Chair of Innovation Economics

Technische Universität Berlin

Introduction to Python for data scientists

workshop, 20.08.2019

Introduction to Python for data scientists

Final exercise:

Dataset: International Energy Statistics
Global energy trade & production 1990-2014

The Energy Statistics Database contains comprehensive energy statistics on the production, trade, conversion and final consumption of primary and secondary; conventional and non-conventional; and new and renewable sources of energy.

https://www.kaggle.com/unitednations/international-energy-statistics/
http://data.un.org/Explorer.aspx

Introduction to Python for data scientists

Final exercise:

1. Explore the dataset:
   a) How many countries are represented? (243)
   b) Which years are represented? (1990-2014)
   c) What are the most common (least common) categories?
   d) Which kinds of bio-fuel-based commodity transactions are represented?
   e) How many countries are represented after 2010? (231)

Introduction to Python for data scientists

Hints:

- Use the notebook “Final exercise workshop 1” to solve the tasks

1a,b,c) Use `value counts()` to aggregate the countries, or `unique()`

1d) Check if x starts with “Bio” using `x.startswith(“Bio”)`

1e) You can filter a DataFrame by its column ‘c’ like this: `df[df[‘c’]>2000]`

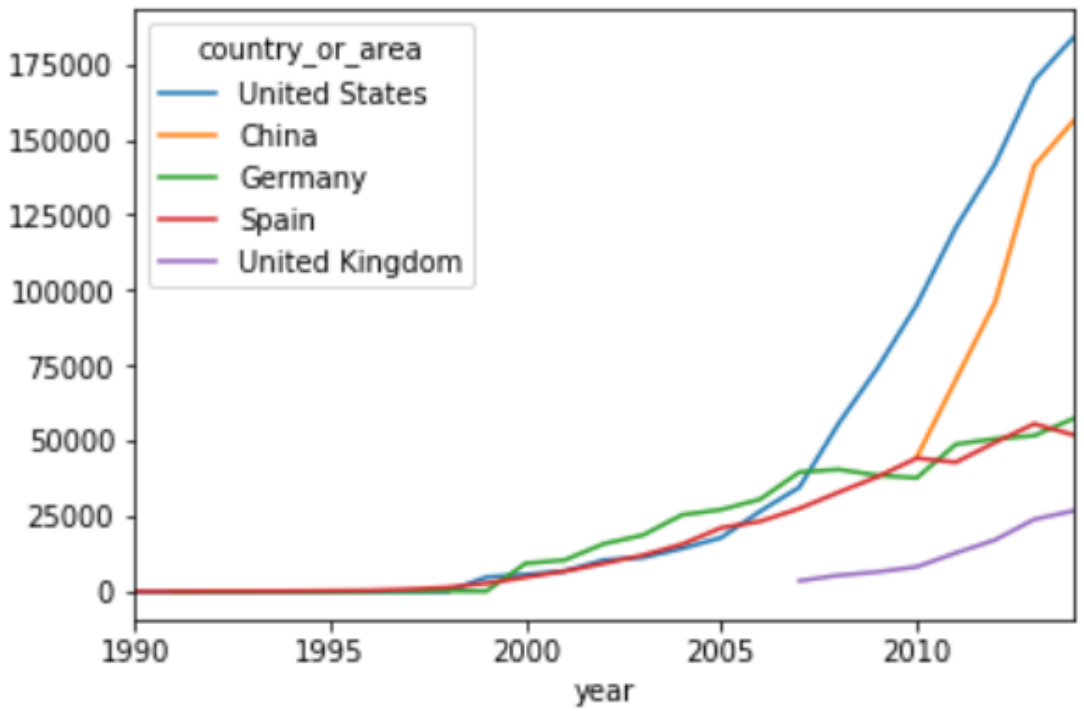
Introduction to Python for data scientists

Final exercise:

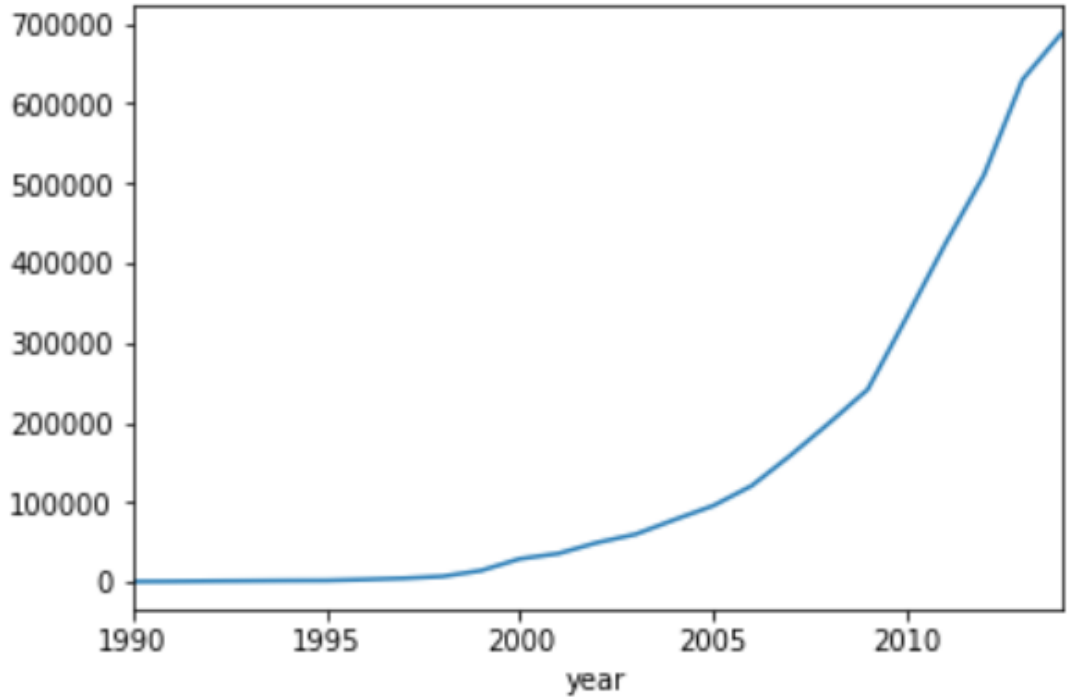
2. Wind energy (commodity_transation “Wind – Main activity”)
   a) Which country is the largest producer of wind energy?
   b) Which country is the largest producer of wind energy after 2010? (US)
   c) Plot the production of wind energy for all years for the TOP5 producers in 2014
   d) Plot the total worldwide wind production for all years
   e) Which countries have the biggest growth from 2010 to 2014 (in %)? Plot a barchart with the TOP10.

