BLU537E 2021-Fall Final Exam

Remarks:

Write the code yourself. ***Cheating is strictly forbidden***.

For each problem write your code in the function format and give the names of the functions as problem numbers, for example for the solution of problem1:

```
def problem1(input):
  return something
```

Put the codes for all problems into one file (jupter notebook file) and name that file using your student username in the following format: badays_blu537e_final.ipynb. The notebook file should definitely contain the outputs of the functions, if applicable. Sample solution file (sample_solution.ipynb) is given to you to show how to organize your solutions.

Also write your name and student number inside the jupyter notebook as well.

Give as much as documentation for your script using comments.

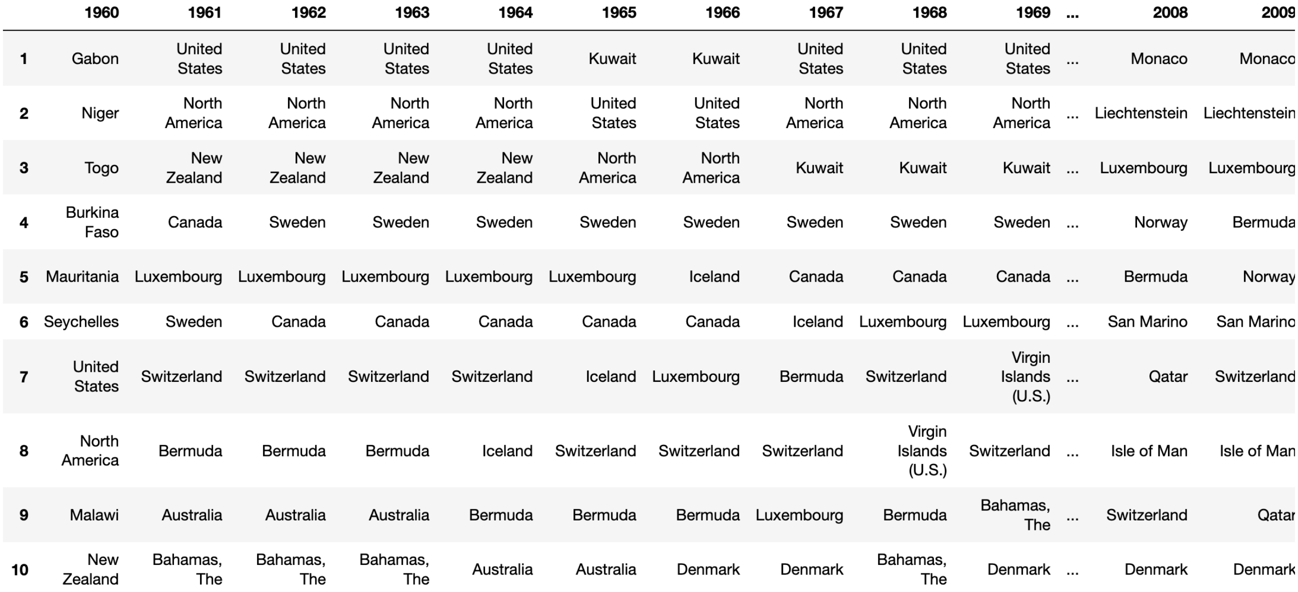
Note1: You can **use** any python library or module for the problems.

Note2: If your homework solution file has problems in structure you can lose up to ***20*** points!!

For example, if you didn't write solutions in a function format or if you did not arrange input arguments properly you may lose points.

Problem 1 (20 Points).

