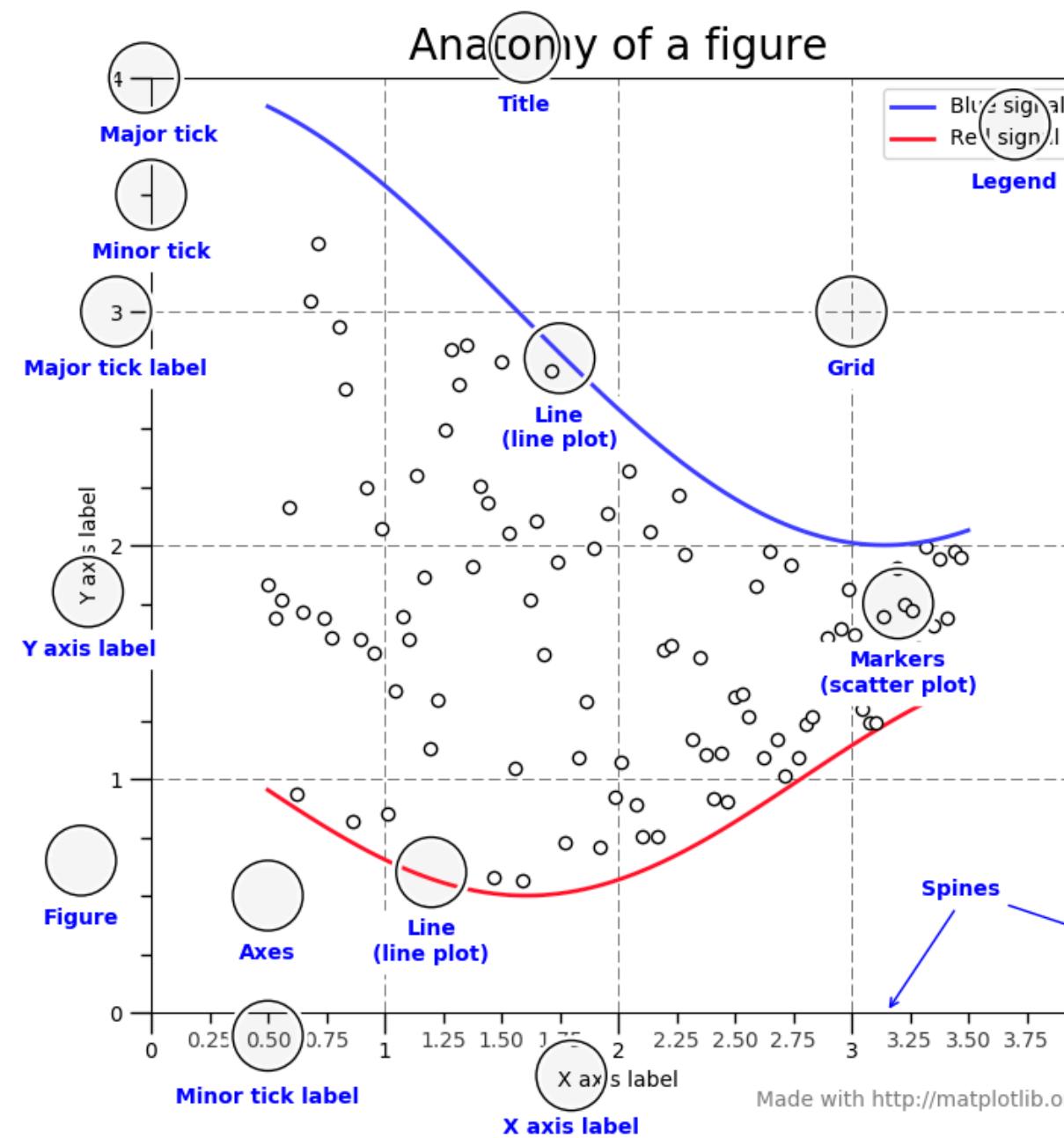
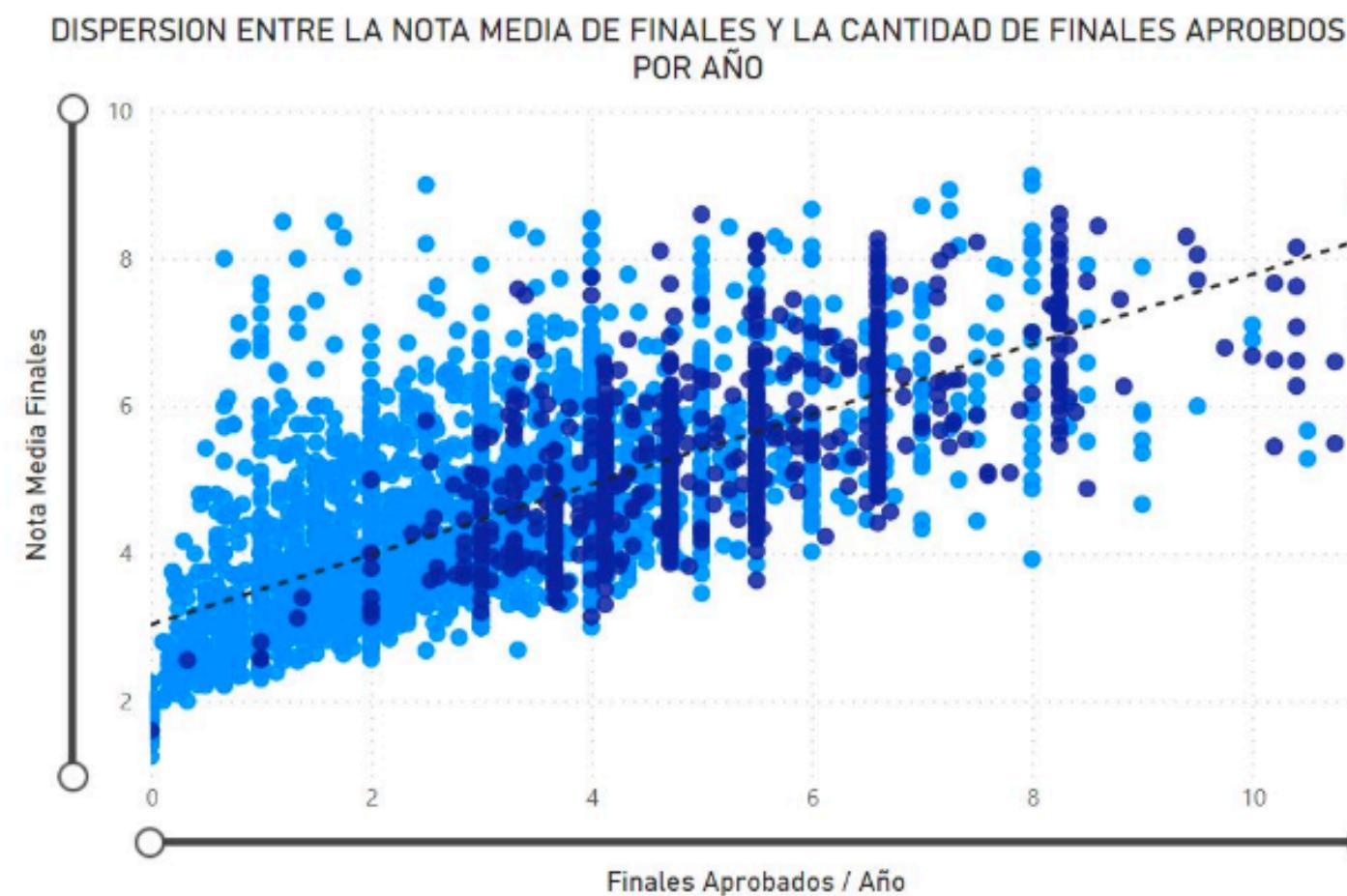
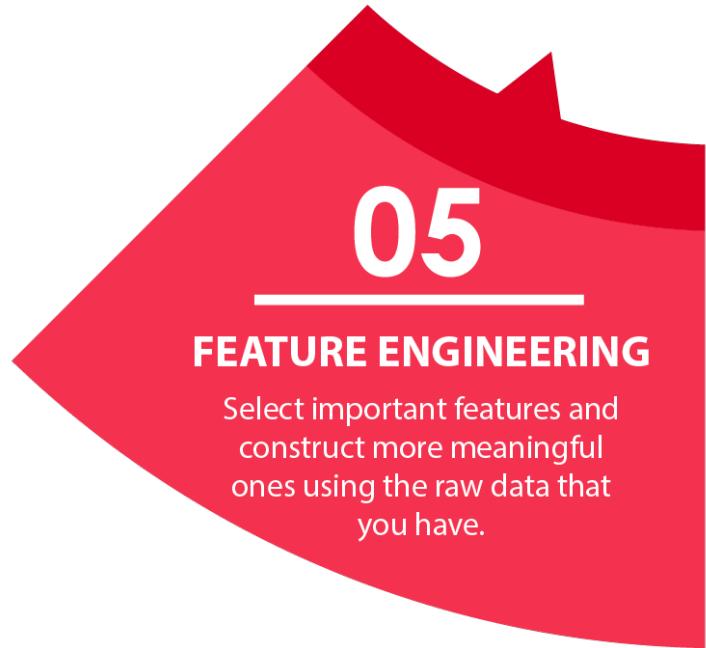


