CBE 5790 Modeling and Simulation (Autumn 2018)

Homework P3

Deadline for uploading program py-file to Carmen: Tue, Oct 2 at 2:20 PM
No quiz this week.

Problem P3 – Profiling a chemical dataset

Your company routinely uses Excel spreadsheets to store data for large numbers of chemical substances; an example of such a file - the Hansen dataset - can be downloaded from Carmen. Write a Python program that will process this Excel file to calculate a variety of summary statistics after filtering the data using options specified by a user. Your program will then write the results to a new worksheet in the Excel file. Pandas is an excellent tool for this type of work, so your program should make good use of this package.

def chemprofile(*filename, datasheet, sheetname, filters*):

	name	data type	default value	description
parameters	*filename*	string		name of input Excel file (should be the full path if file is not in the same directory as the program)
	datasheet	string	'dataset'	name of worksheet in the input Excel file containing the data to be processed
	sheetname	string	'bosheet'	name of the new worksheet in the input Excel file where the summary statistics will be written
	filters	dict	None	dictionary object containing one or more filters (see below for more details)
out	None			

Note that there is no default value for the input parameter *filename*; this parameter must always be provided when the function is called.

The data fields (column names) in the Excel input file are listed in the table below. The number and order of these columns will not necessarily always be the same; i.e., your program should be able to handle an input file in which the column order is different and/or not all of these data fields are present.

Structure of Excel input file:

Name	Description
CAS_ID	Chemical Abstracts Service ID, a unique identifier for each molecule
SMILES	SMILES is a chemical structure "language" that makes it possible to specify the structure of a molecule using a linear sequence of characters; you won't do anything with SMILES strings in this assignment, but this is a very interesting topic. If you want to understand how it works, ask Bryan or Professor Rathman.
Atoms	total number of atoms, including hydrogens
BondsRot	number of rotatable bonds
HAcc	number of hydrogen-bond acceptor atoms
HDon	number of hydrogen-bond donor atoms

Stereo	number of tetrahedral stereocenters
MW	molecular weight
McGowan	McGowan molecular volume
TPSA	topological polar surface area
Dipole	molecular dipole moment
Polariz	mean molecular polarizability
LogS	base-10 log of aqueous solubility
LogP	base-10 log of octanol-water partition coefficient
Diameter	molecular diameter (maximum distance between two atoms)

Your program must calculate the following summary statistics for each field in the input data table: mean, standard deviation, minimum value, maximum value, median, and the 5%, 10%, 25%, 75%, 90%, and 95% quantiles, skewness and kurtosis. (Note that the median is the 50% quantile.) Your program will write these results to a new worksheet in the Excel file. You should also report the number of compounds on which the statistics are based.

The *filters* input parameter allows a user to filter the data so that summary results are presented only for the subset of compounds that meet the specified criteria. A few examples are given below:

***filters* parameter**	**Summary statistics calculated based on**
None	all compounds in input dataset
{'MW': (50, 400)}	compounds with 50 ≤ MW ≤ 400
{'MW': (None, 400)}	compounds with MW ≤ 400
{'Dipole': (2.4, None)}	compounds with Dipole ≥ 2.4
{'MW': (50, 400), 'LogP': (-2.0, 6.5)}	compounds with (50 ≤ MW ≤ 400) **AND** (-2.0 ≤ LogP ≤ 6.5)

When provided, the *filters* parameter will be a dictionary object. The key for each element in the dictionary object is the name of a data field, and the value is a two-element tuple specifying the lower and upper values that define the range of interest. The value **None** indicates no limit, as illustrated in the examples above. Filtering can be done on any of the data fields except CAS_ID and SMILES.

The *filters* parameter may include filtering criteria for one, two, or any number of data fields. When criteria are specified for multiple fields, the subset of compounds is determined by AND filtering; you do not need to design your program to handle OR filtering.

You can test your function using the Excel file "*Hansen chemical dataset.xlsx*", which can be downloaded from the assignments page. The first worksheet in this file is the type of data your program will be expected to import. The second worksheet gives an example of results your program will export.

What to submit: A file named **p3.py** that contains your function, which should be named **chemprofile**. Be sure to include a docstring at the beginning of your function and proper commenting throughout. We will run your program using various calling option to test the full functionality of your program.