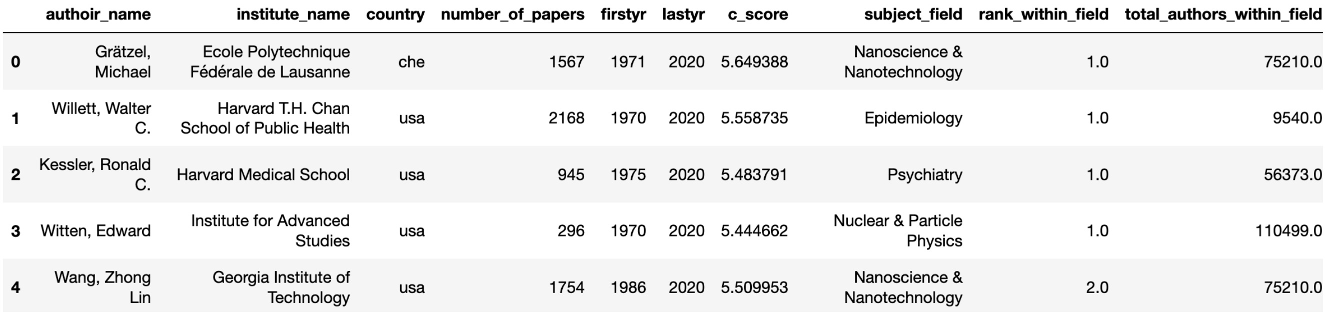
You are given gdp_per_capita.csv file which contains gdp per capita values for countries over the years ranging from 1960 to 2017. Using this file find the top-10 countries having the highest gdp per capita in each year. Produce the following table. Note that index of the table should start from 1. Your function should take gdp_per_capita.csv file as the input.



	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	...	2008	2009
1	Gabon	United States	United States	United States	United States	Kuwait	Kuwait	United States	United States	United States	...	Monaco	Monaco
2	Niger	North America	North America	North America	North America	United States	United States	North America	North America	North America	...	Liechtenstein	Liechtenstein
3	Togo	New Zealand	New Zealand	New Zealand	New Zealand	North America	North America	Kuwait	Kuwait	Kuwait	...	Luxembourg	Luxembourg
4	Burkina Faso	Canada	Sweden	Sweden	Sweden	Sweden	Sweden	Sweden	Sweden	Sweden	...	Norway	Bermuda
5	Mauritania	Luxembourg	Luxembourg	Luxembourg	Luxembourg	Luxembourg	Iceland	Canada	Canada	Canada	...	Bermuda	Norway
6	Seychelles	Sweden	Canada	Canada	Canada	Canada	Canada	Iceland	Luxembourg	Luxembourg	...	San Marino	San Marino
7	United States	Switzerland	Switzerland	Switzerland	Switzerland	Iceland	Luxembourg	Bermuda	Switzerland	Virgin Islands (U.S.)	...	Qatar	Switzerland
8	North America	Bermuda	Bermuda	Bermuda	Iceland	Switzerland	Switzerland	Switzerland	Virgin Islands (U.S.)	Switzerland	...	Isle of Man	Isle of Man
9	Malawi	Australia	Australia	Australia	Bermuda	Bermuda	Bermuda	Luxembourg	Bermuda	Bahamas, The	...	Switzerland	Qatar
10	New Zealand	Bahamas, The	Bahamas, The	Bahamas, The	Australia	Australia	Denmark	Denmark	Bahamas, The	Denmark	...	Denmark	Denmark

Problem 2 (20 Points).

John P. A. Ioannidis and coworkers developed a database which shows top scientists in various scientific fields. You can find the detailed information about the study from the following link: https://doi.org/10.1371/journal.pbio.3000384
You are given worldranking_2020-2.xlsx file showing the top %2 scientists in each field. The table looks like the following picture:



	authoir_name	institute_name	country	number_of_papers	firstyr	lastyr	c_score	subject_field	rank_within_field	total_authors_within_field
0	Grätzel, Michael	Ecole Polytechnique Fédérale de Lausanne	che	1567	1971	2020	5.649388	Nanoscience & Nanotechnology	1.0	75210.0
1	Willett, Walter C.	Harvard T.H. Chan School of Public Health	usa	2168	1970	2020	5.558735	Epidemiology	1.0	9540.0
2	Kessler, Ronald C.	Harvard Medical School	usa	945	1975	2020	5.483791	Psychiatry	1.0	56373.0
3	Witten, Edward	Institute for Advanced Studies	usa	296	1970	2020	5.444662	Nuclear & Particle Physics	1.0	110499.0
4	Wang, Zhong Lin	Georgia Institute of Technology	usa	1754	1986	2020	5.509953	Nanoscience & Nanotechnology	2.0	75210.0

Using this file do the following analyses:

a) Find the number of subject fields in this dataset (4 Points).

