ailys

Python Engineer Task

Overview

Imagine you are working for a company that is developing a night lamp for hotels that is connected to the hotel's wifi. Guests at the hotel can control the night lamp in their room using their smartphone. Your company is considering expanding into the European market but is not sure which national market to launch in first. You are tasked with producing a data set that can be used to make an informed decision about which national market is the best fit.

Task

Your task consists of two parts, the first part is to combine two public data sets into a single CSV file that contains 3 columns: "Country Code", "Percentage of individuals online" and "Number of Bed-places".
The "Country Code" column should be distinct and should not contain missing data.

The second part of your task is to somehow visualize the resulting data set, you may visualize the data in whatever way you see fit, both plots and tables are fine. Please produce at minimum 1 and at maximum 3 visualizations in total.

Datasets

The raw data comes from two separate datasets in .tsv.gz (tab-separated,gzip-compressed) file format:

TOUR_CAP_NAT

Description

Number of establishments, bedrooms and bed-places
URL: http://ec.europa.eu/eurostat/web/products-datasets/product?code=TOUR_CAP_NAT

Vocabulary

accommod	Value type (e.g. bedrooms, bed-places)
unit	Value unit (e.g. absolute, percentage)
nace_r2	Classification of economic activity
geo\time	Country Code
2016	The number of bed-places, in 2016

Columns to use

"accommod,unit,nace_r2,geo\time" and "2016 "
Rows to use: The first column is a composite of "accommod", "unit", "nace_r2" and "geo\time" separated by comma (","). You should only use rows where:
"accommod" is "BEDPL"
"unit" is "NR"
"nace_r2" is "I551"

TIN00083

Description

Individuals using mobile devices to access the internet on the move
URL: https://ec.europa.eu/eurostat/web/products-datasets/-/tin00083

Vocabulary

indic_is	Data type identifier, always the constant “I_IUMD”
ind_type	Population segments, e.g. students, males, females etc.
unit	The data unit, percentage of individuals.
geo\time	Country Code
2016	Percentage of individuals usin` mobile devices to access the internet on the move, in 2016.

Columns to use

“indic_is,ind_type,unit,geo\time” and “2016”

Rows to use

The first column is a composite of “indic_is”, “ind_type”, “unit” and “geo\time” separated by comma (“,”). You should only use rows where:
“ind_type” is “IND_TOTAL”

Special values

· The value “: ” signifies missing data.
· Values suffixed with “ b” should be used (after removing the suffix).
· Values suffixed with “ u” are unreliable and should be considered as missing data.
· Values suffixed with “ bu” should be considered as missing data.
· Country codes “EA”, "EU27_2007", "EU27_2020", "EU28" should be treated the same as missing data and should be ignored.

Submission

Your solution must be written in Python and should include
The merged CSV file, with columns “Country Code”, “Percentage of individuals online” and “Number of Bed-places”.
Plots and/or tables for visualizing the data as separate files.
The code you used to produce the merged CSV and plots/tables.
You can either email your solution as a zip file directly to recruit@solidware.io or upload your solution to github, gitlab or bitbucket and email the link to your repository to recruit@solidware.io.

Notes

If the instructions are unclear do not hesitate to ask us for clarification!
There is no need to make the solution advanced. Doing basic things should be enough.
A tip is to use the pandas library (https://pandas.pydata.org) to help read and work with the data.