CHAPMAN University

Department of Computational and Data Sciences
CS501 Introductory Computation for Scientists
Fall 2019
Homework#7

Date Given: Oct 2, 2019 Due Date: Oct 8, 2019

==========

There are 8 problems in this homework assignment. Write a program in Python to solve these problems. Use 'datascience' package to write the Python code.

To install 'datascience' package on your computer, specify the following command at the 'Anaconda prompt' level (Windows) as an administrator.

```
pip install datascience
```

Documentation of the 'datascience' package is available at the following URL.

http://data8.org/datascience

In the first cell of your Jupyter notebook, please load the libraries using the following script.

```
from datascience import *
import numpy as np
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
%matplotlib inline
```

Use the data file "educ_inc.csv" for all problems of this assignment. This file is uploaded on Blackboard available for download adjacent to the homework assignment#7 (HW07.pdf) file.

	A	B	C	D	E	F
1	Year	Age	Gender	Educational Attainment	Personal Income	Population Count
2	1/1/2008 0:00	00 to 17	Male	College, less than 4-yr degree	C: 10,000 to 14,999	1304
3	1/1/2008 0:00	00 to 17	Female	College, less than 4-yr degree	B: 5,000 to 9,999	1565
4	1/1/2008 0:00	65 to 80+	Male	College, less than 4-yr degree	A: 0 to 4,999	1923
5	1/1/2008 0:00	65 to 80+	Female	No high school diploma	H: 75,000 and over	1981
6	1/1/2008 0:00	00 to 17	Female	No high school diploma	D: 15,000 to 24,999	2009
7	1/1/2008 0:00	00 to 17	Male	No high school diploma	F: 35,000 to 49,999	2227
8	1/1/2008 0:00	00 to 17	Male	No high school diploma	E: 25,000 to 34,999	2606
9	1/1/2008 0:00	00 to 17	Male	College, less than 4-yr degree	D: 15,000 to 24,999	3465
10	1/1/2008 0:00	00 to 17	Male	No high school diploma	D: 15,000 to 24,999	3974

There are 1026 observations in this file. The data is from 2008 to 2014 and displays Age, Gender, Education Attainment, Personal Income and Population Count.

The first 5 variables (Year, Age, Gender, Educational Attainment, Personal Income) are categorical, and the last variable (Population Count) is numerical.

Problem#1: Grouping by a single categorical variable.

Create the following tables which are grouped by a single categorical variable. The second column of the tables below shows the '**count'** of the observations.

Year	count
1/1/08 0:00	145
1/1/09 0:00	149
1/1/10 0:00	147
1/1/11 0:00	151
1/1/12 0:00	145
1/1/13 0:00	144
1/1/14 0:00	145

Age	count
00 to 17	134
18 to 64	448
65 to 80+	444

Gender	count
Female	513
Male	513

Educational Attainment	count
Bachelor's degree or higher	231
College, less than 4-yr degree	258
High school or equivalent	250
No high school diploma	287

Personal Income	count
A: 0 to 4,999	143
B: 5,000 to 9,999	141
C: 10,000 to 14,999	138
D: 15,000 to 24,999	137
E: 25,000 to 34,999	122
F: 35,000 to 49,999	118
G: 50,000 to 74,999	115
H: 75,000 and over	112

==========

Problem#2: Grouping by a single categorical variable with an aggregate function.

Create the following table which are grouped by a single categorical variable 'Year'. The second column of the tables below shows the '**sum'** of the 'Population Count' variable.

Year	Population Count sum
1/1/08 0:00	26532250
1/1/09 0:00	26442817
1/1/10 0:00	26469031
1/1/11 0:00	27160088
1/1/12 0:00	27641508
1/1/13 0:00	27905466
1/1/14 0:00	28215807

==========

Problem#3: Grouping by two categorical variables

Create the following tables which are grouped by two categorical variables 'Year' and 'Age'. The second column of the tables below shows the '**count'** of the observations. Since the 'Year' categorical variable has 7 levels and the 'Age' categorical variable has 3 levels, you should expect to see 21 (7*3=21) rows in you output table.

Year	Age	count
1/1/08 0:00	00 to 17	17
1/1/08 0:00	18 to 64	64
1/1/08 0:00	65 to 80+	64
1/1/09 0:00	00 to 17	21
1/1/09 0:00	18 to 64	64
1/1/09 0:00	65 to 80+	64
1/1/10 0:00	00 to 17	19
1/1/10 0:00	18 to 64	64
1/1/10 0:00	65 to 80+	64
1/1/11 0:00	00 to 17	23
1/1/11 0:00	18 to 64	64
1/1/11 0:00	65 to 80+	64
1/1/12 0:00	00 to 17	19
1/1/12 0:00	18 to 64	64
1/1/12 0:00	65 to 80+	62
1/1/13 0:00	00 to 17	17
1/1/13 0:00	18 to 64	64
1/1/13 0:00	65 to 80+	63
1/1/14 0:00	00 to 17	18
1/1/14 0:00	18 to 64	64
1/1/14 0:00	65 to 80+	63

Problem#4: Grouping by two categorical variables in a 'pivot' table.