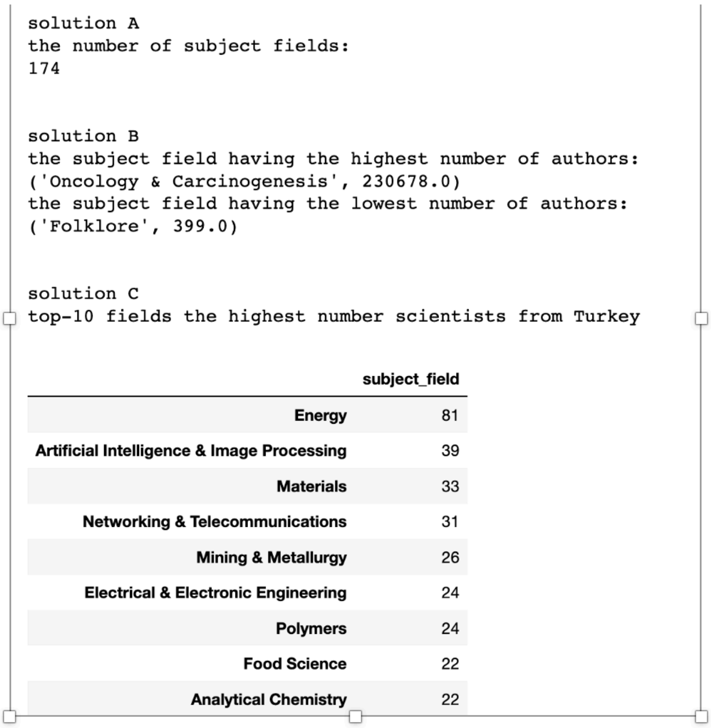
b) Find the subject fields having the the highest and lowest number of authors within that field (4 Points).

c) Find the number of scientists in each field from Turkey. Print top-10 fields the highest number of scientists. (4 Points).

d) First, calculate the seniority level of each scientist by simply substracting firstyr from lastyr. Name this column as "seniority_level". Then, find the number of scientists in each field represented in this data set. You sould not use total authors within the field column. Here, you will calculate how many scientists are given for each field in this dataset. After that, extract data for subject fields having at least 200 scientists. For example, if a subject field has 20 scientists in this data set you will not include that subject field. Then, filter scientists ranked between 1-100 in each subject field. In the final step, calculate minimum, mean, median and maximum for seniority_level and number of papers for each subject field. Display top-10 and bottom-10 tables when data is sorted by mean seniority level. (4 Points).

e) Extract data for engineers. You can search for subject fields having "Engineer" in it. Then, calculate average number of papers for each engineering field. Display this table with total authors within field information. (4 Points).

Your function take worldranking_2020-2.xlsx file as the input. Your output should look like:

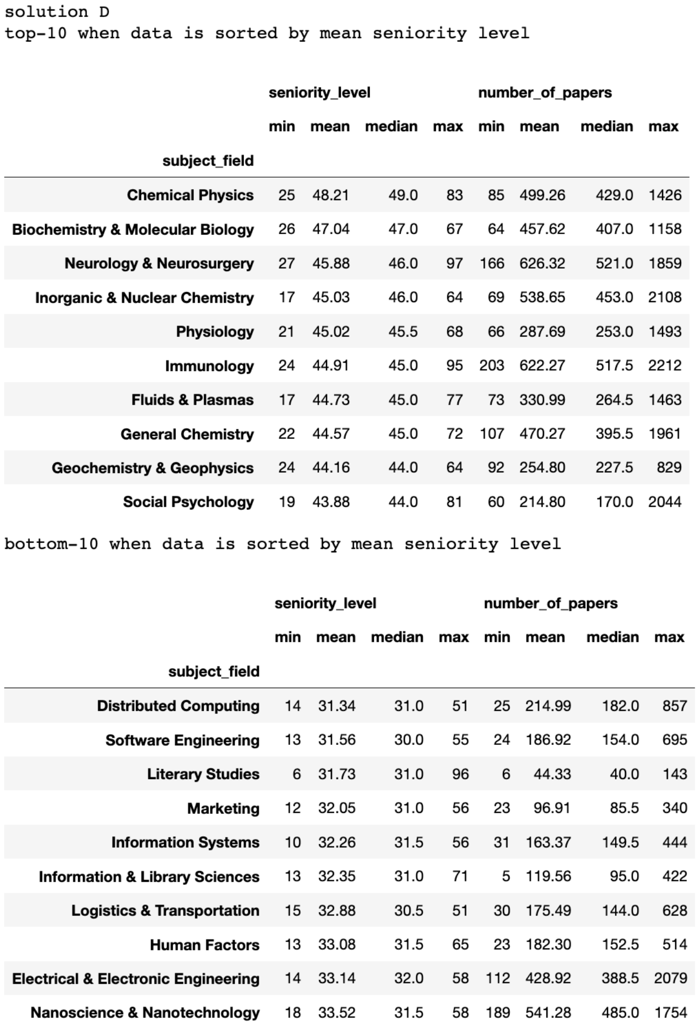


```
solution A
the number of subject fields:
174


solution B
the subject field having the highest number of authors:
('Oncology & Carcinogenesis', 230678.0)
the subject field having the lowest number of authors:
('Folklore', 399.0)


solution C
top-10 fields the highest number scientists from Turkey
```