Figure structure

```
#Start Figure  
fig, ax = plt.subplots(figsize=(5, 5))  
  
#Body of the figure to build and the data to use  
sns.barplot(x=to_plot['gender'], y=to_plot['index'],  
             palette=['#08519c', '#f03b20'])  
  
#Change Axes  
ax.set_xlabel("Gender", fontsize=12)  
ax.set_ylabel("", fontsize=12)  
ax.set_yticklabels(['Male', 'Female'])  
ax.spines['top'].set_visible(False)  
ax.spines['right'].set_visible(False)  
  
# Add Numbers to plot  
for index, row in to_plot.iterrows():  
    ax.text(row.gender + 25, index, row.gender,  
            color='black', ha="center", fontsize=12)  
  
## Add legend  
custom_lines = []  
for el in [('Male', '#08519c'), ('Female', '#f03b20')]:  
    custom_lines.append(  
        plt.plot([],[], marker="o", ms=8, ls="", mec='black',  
                 mew=0, color=el[1], label=el[0])[0]  
    )  
ax.legend(  
    bbox_to_anchor=(0., 1.05, 1., .102),  
    handles=custom_lines, loc='upper center',  
    facecolor='white', ncol=1, fontsize=10, frameon=False  
)  
#Save or show  
plt.show()
```