Create the following pivot table which are grouped by two categorical variables 'Year' and 'Age'. The pivot table shows the '**count'** of the observations.

Age	1/1/08 0:00	1/1/09 0:00	1/1/10 0:00	1/1/11 0:00	1/1/12 0:00	1/1/13 0:00	1/1/14 0:00
00 to 17	17	21	19	23	19	17	18
18 to 64	64	64	64	64	64	64	64
65 to 80+	64	64	64	64	62	63	63

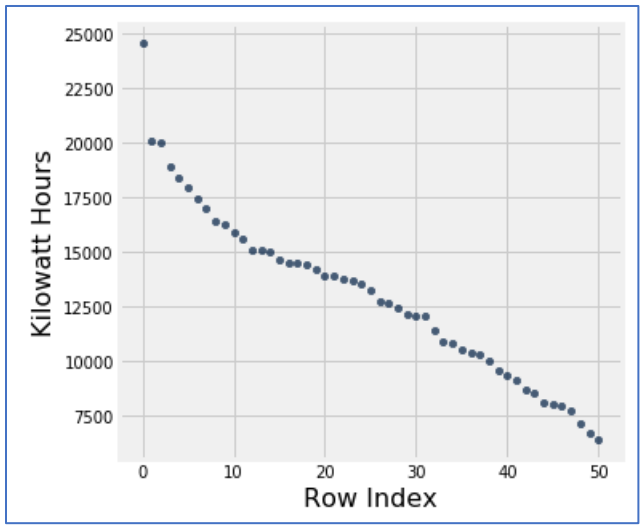
Problem#5: **Join the following 2 tables where table1.a = table2.a**

```
# Problem#5

table1 = Table().with_columns(
     'a', make_array(9, 3, 3, 1),
     'b', make_array(1, 2, 2, 10),
     'c', make_array(3, 4, 5, 6))

print(table1)
print()

table2 = Table().with_columns(
 'a', make_array(9, 1, 1, 1),
 'd', make_array(1, 2, 2, 10),
 'e', make_array(3, 4, 5, 6))

print(table2)
```

```
a    | b    | c
9    | 1    | 3
3    | 2    | 4
3    | 2    | 5
1    | 10   | 6

a    | d    | e
9    | 1    | 3
1    | 2    | 4
1    | 2    | 5
1    | 10   | 6
```

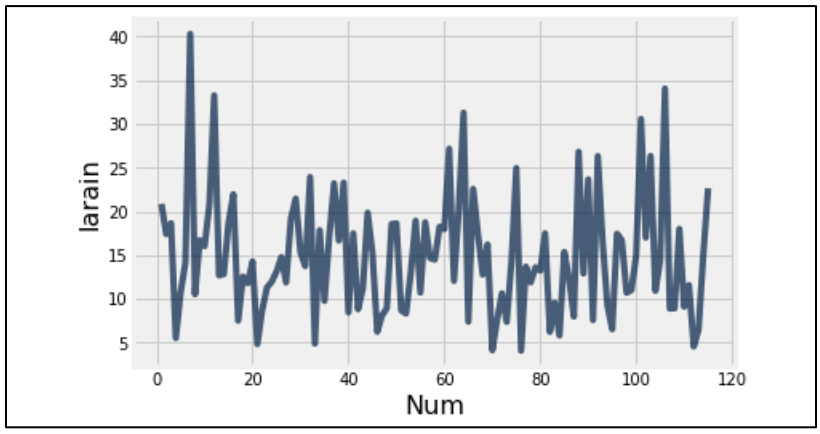
Problem#6: **Create a scatter plot**

The ‘energy.csv’ file contains the energy consumption of all the states in US. Create a scatter plot between the states and the energy consumption of every state.



Problem#7: Create a Line Plot

Create a line plot of the “Los Angeles rain data given in ‘larain.csv’ file.



Problem#8: **Create a bar graph**

In problem#1(a) we grouped the data using ‘Year’ categorical variable.
In problem#2 we grouped the data using ‘Year’ categorical variable with aggregate function ‘sum’.

Draw the bar graph of the tables you created in problem#1(a) and problem#2.