	subject_field
Energy	81
Artificial Intelligence & Image Processing	39
Materials	33
Networking & Telecommunications	31
Mining & Metallurgy	26
Electrical & Electronic Engineering	24
Polymers	24
Food Science	22
Analytical Chemistry	22



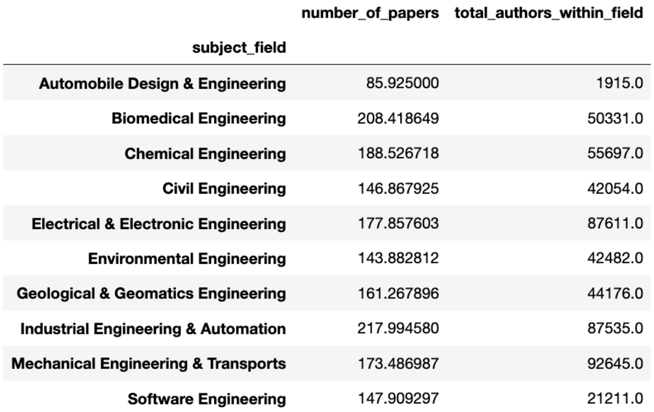
```
solution D
top-10 when data is sorted by mean seniority level
```

	seniority_level				number_of_papers			
	min	mean	median	max	min	mean	median	max
subject_field								
Chemical Physics	25	48.21	49.0	83	85	499.26	429.0	1426
Biochemistry & Molecular Biology	26	47.04	47.0	67	64	457.62	407.0	1158
Neurology & Neurosurgery	27	45.88	46.0	97	166	626.32	521.0	1859
Inorganic & Nuclear Chemistry	17	45.03	46.0	64	69	538.65	453.0	2108
Physiology	21	45.02	45.5	68	66	287.69	253.0	1493
Immunology	24	44.91	45.0	95	203	622.27	517.5	2212
Fluids & Plasmas	17	44.73	45.0	77	73	330.99	264.5	1463
General Chemistry	22	44.57	45.0	72	107	470.27	395.5	1961
Geochemistry & Geophysics	24	44.16	44.0	64	92	254.80	227.5	829
Social Psychology	19	43.88	44.0	81	60	214.80	170.0	2044

```
bottom-10 when data is sorted by mean seniority level
```

	seniority_level				number_of_papers			
	min	mean	median	max	min	mean	median	max
subject_field								
Distributed Computing	14	31.34	31.0	51	25	214.99	182.0	857
Software Engineering	13	31.56	30.0	55	24	186.92	154.0	695
Literary Studies	6	31.73	31.0	96	6	44.33	40.0	143
Marketing	12	32.05	31.0	56	23	96.91	85.5	340
Information Systems	10	32.26	31.5	56	31	163.37	149.5	444
Information & Library Sciences	13	32.35	31.0	71	5	119.56	95.0	422
Logistics & Transportation	15	32.88	30.5	51	30	175.49	144.0	628
Human Factors	13	33.08	31.5	65	23	182.30	152.5	514
Electrical & Electronic Engineering	14	33.14	32.0	58	112	428.92	388.5	2079
Nanoscience & Nanotechnology	18	33.52	31.5	58	189	541.28	485.0	1754