Review of important topics - Feature engineering

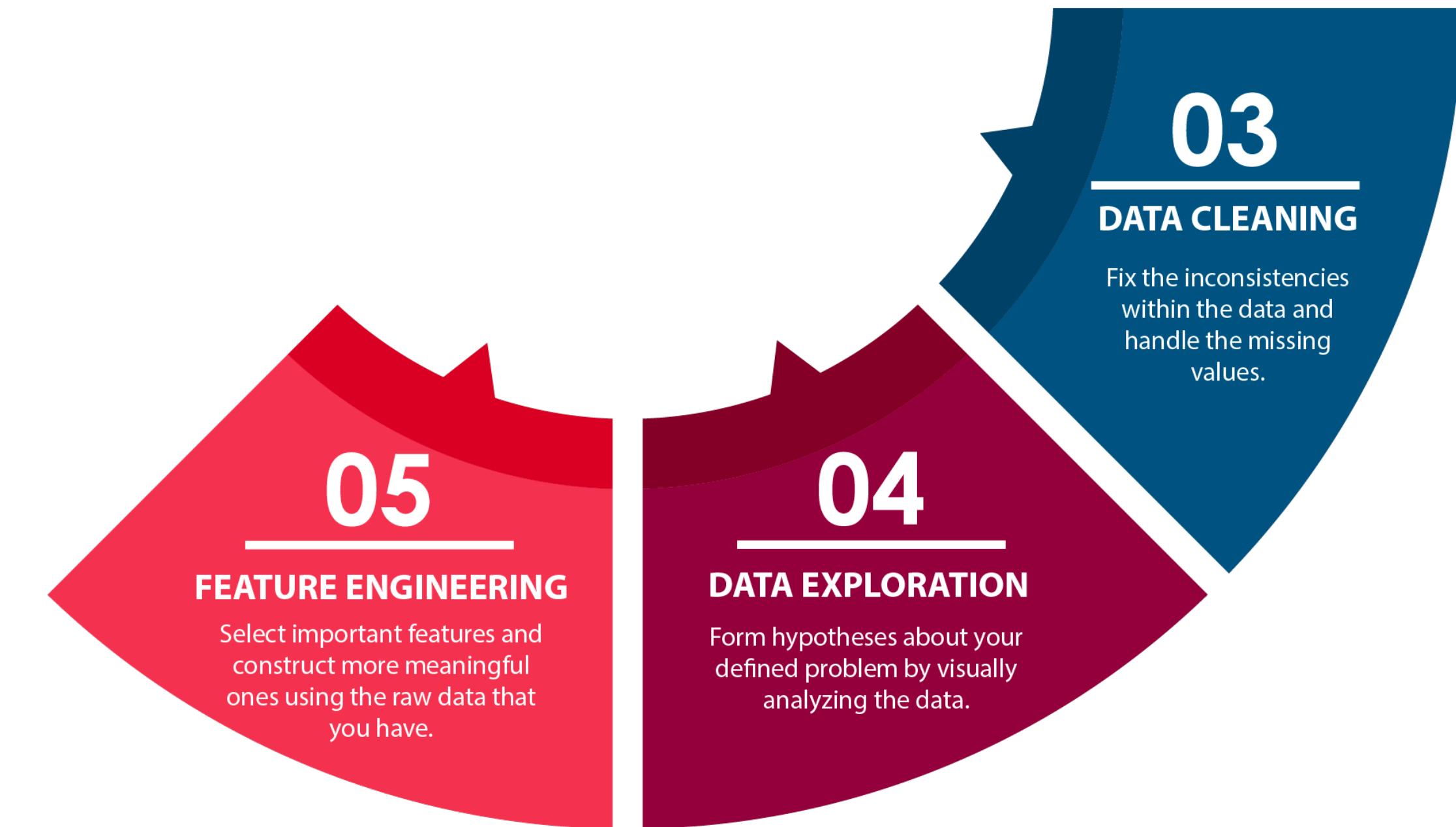


- Do you think maybe some columns could provide more information?
- Examples of current projects:
 - Chatbot: #TimesAsking4help, #Words
 - Universities: #PassedSubjects, GradeMean
- Obtain then figures with those new features

Review of important topics



- Until here you have done 90 % - 95 % of your coding work



Review of important topics - Data Preparation & Model



- Data Encoding
- Data Normalization
- Train Test Split
- Model Selection. Classification or Regression
- Code model, analyze model & get outputs

Remember

Final Project: Housing Dataset

Pepe Bonet Giner

Table of contents:

1. [Load Data](#)
2. [Initial Exploration](#)
3. [Data Cleaning](#)
 - A. Remove irrelevant data
 - B. Deduplicate your data
 - C. Fix structural errors
 - D. Deal with missing data
 - E. Deal with or Filter out data outliers
 - F. Validate your data
4. [EDA](#)
 - A. Categorical EDA
 - B. Numerical EDA
 - a. Univariate Analysis
 - b. Bivariate & Multivariate Analysis
5. [Feature Engineering](#)
6. [Data Analysis Results](#)
7. [Data Preparation](#)
8. [Predictive Model](#)

