



Economia, Finanza e Analisi dei Dati

Laurea Magistrale

Data Analysis for Economics

01 - Introduction

Marco Nieddu

Fall 2025

This course: what it is

- ▶ **New** (2nd edition) course on coding (mainly Python) to collect, organize, and analyze data in economics
- ▶ **Course goal:** provide you with knowledge to
 - ▶ go from a research idea to the final dataset that allows you to test it
 - ▶ write results of your analysis in an 'economic report/paper' structure, with appropriate **charts** and tables
- ▶ **Skills you will develop:** data visualization, cleaning, wrangling, programming, etc.
 - ▶ practical (and often neglected) skills useful for dissertation and career
 - ▶ important prerequisite for 'standard' metrics/policy evaluation course in next term(s)

This course: what it is

- ▶ Besides coding, this is also a **Topics in Applied Economics** course
- ▶ We will work with real data to answer economic questions such as:
 - ▶ Gender pay gap (Labor econ)
 - ▶ Prices and taxation (Public econ / IO)
 - ▶ Religious holidays and productivity (Political econ)
- ▶ At the beginning of each topic, we devote space to economic theory and related literature (before data work)

This course: what it is *not*

- ▶ Econometrics is **not** the core of this course
 - ▶ Although we may cover some econometrics/statistics concepts and intuitions
- ▶ We fill the gaps left by traditional econometrics and methods classes
 - ▶ e.g., how to actually find datasets in the wild and clean them

Course preliminaries I: People

- ▶ Me: [Marco Nieddu](#)
 - ▶ Associate Professor in Public Economics (*Scienza delle finanze*)
 - ▶ I work on applied microeconomics (mostly education/development)
 - ▶ I'm an (applied) economist using data science tools for economics research, **not a data scientist!**
- ▶ Teaching Assistant: [Alessio Garau](#)
 - ▶ He will help with projects (more later) + give extra classes on programming basics and syntax

Course preliminaries II: Course Material

- ▶ No single textbook; we borrow material from several sources
- ▶ The syllabus combines my own material with that of other authors (**which I fully acknowledge**)
- ▶ All teaching material will be available in the course shared folder*
 - ▶ *Instructions on setup in next slides
 - ▶ New material posted online before each class

Course preliminaries II: Course Material

- ▶ We use OneDrive and/or Dropbox
- ▶ Slides and material posted before class
- ▶ Files update live during classes, so you can follow scripts if you get lost
- ▶ **Please make sure you are added to the shared folder(s), and that you can see them on your computer (synced).** In case you cannot, please send me an email **ASAP**.

Course preliminaries II: Course Material

Some notes/books I borrow from:

- ▶ Pedro H.C. Sant'Anna (Emory University), "Data Science for Economics", [link](#)
- ▶ Jared Hutchins (University of Illinois), "Data Science for Applied Economics", [link](#)
- ▶ Grant McDermott (Amazon), "Data Science for Economists", [link](#)
- ▶ Tyler Ransom (University of Oklahoma), "Data Science for Economists", [link](#)
- ▶ Florian Heiss and Daniel Brunner, "Using Python for Introductory Econometrics", [link](#)
- ▶ Gábor Békés & Gábor Kézdi, "Data Analysis for Business, Economics, and Politics". Cambridge UP, 2021

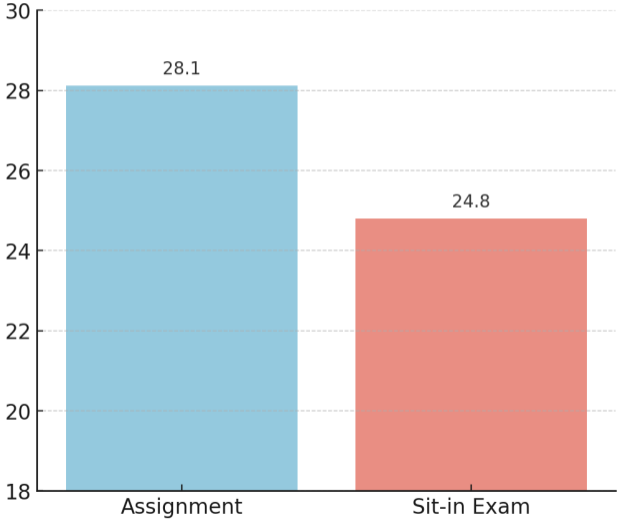
Exam/Grading

	Component	Weight	When	Rules
1	Individual project	50%	Nov. 16th	Open book (GPT ✓)
2	Sit-in exam (PC)	50%	Jan. 19th*	Closed book (GPT ✗)

OR

	Component	Weight	When	Rules
1	Sit-in exam (PC)	100%	Jan. 19th	Closed book (GPT ✗)

Which option to choose?