Introduction to Python for data scientists

2c)

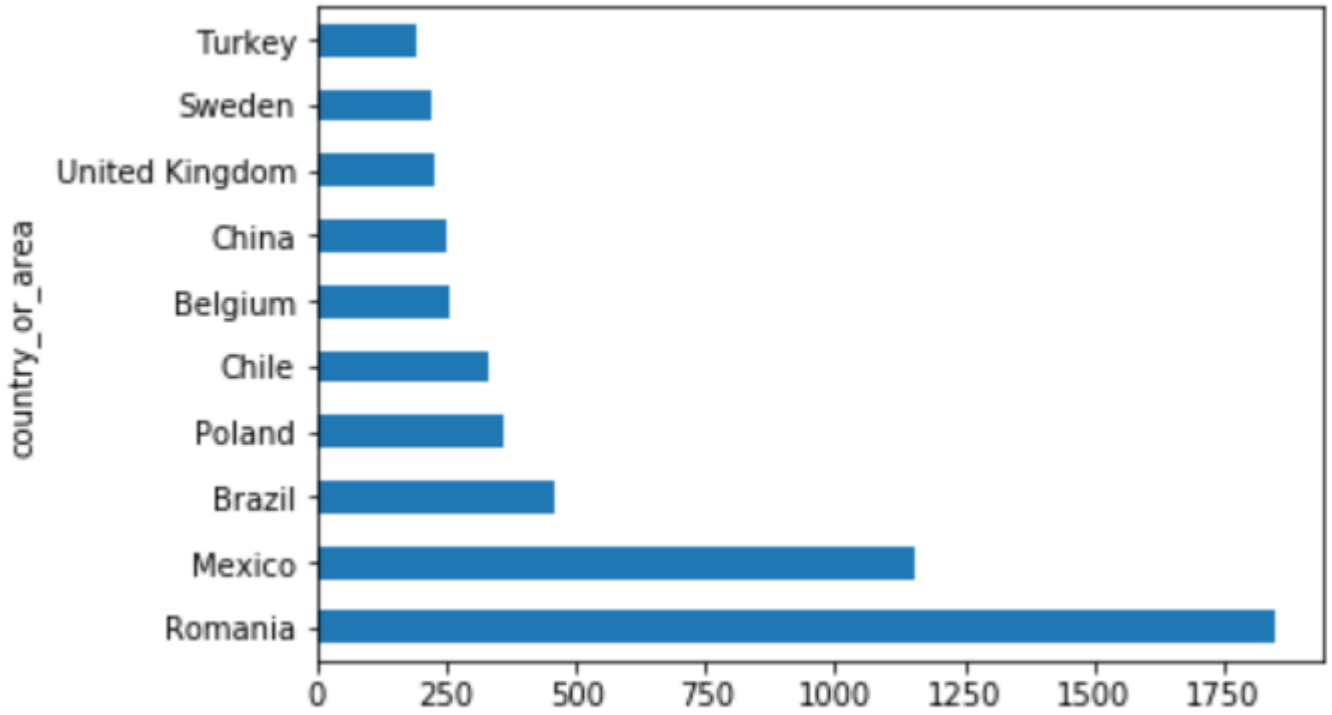


2d)



Introduction to Python for data scientists

2e)



Introduction to Python for data scientists

Hints:

2)

- Reduce your dataset to wind-data: `energy[energy['commodity_transaction']=="Wind – Main activity"]`
- Use `pivot_table()` to restructure the energy data to your needs, e.g. `pivot_table(index="country_or_area", columns="year", values="quantity")`
- Use `sort_values(by=year, ascending=True/False)` to sort the table
- Use `apply(lambda x: x…, axis=0/1)` to calculate custom aggregations (e.g. a growth rate)
- Watch out: column names can be strings (“2014”) or integers (2014)

Introduction to Python for data scientists

Tutorials:

- The official Python tutorial: https://docs.python.org/3/tutorial/
- Introduction to Pandas: https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html
- An overview over available Python tutorials: https://hackr.io/tutorials/learn-python
- E.g.:
  - **Python Tutorial for Beginners [Full Course] 2019** https://www.youtube.com/watch?v=_uQrJ0TkZlc&ref=hackr.io
  - **Welcome to Python for you and me** https://pymbook.readthedocs.io/en/latest/index.html
  - Udemy course