Session 5. A date with EDA: Bring your own data

2026-02-18

Highlights:

This is my mini-reflection. Paragraphs must be indented.
It can contain multiple paragraphs.

> "Errors using inadequate data are much less than those using no data at all."
>
> — Charles Babbage

> "Data that is loved tends to survive."
>
> — Kurt Bollacker

Session outline

- Reading external data
- Hands-on practice
- Some additional notes on working with `Quarto` / `Jupyter`

Reminder

The human brain is a highly evolved machine with a specialization in visual recognition of spatial patterns, including shapes, areas, lengths, directions, and colors. This is why well-designed statistical plots are so effective for conveying complex information.

Preliminaries

Load packages. Remember, packages are units of shareable code that augment the functionality of base `Python`. For this session, the following package/s is/are used:

```
import pandas as pd
import geopandas as gpd
import matplotlib.pyplot as plt

# Set display options to show all rows and columns
pd.set_option('display.max_rows', None)    # Show all rows
pd.set_option('display.max_columns', None) # Show all columns
pd.set_option('display.width', None)       # Auto-detect width
pd.set_option('display.max_colwidth', None) # Show full column content
```

We also will utilize some data from the `edashop` R package. To convert these R data files to `Python` files, we will use the `reticulate` package:

```r
library(edashop)  # A Package for a Workshop on Exploratory Data Analysis
library(reticulate)
```

Reading external data

In this workshop we have used data that was meticulously prepared and documented to make it analysis-ready. But often we need to read data from so-called external sources, that is, files that are not in native `R` format, including Excel files, csv files, Stata files, shape files, and so on.

The original files of the data sets provided in {edashop} are shared with the package. For example, check the following help file:

```r
?i40_index_rank
```

This is one of two data frames created using original files from Honti et al. (2020).

The path to the source CSV file can be obtained in `R` and then passed to `Python`:

```r
csv_path  <- system.file("extdata",
                         "rankings.csv",
                         package = "edashop")
```

You can check that this object is just a string with the path to the file:

```python
csv_path = r.csv_path
print(csv_path)
```

```
C:/Users/brdia/AppData/Local/R/cache/R/renv/library/edashop-3dfdfba5/windows/R-4.4/x86_64-w64-mingw32/e
```

One way to read files in CSV format is with `pandas.read_csv()`:

```python
imported_csv = pd.read_csv(csv_path)
print(imported_csv.head())
```

```
  Regio  GDPrank  PrometheeRank  RII
0  UKN0     68.0          149.0  NaN
1  UKM9    165.0          193.0  NaN
2  UKM8     26.0          116.0  NaN
3  UKM7     16.0           70.0  NaN
4  UKM6    149.0          107.5  NaN
```

Once the file is imported it can be saved as a file of type rda for ease of access in the future. For reproducibility purposes, it is highly recommended to keep the source files intact and document any data processing done in a script or Quarto document. For example, here we change the names of the columns in the table we just imported:

```python
imported_csv = imported_csv.rename(columns = {
    'Regio': 'NUTS_ID',
    'GDPrank': 'gdp_rank',
    'PrometheeRank': 'promethee_rank',
    'RII': 'regional_innovation_index'
})
print(imported_csv.head())
```

```
  NUTS_ID  gdp_rank  promethee_rank  regional_innovation_index
0    UKN0      68.0           149.0                        NaN
1    UKM9     165.0           193.0                        NaN
2    UKM8      26.0           116.0                        NaN
3    UKM7      16.0            70.0                        NaN
4    UKM6     149.0           107.5                        NaN
```

Then we save the file (optional):

```
# Save to a pickle file for faster loading later
imported_csv.to_pickle("external_csv.pkl")

# Or save as CSV
imported_csv.to_csv("external_csv.csv", index = False)
```

This makes it easy to retrieve our prepared data.
An external file in Stata's dat format is also shared:

```
dta_path <- system.file("extdata", "phd_italy.dta", package = "edashop")
```

