Mini-project #2 – Tools

NAYA – Data Scientist Professional – November 2016

Background

In this project we will explore and visualize some facts about mountains and peaks around the world. The project is entirely based on the data available in the site PeakWare.com, which entitles itself as the "World Mountain Encyclopedia".

Submission

You should submit 3 files:

1. A py or a notebook file with the implementation of part I
2. A csv file with the result of part I
3. A py or a notebook file with the implementation of part II

Part I – pre-processing

Do whatever you need to create a DataFrame with the structure as illustrated below. Make sure the dtypes are reasonable, namely numbers are numbers and strings are strings.

	Continent	Country	Difficulty	Height [f]	Height [m]	Latitude	Longitude	Nearest major airport	Year first climbed
Name									
Deadwood Peak	North America	United States	scramble	6290	1917	46.881610	-121.519100	Seattle-Tacoma	None
Rampart Ridge	North America	United States	scramble	5870	1789	47.411000	-121.344400	Seattle-Tacoma	None
Montana Peak	North America	United States	scramble	6907	2105	61.900000	-149.000000	Anchorage	None
Tschingelhorn	Europe	Switzerland	basic snow/ice climb	11732	3576	46.478694	7.848194	Zurich or Geneva	1865
Pelister	Europe	Macedonia	walk up	8533	2601	41.003611	21.186667	Ohrid, Skopje	None
Great End	Europe	United Kingdom	walk up	2985	910	54.464900	-3.194140	Manchester	None
Columbia Peak	North America	United States	scramble	7172	2186	47.961537	-121.361332	Seattle-Tacoma	1897
Oestliche Knotenspitze	Europe	Austria	scramble	10174	3101	47.050000	11.166700	Innsbruck/Munich	None
Cougar Peak	North America	United States	scramble	7894	2406	42.307101	-120.631078	Eugene Airport	None
Mount Saint Helens	North America	United States	walk up	8364	2549	46.197800	-122.191000	Seattle, Washington or Portland, Oregon	1853

Notes:

- The order of the columns is not important.
- The names of the peaks are the index of the DataFrame.
- Missing data
    - Not all the peaks have all the features. Put None where the data is missing.
    - If only one of the elevations is given, then fill the missing data.
    - If both elevations are missing, then drop the peak record.
- Data manipulation
    - If a peak is listed with more than a single country, then use the first country.
    - If the date of the first climbing is recorded as an irregular date, then try to manipulate it so that only the year data is preserved. If the data is too obscure, then put None instead.
    - Note that some terms are written differently for different peaks. Use the *lower()* method to avoid mistakes.

Tips:

- There are several thousands of peaks in the website, so it might take a relatively long time to retrieve the data. Experiment with smaller datasets (e.g. only Antarctica) before you run over the entire dataset.
- The textual data retrieved by the *BeautifulSoup* module is sometimes an instance of the *NavigableString* class. You should apply the *unicode()* converter to make it more manageable.
- The numerical data of the heights contains ',' (comma) for thousands separation, preventing the simple *int()* converter from working. The simplest work-around is to apply the *string.replace(",", "")* method.

Part II – exploration

Answer the following questions and create the described graphs.

1. How many countries are listed?
2. Which mountains from Israel are listed?
3. Make a histogram of the peaks heights in Europe.
4. Which country has the highest number of peaks above 6000m?
5. Make a pie chart for the number of peaks in each continent.
6. Sort the continents by their average peak height.
7. Find the highest mountain in each continent.
8. How many peaks are in “islands” countries? which is the highest of them?
9. Which country has the largest number of peaks listed? And per continent?
10. What is the first mountain that was climbed in each century?
11. Of all the peaks with climbing difficulty “walk up”, which is the highest?
12. How many peaks are there on the equator (no more than 1 degree away from it?)
13. Find the highest peak for each combination of continent and difficulty.

Further ideas for exploration are more than welcome ☺

Good luck!