- Clean & Organized code
- With different sections & comments in your code
- Ordered workflow (Minor exploration, Cleaning, EDA, etc...)
- Show me what you have learned

```
: # Load Packages that I will use
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1. Load Data

[Go up](#)

Let's start working on it

First, are there any questions regarding
what you guys have to do?

Let's start working on it

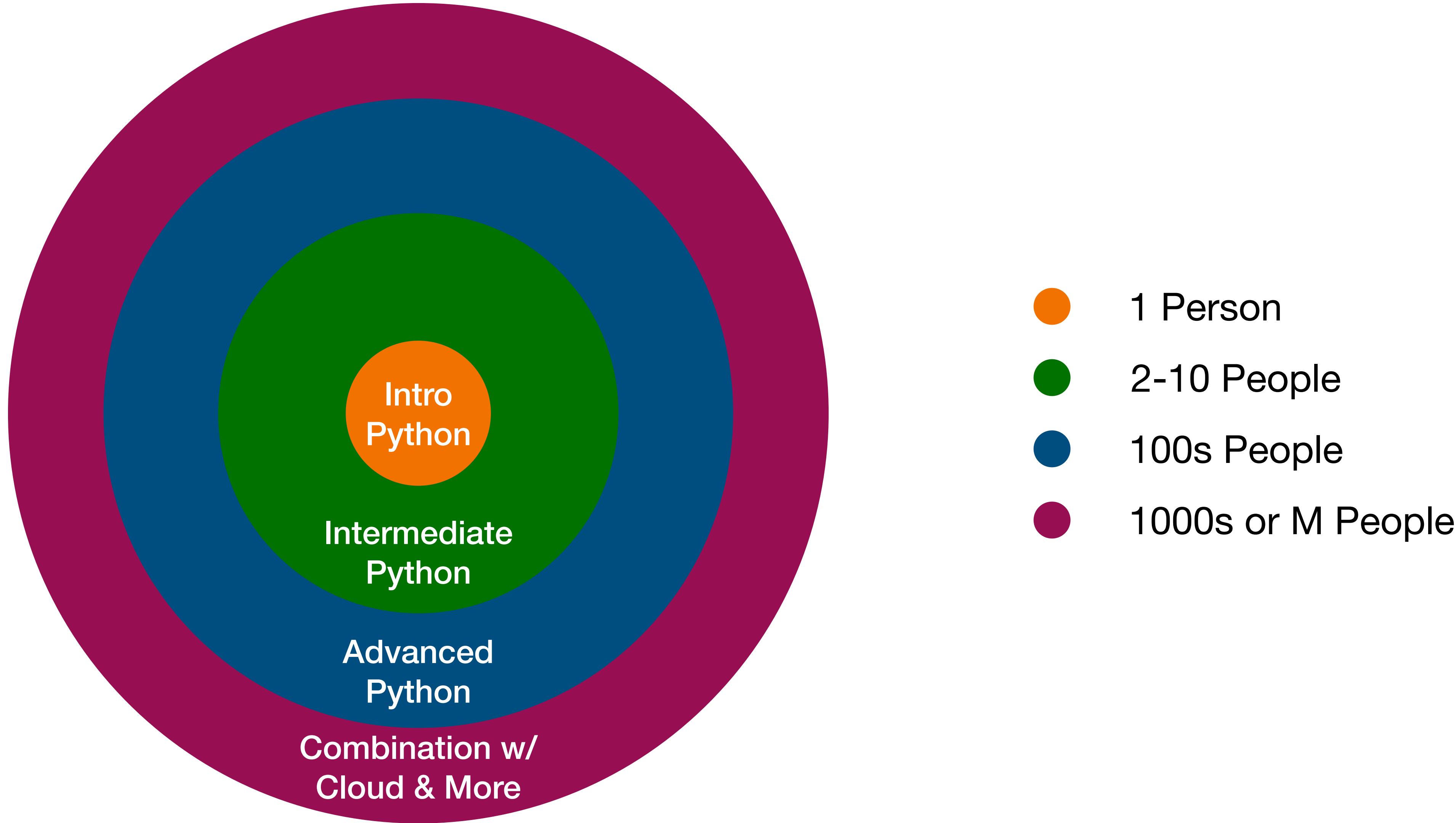
First, are there any questions regarding what you guys have to do?

Mini-warning: to use the bank marketing response dataset you load it with:

```
pd.read_csv('bank_marketing_response.tsv', sep='\t')
```

Topic 4: Advanced Python

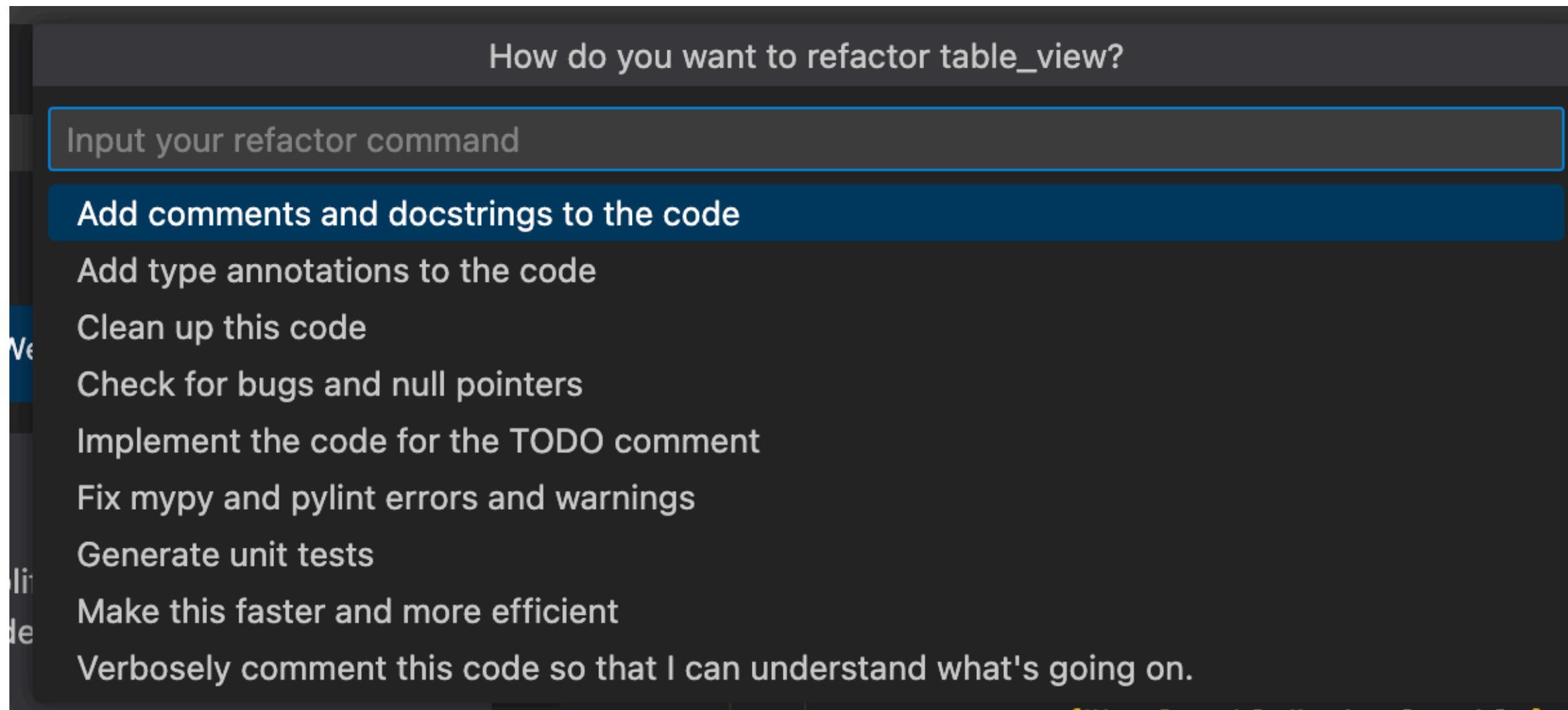
Scaling your code



Advanced Python main topics

- Scripting (VStudio)
- Classes
- Click
- Debugging
- GitHub 1 (Version Control)
- GitHub 2 (CD & CI)
- Project Structure
- Testing
- Multiprocessing
- Exceptions + assert
- Linting + Black + Cleaning Code

Example of the power of AI in VS code