As before `dta_path` is a string with the path to the file:

```
dta_path = r.dta_path
print(dta_path)
```

```
C:/Users/brdia/AppData/Local/R/cache/R/renv/library/edashop-3dfdfba5/windows/R-4.4/x86_64-w64-mingw32/e
```

pandas can read Stata files directly with read_stata():

```
imported_dta = pd.read_stata(dta_path)
print(imported_dta.head())
```

```
           id  q01_annodinascita  q02_sesso  q03_cittadinanzaest q04_provvive  \
0  QSBBFT60BN               1985        0.0                  0.0           RM
1  QS1T4P4F5E               1982        0.0                  0.0           CA
2  QSDPZCAFOM               1960        0.0                  0.0           FI
3  QSO7842WKF               1981        1.0                  0.0       ESTERO
4  QS6DGKNLT2               1984        1.0                  0.0           RO

   q05_genitorelaureato  q06_genitoreaccad  q07_genitoreimpr  concludott_cert  \
0                   1.0                0.0               0.0              1.0
1                   1.0                0.0               0.0              1.0
2                   0.0                0.0               0.0              0.0
3                   1.0                0.0               0.0              0.0
4                   1.0                1.0               0.0              1.0

   q08_phd_clean  q08a_adj1  q08a_adj2  q08a_adj3  q08a_adj4  q08a_adj5  \
0            1.0        0.0        0.0        0.0        0.0        0.0
1            1.0        0.0        0.0        0.0        0.0        0.0
2            0.0        1.0        0.0        0.0        0.0        0.0
3            0.0        1.0        0.0        0.0        0.0        0.0
4            1.0        0.0        0.0        0.0        0.0        0.0

   q08a_adj6  q08a_adj7  q08a_adj8               struttura code_un  \
```

```
0        1.0        0.0        0.0  Univ. ROMA LA SAPIENZA      26
1        0.0        0.0        1.0          Univ. CAGLIARI      04
2        0.0        0.0        0.0           Univ. FIRENZE      10
3        0.0        0.0        0.0           Univ. PALERMO      20
4        0.0        1.0        0.0  Univ. ROMA TOR VERGATA      27

  nome_provincia nome_regione  q13a_phdgiudizio  q13b_phdgiudizio  \
0           Roma        Lazio               5.0               4.0
1       Cagliari     Sardegna               6.0               5.0
2        Firenze      Toscana               6.0               6.0
3        Palermo      Sicilia               5.0               6.0
4           Roma        Lazio               6.0               5.0

   q13c_phdgiudizio  q13d_phdgiudizio  q13e_phdgiudizio  q13f_phdgiudizio  \
0               4.0               3.0               4.0               3.0
1               3.0               4.0               4.0               6.0
2               3.0               3.0               5.0               NaN
3               4.0               5.0               6.0               5.0
4               5.0               5.0               5.0               4.0

   q13g_phdgiudizio  q13h_phdgiudizio  q13i_phdgiudizio  q14_phdfinpriv  \
0               1.0               4.0               3.0             0.0
1               5.0               5.0               3.0             0.0
2               4.0               NaN               NaN             0.0
3               6.0               4.0               6.0             0.0
4               1.0               6.0               5.0             0.0

   q18_phdimprese  q21_ricbase  q29_brevettisn  q29a_brevettin  q30_papernaz2  \
0             0.0        100.0             0.0               0            NaN
1             0.0        100.0             0.0               0            NaN
