M1 – Peergrade assignment 1

You are given 2 datasets from https://nomadlist.com/ - A community page for remote workers worldwide. Further, you are provided with a countries list, a dataset containing information on countries, regions and countrycodes.

- **Trips** data: holds ~46k individual trips of travelers on the platform.
  URL: https://sds-aau.github.io/SDS-master/M1/data/trips.csv

- **People** data: contains some personal information on 4k travelers
  URL: https://sds-aau.github.io/SDS-master/M1/data/people.csv

- **Country** data: Holds countrycodes, contrynames and region-associations
  URL: https://sds-aau.github.io/SDS-master/M1/data/countrylist.csv

Use the pandas read csv function to read the URLs into your working notebook, as shown in SDS examples.

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: Thursday 10.09.2020 23:59:00. Peergrade.io (link + submission details will be available by Wednesday, 09.09.2020.)

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 a least a “High School” diploma?
   Hint: For this calculation remove missing value-rows or fill with “False”.

   b. How many “Startup Founders” have attained a “Master’s Degree”? Bonus: compared to people who don’t have a formal higher education (e.g. by using the “False” occurrences)?

   c. Who is the person with a Master’s Degree that has the highest number of followers? Bonus: Explore the individual further, what else can you find out?

3. **Trips**

   a. Which country received the highest number of trips? – And which the lowest?

   b. Which region received the highest number of trips in 2017? Use the start of trips as a time reference.

   c. Which country in “Western Europe” did travelers spent least time? – Provide visualization

   d. Do nomad Startup Founders tend to have shorter or longer trips on average?

   e. visualize over-time median trip duration overall (bonus: and split by world-region) The plot will look weird ^^. PyHint: Resample by week (‘W’) and calculate the size of observations. RHint: Use the floor_date function to reset dates by week.