```
# Write a function that computes the nth fibonacci number
Generate tests for the below function
def fib(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return fib(n-1) + fib(n-2)
```

More on this, in the advanced course if you want

Topic 5: Learning Python & Documentation

How do we learn it?

How do we learn it?

Practice

How do we learn it?

Job (Projects)

University
Courses

Videos

Own Projects

Books

Online
Courses

Practice

How do we learn it?



COURSES

ABOUT

The online coding school that invests in you

Train remotely to become a software engineer or data scientist and pay nothing upfront until you are earning \$50k or more.

START APPLICATION



Datacamp Courses

Data Scientist with Python

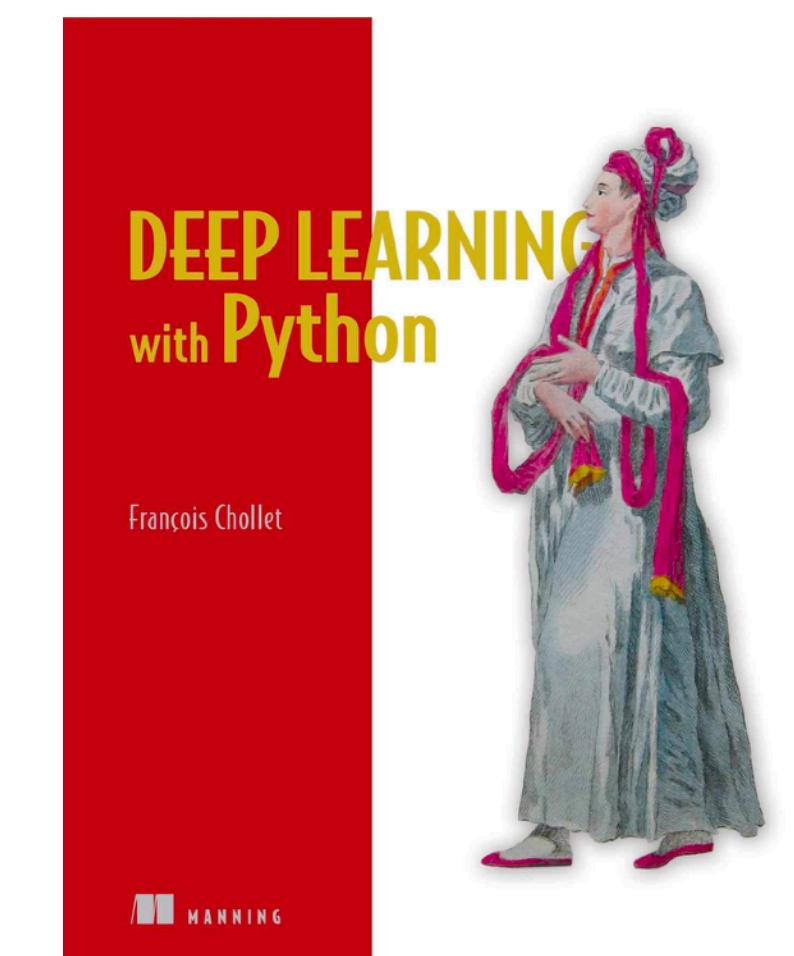
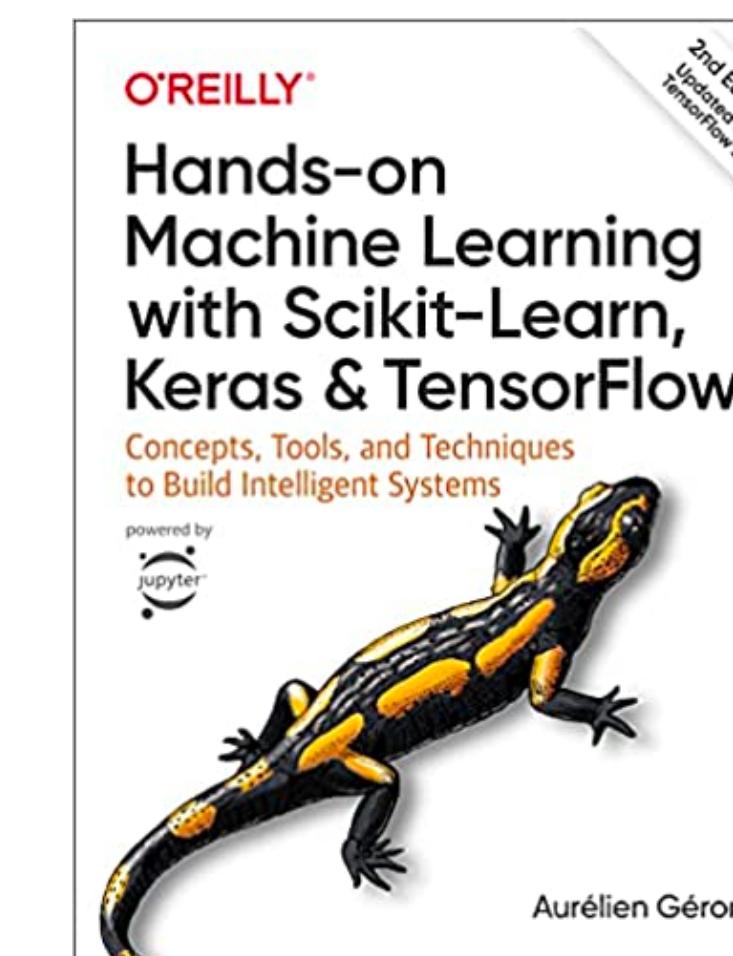
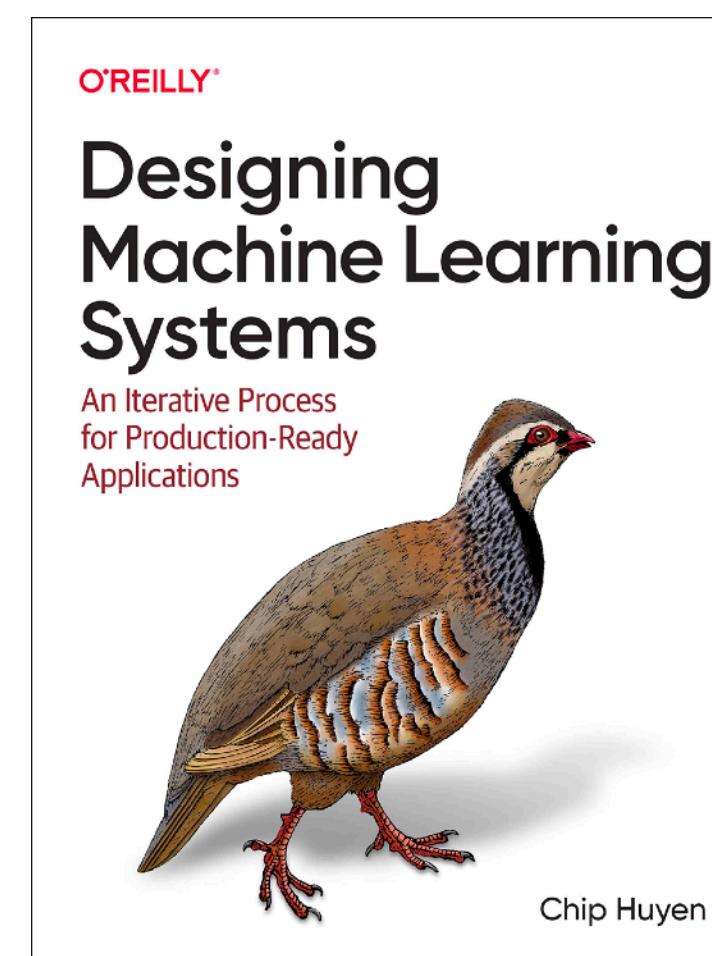
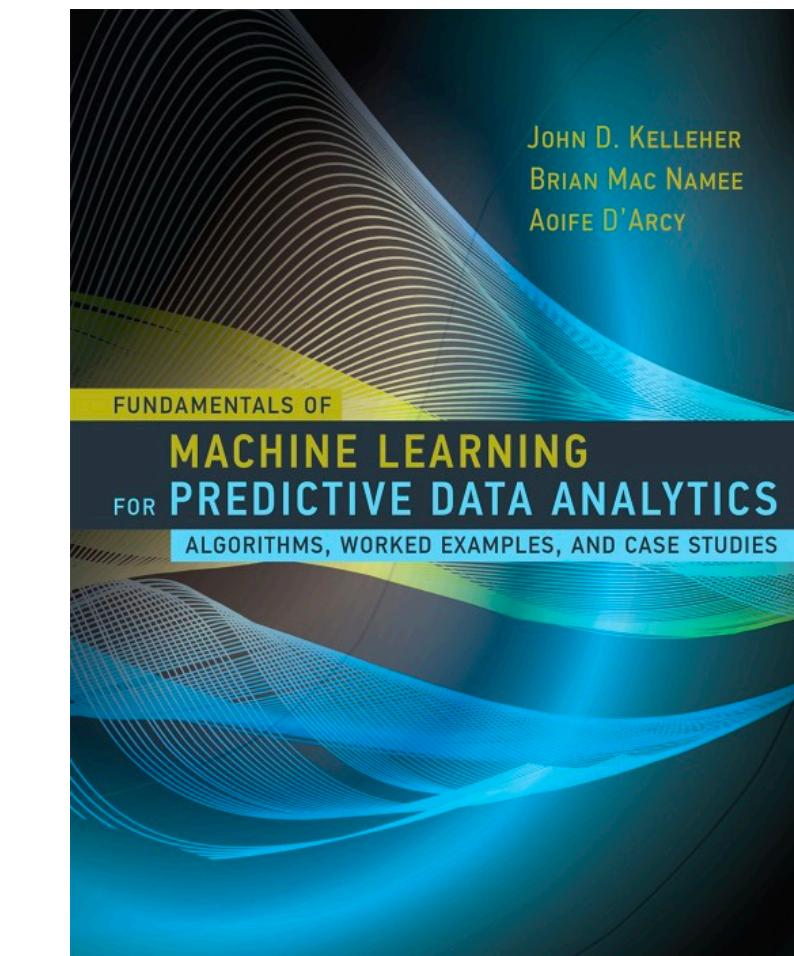
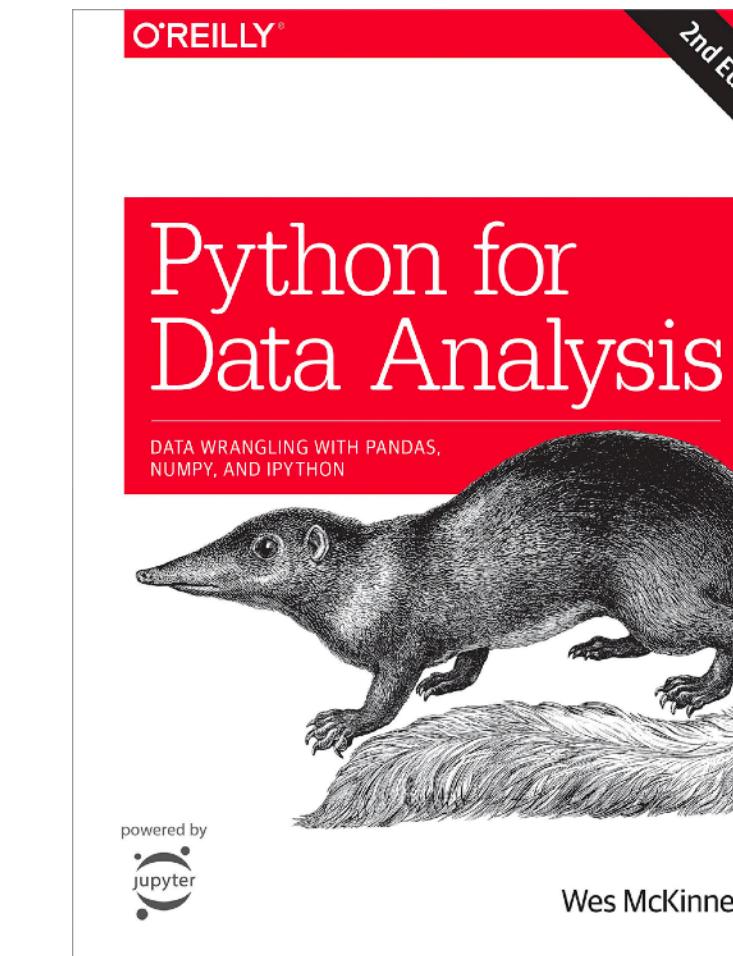
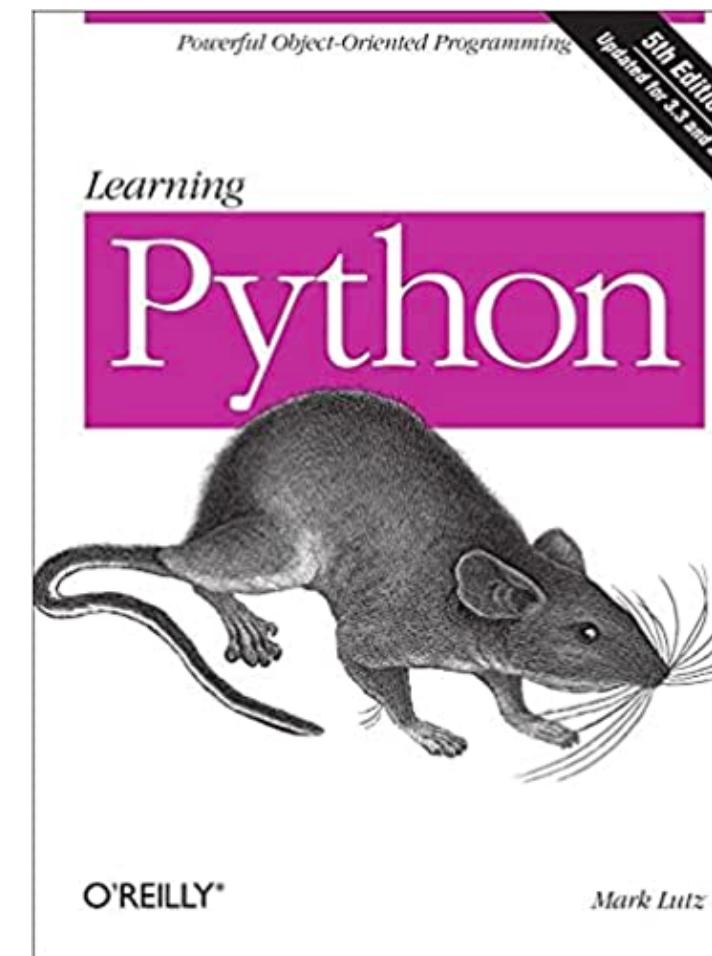
Gain the career-building Python skills you need to succeed as a data scientist. No prior coding experience required.

In this track, you'll learn how this versatile language allows you to import, clean, manipulate, and visualize data—all integral skills for any aspiring data professional or researcher. Through interactive exercises, you'll get hands-on with some of the most popular Python libraries, including pandas, NumPy, Matplotlib, and many more. You'll then work with real-world datasets to learn the statistical and machine learning techniques you need to train decision trees and use natural language processing (NLP). Start this track, grow your Python skills, and begin your journey to becoming a confident data scientist.