2             0.0        100.0             0.0               0            NaN
3             0.0         10.0             0.0               0            NaN
4             0.0         50.0             0.0               0            9.0

   q31_paperint2  q371_lavora  q372_impresa  q372a_impattiva  \
0            3.0          1.0           0.0              0.0
1            2.0          1.0           0.0              0.0
2            NaN          1.0           0.0              0.0
3            NaN          0.0           0.0              0.0
4            NaN          1.0           0.0              0.0

   q373_impabbandono  q375_posizuni  q377_posizauto q3711_impprov  \
0                NaN            1.0             0.0
1                0.0            1.0             0.0
2                NaN            0.0             0.0
3                0.0            0.0             0.0
4                0.0            1.0             1.0

   q3713_impcorso  q3715_impaddetti  q379_impanni  q53a_uniimp  q53b_uniimp  \
0             0.0               NaN           NaN          NaN          NaN
1             0.0               NaN           NaN          NaN          NaN
2             0.0               NaN           NaN          NaN          NaN
3             0.0               NaN           NaN          6.0          5.0
4             0.0               NaN           NaN          NaN          NaN
```

```
   q53c_uniimp  q53d_uniimp  q53e_uniimp  q53f_uniimp  q53g_uniimp  \
0          NaN          NaN          NaN          NaN          NaN
1          NaN          NaN          NaN          NaN          NaN
2          NaN          NaN          NaN          NaN          NaN
3          4.0          6.0          4.0          5.0          6.0
4          NaN          NaN          NaN          NaN          NaN

   q53h_uniimp  q53i_uniimp  dimensione  med_school  polytech  geo_n  geo_c  \
0          NaN          NaN         4.0         1.0       0.0    0.0    1.0
1          NaN          NaN         3.0         1.0       0.0    0.0    0.0
2          NaN          NaN         4.0         1.0       0.0    0.0    1.0
3          5.0          6.0         4.0         1.0       0.0    0.0    0.0
4          NaN          NaN         3.0         1.0       0.0    0.0    1.0

   geo_s  public_uni  vqr_average  utt_sn2006  utt_mission_d2006  disocc2006  \
0    0.0         1.0         0.99         1.0                1.0    7.188093
1    1.0         1.0         0.78         1.0                1.0   11.048825
2    0.0         1.0         1.05         1.0                1.0    4.425923
3    1.0         1.0         0.80         1.0                1.0   18.300277
4    0.0         1.0         0.83         1.0                1.0    7.188093

   area1  area2  area3  area4  area5  area6  area7  area8  area9  area10  \
0      0      0      1      0      0      0      0      0      0       0
1      0      0      0      0      0      0      0      1      0       0
2      0      0      0      0      0      0      0      0      0       1
3      0      0      0      0      0      0      0      0      0       1
4      0      0      0      0      0      0      0      0      0       0

   area11  area12  area13  area14 q378_imptipo_clean  so_reg2006  spinoff0506
0       0       0       0       0                          1.0          0.0
1       0       0       0       0                          1.0          3.0
2       0       0       0       0                          0.0          3.0
3       0       0       0       0                          0.0          0.0
4       0       1       0       0                          1.0          2.0
```