1. Individual project

- ▶ Output: 5-10 page essay on topic of your choice (economics-related)
- ▶ Includes finding data, collecting/cleaning, and performing brief analysis, possibly following this structure:
 1. Abstract (whole story in 150 words)
 2. Research question (why important)
 3. Brief review of literature (what we know)
 4. Data description (which data/why/how they look)
 5. Results: 3-4 objects to summarize findings (+ comments)
 6. Scripts (and source file) to **reproduce*** figures/tables/analyses

*you may be asked questions about the code (No “IDK, GPT did this”)

1. Individual project (cont'd)

- ▶ To get the grade, you must also give short (10-15 min) presentation on your project at a later (but not final) stage
 - ▶ For me, needed to ensure substantial human contribution (I don't grade GPT)
 - ▶ For you, chance to receive feedback and improve project before submission
- ▶ **Important dates:**
 - ▶ Deadline to present research idea: October 8th
 - ▶ Presentations: November 3rd and 4th
 - ▶ Deadline to submit final project: November 15th

2. Sit-in Exam

The exam is a mini-version of the research project

- ▶ You are expected to download/import some data, analyze, and write a very short comment
- ▶ Task simple enough to be completed in three hours
- ▶ **Closed-book exam** (No GPT!), but you may look at course material
- ▶ Rationale: incentivize using GPT as learning tool (a complement), not as substitute

Course roadmap

- ▶ Lectures: Mon/Tue/Wed 12-14 (last class: November 5th)
- ▶ Additional TA classes (Thu, time TBA)

Week	Day	Title	Data
W1	29sep - 01oct	Intro and Python basic syntax	
W2	06oct - 08oct	Topic 1: Gender differences in labor market	Eurostat
W3	13oct - 15oct	Topic 2: Gas prices and tax policies	MISE
W4	20oct - 23oct	Topic 2: Gas prices and tax policies	MISE
W5	27oct - 29oct	Topic 3: Religious festivals and development	Wikipedia
W6	03nov - 05nov	Presentations and Mock Exam	

Our schedule is ambitious and may be adjusted during the course!

Getting started

- ▶ Download and install Anaconda:
 - ▶ <https://www.anaconda.com/download>
 - ▶ Anaconda includes Python, Spyder, and Jupyter
- ▶ Download STATA (you have access as UNICA students)

Why are we using Python?

We are going to focus on one language: **Python**. Why?

1. General purpose
2. Open source
3. Relatively easy to use
4. Bonus: GPT can help a lot when coding in Python

But we will also see the **STATA** equivalents of scripts we write for each project/topic

1. General Purpose

- ▶ **R and Stata** aimed mainly at data analysis and processing
- ▶ **Python** not specific to data analysis. Used for ML, web scraping, even apps

Advantage: Python can do nearly all tasks

Disadvantage: Its econometrics support is weaker than R and Stata

2. Open Source

- ▶ **Stata** costs a lot, sold by private company
- ▶ **Open Source languages** (R, Python) free to use
- ▶ Maintained by a community, packages exist for almost everything

Disadvantage: Packages do not always work well together

3. Easy to use

- ▶ Compared to other languages, **Python** is intuitive and easy
- ▶ High-level, object-oriented, interpreted → easy to debug

Disadvantage: Slow, not always memory efficient

4. Increasingly popular

- ▶ <https://pypl.github.io/PYPL.html>
- ▶ <https://r4stats.com/articles/popularity/>

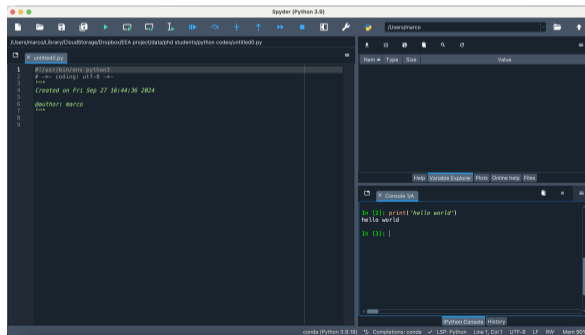
Getting started II

Once you have installed Anaconda

1. Launch Spyder

- ▶ We are going to use Spyder (Python scripts) as it makes reproducibility easier

2. Once open, in the console (bottom right), type: `print('hello world')`



```
1 #!/usr/bin/env python3
2 # -*- coding: utf-8 -*-
3 """
4 Created on Fri Sep 27 16:44:36 2024
5
6 @author: marco
7 """
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

If you see 'hello world' below the command line, you are good to go.

This is where we will start tomorrow.

Addendum: Data visualization

- ▶ When it is beautiful
 - ▶ <https://www.reddit.com/r/dataisbeautiful/>
- ▶ When it is not
 - ▶ <https://www.reddit.com/r/dataisugly/>