[Resume Track](#)

⌚ Python ⏸ 88 hours 📚 23 Courses 💾 6 Projects 🗂️ 3 Skill Assessments

Free Books to learn Python & data analytics



Topic 6: Final Words

You are already ahead

- I know writing code is frustrating
- Having somebody telling you how bad you are doing it every minute is not nice
- Especially if you spend hours stuck in a problem
- But you know a lot of Python already

You are already ahead

- I know writing code is frustrating
- Having somebody telling you how bad you are doing it every minute is not nice
- Especially if you spend hours stuck in a problem
- But you know a lot of Python already

Me

20-21

Coding?
What is that?

You are already ahead

- I know writing code is frustrating
- Having somebody telling you how bad you are doing it every minute is not nice
- Especially if you spend hours stuck in a problem
- But you know a lot of Python already

Me

20-21

Coding?
What is that?

Python: 22
Pandas & JN: 24
AI & ML: 24-25
DL: 25-26
Teaching & Workshops:
27-28

You are already ahead

- I know writing code is frustrating
- Having somebody telling you how bad you are doing it every minute is not nice
- Especially if you spend hours stuck in a problem
- But you know a lot of Python already

Me

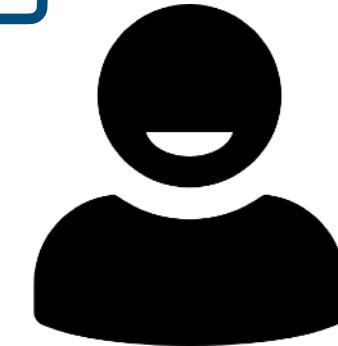
20-21

Coding?
What is that?

Python: 22
Pandas & JN: 24
AI & ML: 24-25
DL: 25-26
Teaching & Workshops:
27-28

You

20-21



- Business & Economy knowledge