Package `pandas` also has utilities to read Excel files (requires `openpyxl` or `xlrd`):

```
xlsx_path <- system.file("extdata", "italy-nuts-codes.xlsx", package = "edashop")
```

```
xlsx_path = r.xlsx_path
imported_xlsx = pd.read_excel(xlsx_path)
print(imported_xlsx.head())
```

```
                NUTS_3 N3_Code    NUTS_2 N2_Code           NUTS_1 N1_Code
0               Torino   ITC11  Piemonte    ITC1  Northwest Italy     ITC
1             Vercelli   ITC12  Piemonte    ITC1  Northwest Italy     ITC
2               Biella   ITC13  Piemonte    ITC1  Northwest Italy     ITC
3  Verbano-Cusio-Ossola   ITC14  Piemonte    ITC1  Northwest Italy     ITC
4               Novara   ITC15  Piemonte    ITC1  Northwest Italy     ITC
```

As a last example, we use `geopandas` to read a shapefile (which requires `geopandas` and its dependencies `fiona`, `shapely`, etc.):

```r
shp_path <- system.file("extdata", "nuts2.shp", package = "edashop")
```

```python
shp_path = r.shp_path
imported_shp = gpd.read_file(shp_path)
print(imported_shp.head())
```

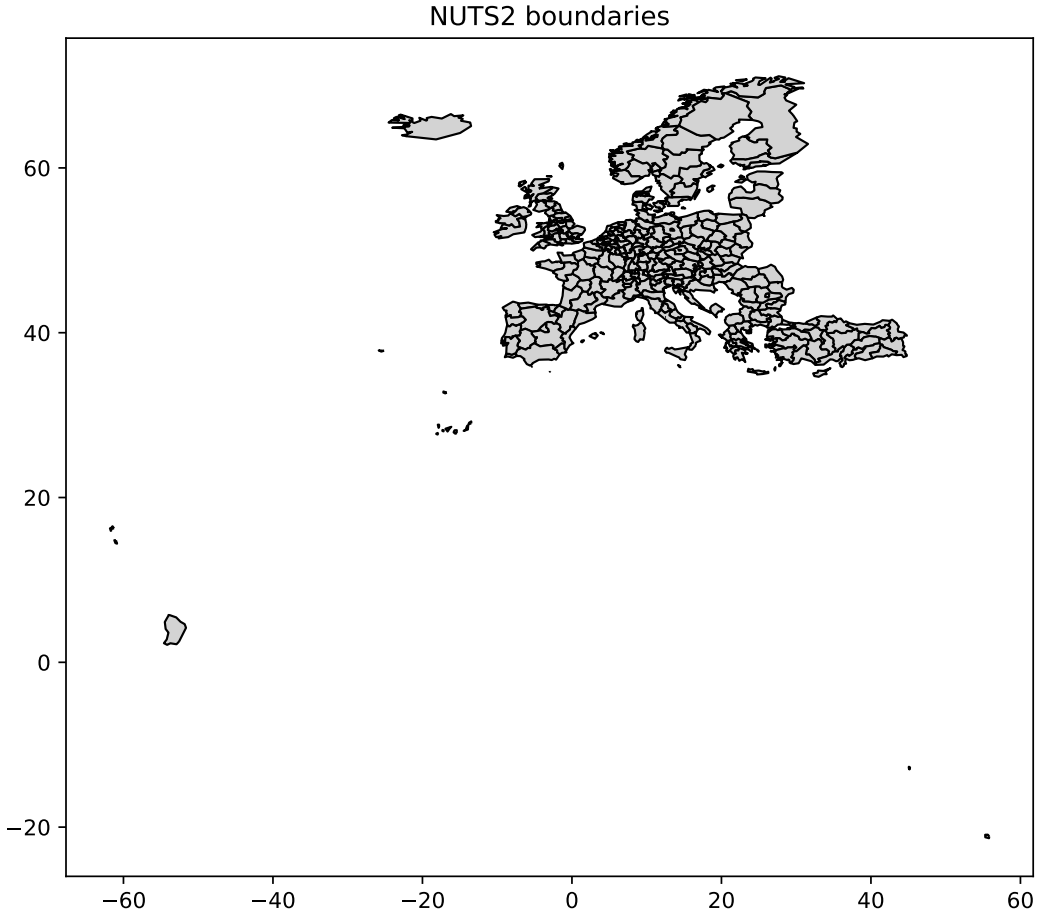
```
  NUTS_ID  LEVL_CODE CNTR_CODE         NAME_LATN         NUTS_NAME   FID  \
0    AT11          2        AT   Burgenland (AT)   Burgenland (AT)  AT11
1    AT12          2        AT  Niederösterreich  Niederösterreich  AT12
2    AT13          2        AT              Wien              Wien  AT13
3    AT21          2        AT           Kärnten           Kärnten  AT21
4    AT22          2        AT        Steiermark        Steiermark  AT22


0
1  POLYGON ((15.54200 48.90800, 15.75400 48.85200, 16.94000 48.61700, 16.95000 48.53600, 16.85100 48.43
2
3
4                                                                                 POLYGON ((16.17200 47.
```

Hands-on practice

If you brought a data file of your own in an external format, try reading it here. Then, you can play with it.