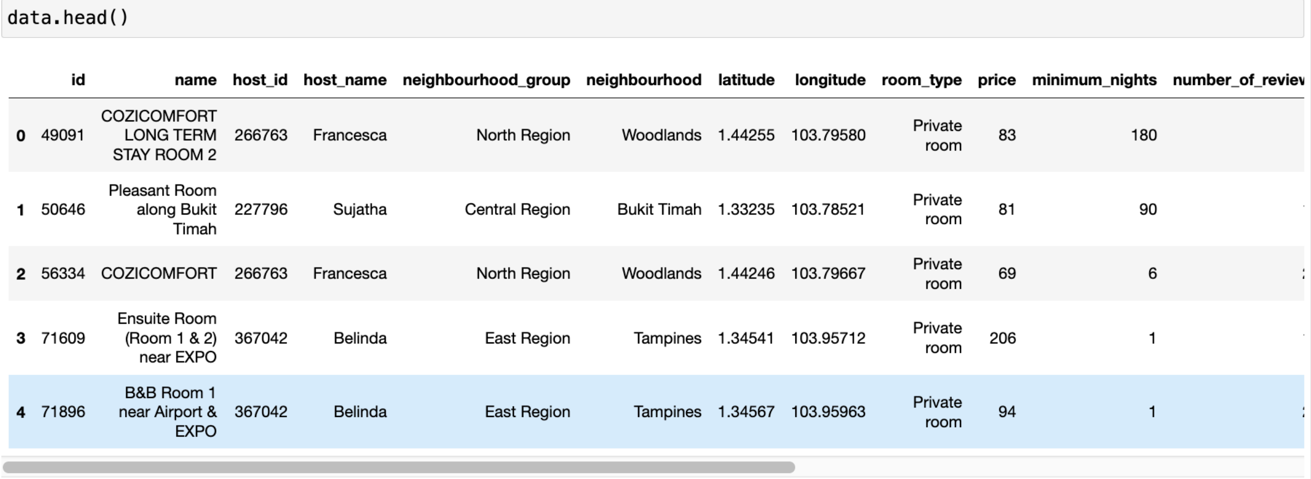
solution E



	number_of_papers	total_authors_within_field
subject_field		
Automobile Design & Engineering	85.925000	1915.0
Biomedical Engineering	208.418649	50331.0
Chemical Engineering	188.526718	55697.0
Civil Engineering	146.867925	42054.0
Electrical & Electronic Engineering	177.857603	87611.0
Environmental Engineering	143.882812	42482.0
Geological & Geomatics Engineering	161.267896	44176.0
Industrial Engineering & Automation	217.994580	87535.0
Mechanical Engineering & Transports	173.486987	92645.0
Software Engineering	147.909297	21211.0

Problem 3 (20 Points).

You are given “listings.csv” file containing Airbnb information for signapure. The data looks like below figure.