- Intermediate Python
- Pandas & JN
- Work with Datasets
- AI Intuition

You are already ahead

- I know writing code is frustrating
- Having somebody telling you how bad you are doing it every minute is not nice
- Especially if you spend hours stuck in a problem
- But you know a lot of Python already

Me

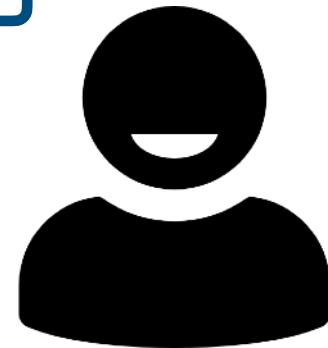
20-21

Coding?
What is that?

Python: 22
Pandas & JN: 24
AI & ML: 24-25
DL: 25-26
Teaching & Workshops:
27-28

You

20-21



- Business & Economy knowledge
- Intermediate Python
- Pandas & JN
- Work with Datasets
- AI Intuition

If you enjoy it, the only thing you need is to work on it

My Take reviewed

- I hope you are more aware after the course that AI is changing the world and you will have to live in it
- Chances are your work may be affected by it, but ideally, this course has prepared you a bit more for it. It can be programming, but it can also be possibilities to exploit AI and hire people to do it or related projects where you have to understand a bit of the technical/coding part as well
- I hope this course has taken you closer to a better understanding of python, how to use it to extract information for datasets, and the AI world and its possible applications
- Essentially, I hope you learned something and enjoyed it. It is been a pleasure!

Thank you!

Feel Free to contact me!

- Linkedin: Jose Bonet Giner
- Personal mail: pepebogi5@gmail.com
- Company website: hyntsanalytics.com
- Blog: pepesjourney.com



Thank you! / Last exercise ;)

Fill in the survey / Feedback form!