As an example, I am going to plot the shapefile that we just read. geopandas objects can be plotted directly with `.plot()`:

```python
fig, ax = plt.subplots(figsize = (8, 8))
imported_shp.plot(ax = ax, color = 'lightgrey', edgecolor = 'black')
ax.set_title("NUTS2 boundaries")
plt.show()
```

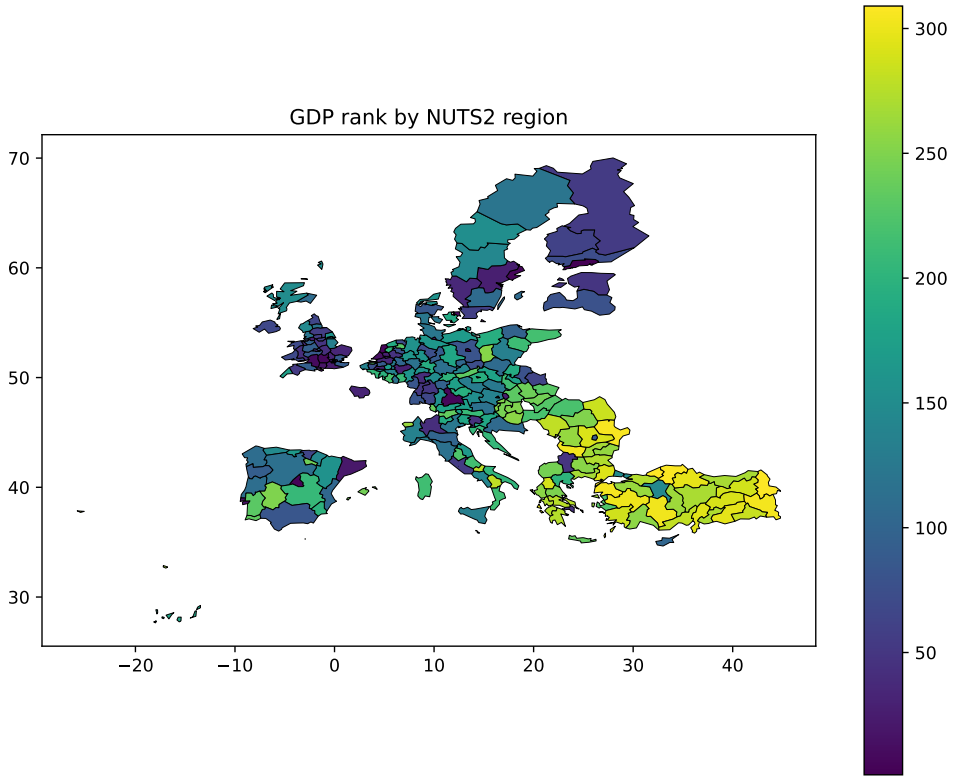


Next we can join the geometry of the boundaries to the rankings of industrial readinness, and convert to `geopandas` GeoDataFrame:

```
# Merge the CSV data with the shapefile on NUTS_ID
i40_rankings = imported_shp.merge(imported_csv, on = 'NUTS_ID', how = 'left')

# Convert to GeoDataFrame
i40_rankings = gpd.GeoDataFrame(i40_rankings, geometry = 'geometry')
```

Now we can map something else, for instance, by encoding it to the color of the polygons (choropleth map):

```
fig, ax = plt.subplots(figsize = (10, 8))
i40_rankings.plot(column = 'gdp_rank', cmap = 'viridis', legend = True,
                  edgecolor = 'black', linewidth = 0.3, ax = ax)
ax.set_title("GDP rank by NUTS2 region")
plt.show()
```



A great resource to learn about working with spatial data in Python is the Geopandas documentation and the book Geographic Data Science with Python by Rey, Arribas-Bel, and Wolf.

Some additional notes on working with Rmarkdown

Quarto is a great invention. It implements principles of literate programming and enforces good discipline when writing and documenting not only code but also analysis processes.

