M1 Mini-Assignment 1

You are given 2 datasets from https://nomadlist.com/ - A community page for remote workers worldwide.

- The *trips* data holds ~46k individual trips of travellers on the platform
  https://github.com/SDS-AAU/M1-2019/raw/master/data/trips.csv

- *People* contains some personal information on 4k travelers
  https://github.com/SDS-AAU/M1-2019/raw/master/data/people.csv

- Finally, you find a *countrylist* file that holds countrycodes, contrynames and region-associations
  https://github.com/SDS-AAU/M1-2019/raw/master/data/countrylist.csv

Your solution approach is more important than the results obtained!
Comment your notebook well, explaining all the steps of your analysis. Small technical explanations can go as comments in the code. Broader explanations should be inserted as markdown cells.
Remember that notebooks execute sequentially.

Submission: Wednesday 11.9. 12:00. Peergrade.io (link + submission details will be sent out on Monday, 9.9)

1. Preprocessing
   a. Trips: transform dates into timestamps (note: in Python, you will have to 'coerce' errors for faulty dates)
   b. Calculate trip duration in days (you can use loops, list comprehensions or map-lambda-functions (python) to create a column that holds the numerical value of the day. You can also use the "datetime" package.)
   c. Filter extreme (fake?) observations for durations as well as dates - start and end (trips that last 234565 days / are in the 17th or 23rd century)
      The minimum duration of a trip is 1 day!
      Hint: use percentiles/quantiles to set boundaries for extreme values - between 1 and 97, calculate and store the boundaries before subsetting.
      Rhint: Use percent_rank(as.numeric(variable)) to create percentiles
   d. Join the countrylist data to the trips data-frame using the countrycode as a key
   e. [*Only for python users*] Set DateTime index as the start date of a trip

2. People
   a. How many people have at least a "High School" diploma?
      Hint: For this calculation remove missing value-rows or fill with "False"
   b. How many people working with "Software Dev" have a "Master's Degree"?

c. Who is the person with a Master's Degree that has the highest number of followers? [Explore who this person is. :-)]

3. Trips
    a. Which country received the highest number of trips?
    b. Which country received the highest number of trips in 2017? Use the start of trips as a time reference. (python: use datetimeindex created in 1 as a selector)?
    Rhint: Use functions from lubridate package to extract year.
    c. Which is the country in 'Eastern Asia' where travellers spent on average least time when going there? Provide a visualization.
    d. Do nomads that indicate working in “Software Dev” tend to have shorter or longer trips on average?
    e. Visualize over-time median trip duration overall (bonus: and split by world-region).
    You will get a weird looking plot :-)
    Hint: Python – resample by week (‘W’) and calculate the size of observations.
    Rhint: Use the floor_date function to reset dates by week