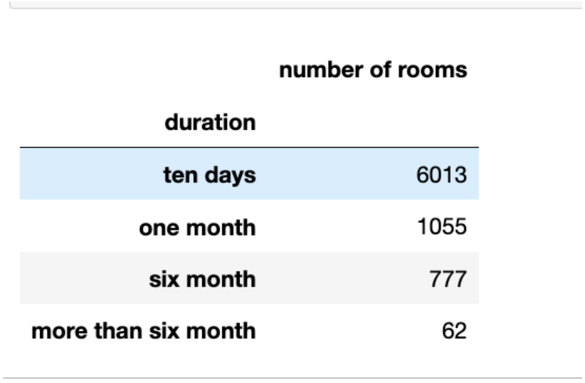


```
data.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviev
0	49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1.44255	103.79580	Private room	83	180	
1	50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1.33235	103.78521	Private room	81	90	
2	56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1.44246	103.79667	Private room	69	6	
3	71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1.34541	103.95712	Private room	206	1	
4	71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	94	1	

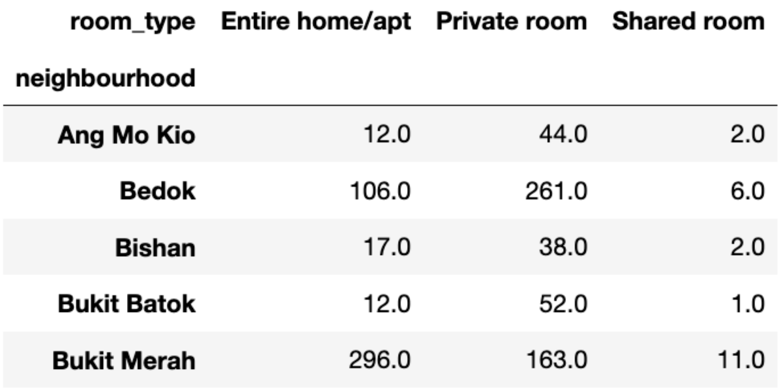
Use this datafile as the input and do the following analyses.

a) divide minimum night information into windows of (0-10), (11-30),(31,180),(181,1000) and give these intervals the following category names ten days, one month, six month and more than six month, respectively. Print the following table which shows the number of rooms for each category. (5 Points)

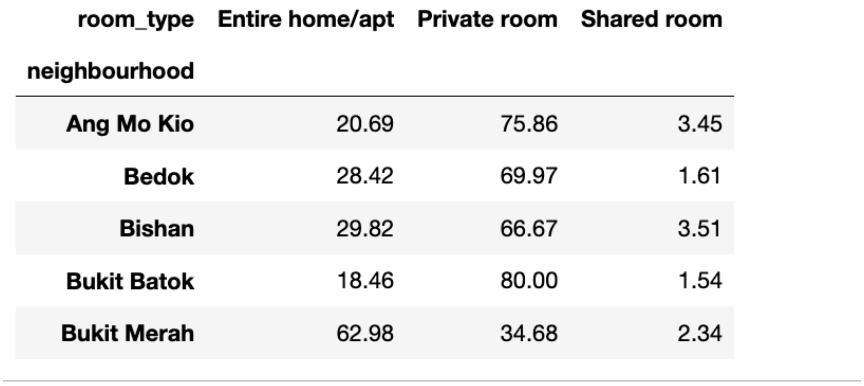


	number of rooms
duration	
ten days	6013
one month	1055
six month	777
more than six month	62

b) Calculate the number of room type for each neighbourhood. Also print the same table in terms of percentages. (5 Points)

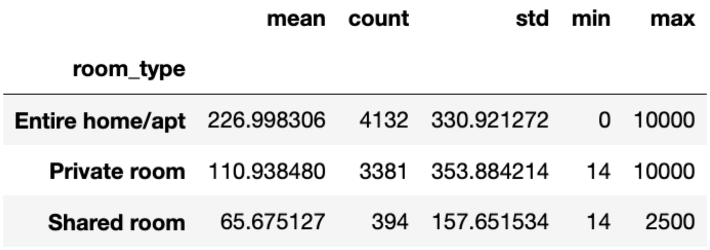


room_type	Entire home/apt	Private room	Shared room
neighbourhood			
Ang Mo Kio	12.0	44.0	2.0
Bedok	106.0	261.0	6.0
Bishan	17.0	38.0	2.0
Bukit Batok	12.0	52.0	1.0
Bukit Merah	296.0	163.0	11.0



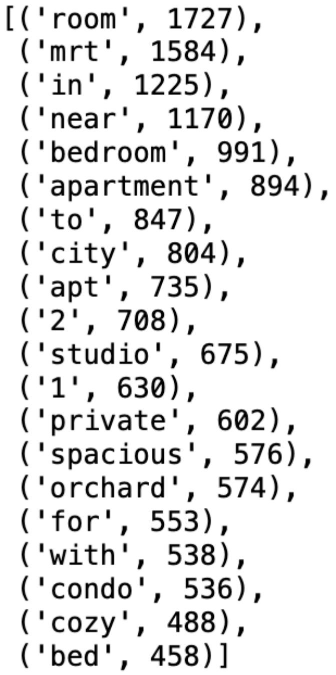
room_type	Entire home/apt	Private room	Shared room
neighbourhood			
Ang Mo Kio	20.69	75.86	3.45
Bedok	28.42	69.97	1.61
Bishan	29.82	66.67	3.51
Bukit Batok	18.46	80.00	1.54
Bukit Merah	62.98	34.68	2.34

c) Print the average, min, max and standart deviation of the prices for each room type. Provide the count of each room type as well. (5 Points)



	mean	count	std	min	max
room_type					
Entire home/apt	226.998306	4132	330.921272	0	10000
Private room	110.938480	3381	353.884214	14	10000