Here is some additional information of value. The chunks of code that we use to communicate with the computer accept a number of options that stipulate the behavior of the chunk. Chunk options are written inside the curly brackets. They can also be named:

```
# Option eval controls whether the chunk is evaluated (i.e., run).
# This chunk of code can only be run manually, by clicking the play icon,
```

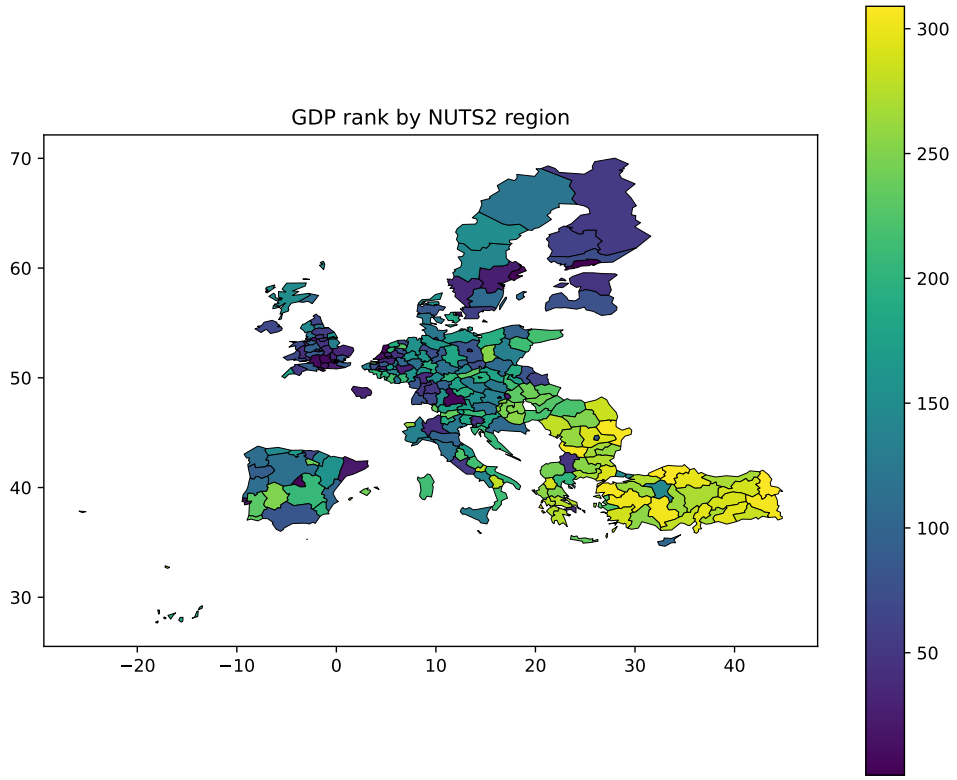
```
# but will be ignored when running the document or when knitting
help(pd.read_csv)
```

Some chunks, like eval, control whether the code is run. Others, like echo, control whether the chunk itself is printed in the output document when knitting:

```
This chunk of code will not be printed in the output. The result of running the code will
```

Chunk `include` is used to include code and results of the code, or to suppress all. For example, when reading files or loading data, we may not wish to have those things printed in the output document:

For example, let us say that we wish to show a plot in a paper or report, but we do not need to show the code:



Quarto is a great format for creating reproducible research reports and papers. These are some examples (originally in `R`, but the same principles apply in `Python` with `Quarto`):

- Paez, Antonio, et al. "A spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain." Geographical analysis 53.3 (2021): 397-421. The repository with the reproducible document is here: https://github.com/paezha/covid19-environmental-correlates
- Paez A, Higgins CD (2021) The Accessibility Implications of a Pilot COVID-19 Vaccination Program in Hamilton, Ontario. Findings (doi.org/10.32866/001c.24082). The repository with the reproducible document is here: https://github.com/paezha/Accessibility-Pharmacies-Hamilton-Vaccines

- Higgins, C.D., Páez, A., Kim, G., Wang, J. (2021) Changes in accessibility to emergency and community food services during COVID-19 and implications for low income populations in Hamilton, Ontario. Social Science and Medicine, 291:114442 (doi.org/10.1016/j.socscimed.2021.114442). The repository with the reproducible document is here: https://github.com/paezha/Accessibility-Food-Banks-Hamilton

For your practice, you can create a new `Quarto` document and write a small exploratory data analysis report using your own data set or one of the data sets provided.

Happy analysis!

Practice

1. Create a new Quarto document (`File` > `New File` > `Quarto Document` in `RStudio`, or use the command line with quarto create project). Choose a template of your liking.

2. Transfer your data analysis exercise to the template and write a small report.