Assignment 3 – Data Analytics with pandas

Total Points: **7**

General Instructions: You must use the Python template files, located within `Assign` subfolder of the `07_Data_Analytics` folder, and then follow the general instructions shown below and the problem specific notes provided in the comments of the files themselves.

You must follow the additional instructions provided in the notes for each of the **7 problems**, worth **1 point** apiece. The assignment will be graded primarily on whether the script for each problem runs and produces the correct result, which must be documented in the provided `Assign3_pandas.xlsx` Excel file using any method that works best for you. Providing the code for writing to CSV or Excel files is NOT required. However, the submission is not considered complete without providing both the Python code and results in the Excel file. Missing, incorrect or inadequately formatted (column splitting and number formatting at the minimum) output in the Excel file will result in 0.1-point deduction for the particular question. If the script does not run for a particular problem, this will result in a 0.2-point deduction for that question. Partial credit will be awarded based on the degree of problem completion and will depend heavily on how accurately you followed the instructions provided in the notes for the question.

As you work through this assignment, execute each script carefully, and verify you got the correct results by comparing against the output provided in the rest of this document. For any errors you encounter, you are expected to try to resolve them yourself first. If, after repeated attempts, the script does not work, message me or the TA and consider coming to office hours.

Problems 1 – 4 are located in `Assign3_US_Energy.py` file.

The overall objective of this set of problems is to analyze the US energy data provided in **US_Energy.csv** file. The file contains information on total energy consumption, production (both in billions of BTUs) and expenditure (in millions of dollars) by state for year 2014. In addition, the GDP (in millions of dollars), estimated population, as well as the region and division a state belongs to, is provided as well.

Problem 1 (**1 point**) List states from **West North Central** division (Midwest region) showing the state name and all the numerical columns. You should get **7 rows** in the result.

	State	GDP	PopEst	TotCons	TotProd	TotExpnd
14	Iowa	170715.00	3107126	1541900	757014	18021.8
15	Kansas	146562.00	2904021	1132354	880650	14646.3
22	Minnesota	320381.25	5457173	1912065	466967	25312.2
24	Missouri	283280.25	6063589	1903839	199586	26719.0
26	Nebraska	110663.25	1881503	864347	401699	10330.1
33	North Dakota	58230.00	739482	640095	3261360	8209.6
40	South Dakota	45600.00	853175	391857	249789	4819.4

Problem 2 (**1 point**) List **Midwest** states with population over **5 million** showing the state name, population, GDP, total consumption, total production and total expenditures. You should get **7 rows** in the result.

	State	PopEst	GDP	TotCons	TotProd	TotExpnd
12	Illinois	12880580	742027.75	4042313	2683815	51550.0
13	Indiana	6596855	324289.00	2931630	1123098	34151.9
21	Michigan	9909877	447221.25	2881550	682788	40740.1
22	Minnesota	5457173	320381.25	1912065	466967	25312.2
24	Missouri	6063589	283280.25	1903839	199586	26719.0
34	Ohio	11594163	588827.75	3809648	1547368	51385.5
48	Wisconsin	5757564	293341.25	1868867	298603	26448.1

Problem 3 (**1 point**) List the states with over 10 million people that either have their total consumption or total production over 4 quadrillion BTUs. Show the state name, region, population, total consumption and total production, sorted descending on population. You should get **5 rows** in the result.
Note: The consumption and production data is in billions of BTUs, and 1 quadrillion = 1,000 trillion = 1,000,000 billion.

	State	Region	PopEst	TotCons	TotProd
4	California	West	38802500	7620082	2413494
42	Texas	South	26956958	12899498	17597105
8	Florida	South	19893297	4121680	553738
12	Illinois	Midwest	12880580	4042313	2683815
37	Pennsylvania	Northeast	12787209	3902434	7087392

Problem 4 (**1 point**) Create a new column called **NetExport** defined as production minus consumption. The states with positive NetExport are "**net producing**", the ones with negative NetExport are "**net consuming**". List all the "net producing" states with the population over **5 million** people. Show the state name, region, population, total production, total consumption, and net export, sorted descending on net export. You should get **4 rows** in the result..

	State	Region	PopEst	TotProd	TotCons	NetExport
42	Texas	South	26956958	17597105	12899498	4697607
37	Pennsylvania	Northeast	12787209	7087392	3902434	3184958
5	Colorado	West	5355866	3041634	1477177	1564457
3	Arkansas	South	6731484	1454325	1114409	339916

Problems 5 – 7 are located in `Assign3_HomeSales.py` file.

The overall objective of this set of problems is to analyze local real estate data provided in **HomeSales.xls** file. The file contains information on house selling prices, their size in square feet, age, and number of features. It also includes two location-related Yes/No variables describing whether or not a house is in the Northeast sector of the town, and whether it is on a corner lot (or not). Another Yes/No variable shows whether or not an offer is pending on a house, and the last column lists its annual property tax . Both the selling price and annual tax are in dollars. Homes with high number of features represents houses with more amenities (what those are exactly is unknown).

Problem 5 (**1 point**) List the houses that are both in the **Northeast sector** and on a **corner lot**. Show the selling price, size, number of features, offer pending and the annual tax, sorted descending on the number of features. You should get **19 rows** in the result.

	Price	SquareFeet	Features	OfferPending	AnnualTax
90	117000	1928	8	No	600
66	205000	2650	7	No	1639
42	215000	2921	6	No	1635
47	215000	2848	6	No	1487
39	156000	1920	5	No	1161
91	159900	2440	5	No	1265
71	123500	1894	5	No	1112
52	215000	2664	5	No	1193
105	129900	2743	5	Yes	1232
59	208000	2600	4	No	1088
86	112500	1710	4	No	800
44	199900	2580	4	No	1732
101	135000	2253	4	No	939
104	125000	2277	4	No	920
63	130000	2000	3	No	1076
40	144900	1710	3	No	1010
94	95500	1565	3	No	648
82	69000	1348	2	No	520
97	89900	1464	2	No	566

Problem 6 (**1 point**) Find the average price, average size, and average age of homes by **Northeast** sector.

NESector	Price	SquareFeet	Age
No	97282.051282	1546.897436	14.666667
Yes	110769.230769	1707.333333	16.217949

Problem 7 (**1 point**) Create a new column called **PPSQF** defined as **Price / SquareFeet**. Find the average price, average size and average PPSQF by **Northeast** sector and **corner lot**. Sort the resulting data frame by average PPSQF descending.

Hint: Remember to reset the index of the resulting data frame before sorting on the average PPSQF.

	NESector	CornerLot	Price	SquareFeet	PPSQF
3	Yes	Yes	149789.473684	2185.000000	67.822695
1	No	Yes	132537.500000	2063.625000	67.052742
2	Yes	No	98203.389831	1553.508475	63.896127
0	No	No	88183.870968	1413.548387	62.847091

Submission: You must submit all Python and data files, so four (4) files altogether, zipped up into a single folder on Canvas by the designated due date. Failing to include the data files will result in 0.5 point deduction off the top.