Shared room	65.675127	394	157.651534	14	2500

d) Print most frequent 20 words in the name column. First normalize the names by converting them into lower letter. (5points)



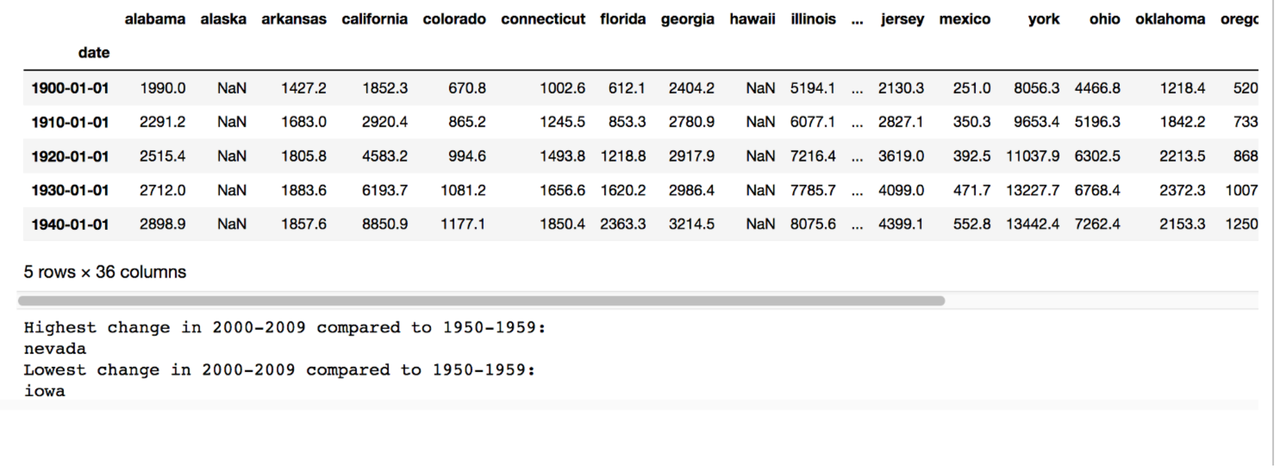
```
[('room', 1727),
 ('mrt', 1584),
 ('in', 1225),
 ('near', 1170),
 ('bedroom', 991),
 ('apartment', 894),
 ('to', 847),
 ('city', 804),
 ('apt', 735),
 ('2', 708),
 ('studio', 675),
 ('1', 630),
 ('private', 602),
 ('spacious', 576),
 ('orchard', 574),
 ('for', 553),
 ('with', 538),
 ('condo', 536),
 ('cozy', 488),
 ('bed', 458)]
```

Problem 4 (20 Points)

You are given files (compressed under us_state_pop_data.zip file) ,showing the population of the States in US as timeseries. For each state, there is a csv file in the data folder. Using these data, do the following tasks. Note that the answer should be in function format as usual (like in the other homeworks) and it should take the data directory as the input.

a) Calculate ten-year average population for each state. Construct a data frame for this table and print the top-five rows (10 Points).

b) Find the states showing the highest and lowest changes in 2000-2009 decade average compared to 1950-1959 decade average (10 Points).



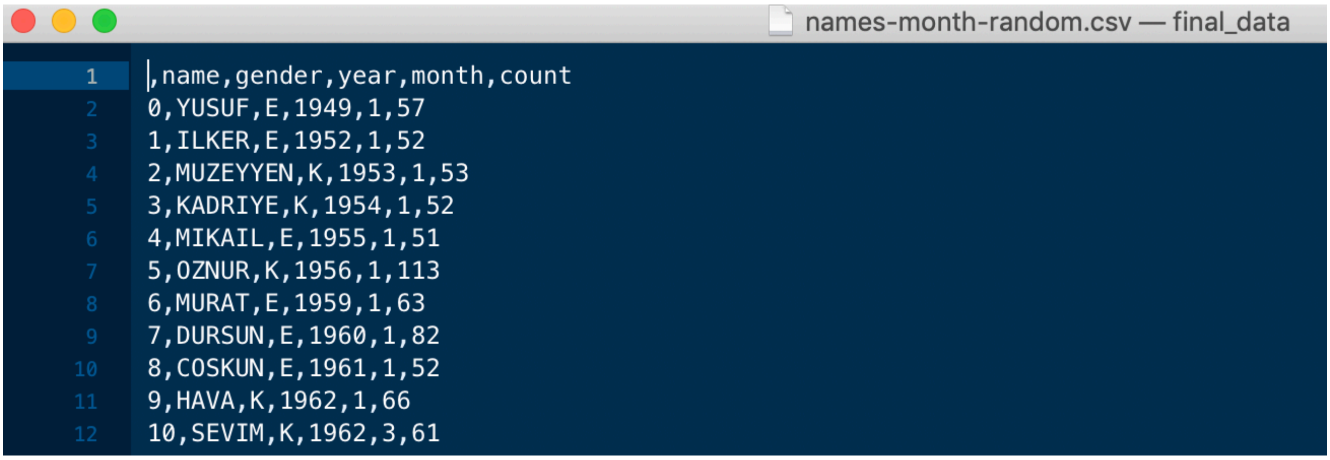
	alabama	alaska	arkansas	california	colorado	connecticut	florida	georgia	hawaii	illinois	...	jersey	mexico	york	ohio	oklahoma	oregc
date																	
1900-01-01	1990.0	NaN	1427.2	1852.3	670.8	1002.6	612.1	2404.2	NaN	5194.1	...	2130.3	251.0	8056.3	4466.8	1218.4	520
1910-01-01	2291.2	NaN	1683.0	2920.4	865.2	1245.5	853.3	2780.9	NaN	6077.1	...	2827.1	350.3	9653.4	5196.3	1842.2	733
1920-01-01	2515.4	NaN	1805.8	4583.2	994.6	1493.8	1218.8	2917.9	NaN	7216.4	...	3619.0	392.5	11037.9	6302.5	2213.5	868
1930-01-01	2712.0	NaN	1883.6	6193.7	1081.2	1656.6	1620.2	2986.4	NaN	7785.7	...	4099.0	471.7	13227.7	6768.4	2372.3	1007
1940-01-01	2898.9	NaN	1857.6	8850.9	1177.1	1850.4	2363.3	3214.5	NaN	8075.6	...	4399.1	552.8	13442.4	7262.4	2153.3	1250

5 rows × 36 columns

```
Highest change in 2000-2009 compared to 1950-1959:
nevada
Lowest change in 2000-2009 compared to 1950-1959:
iowa
```

Problem 5 (20 Points).

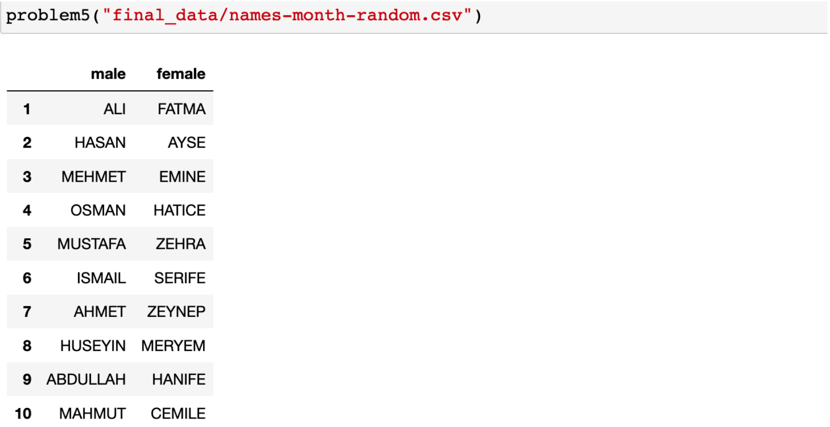
The datafile names-month-random.csv file contains a name statistic for a random sample population. The data set comprises of the names of the people and birthdate information. The dataset is organized as the names, gender, birth year and birth month and the number of people (count). For example, in 1949 january there were 57 babies named as “YUSUF”. In gender column “E” and “K” male and female respectively. Using this dataset do the following analysis:



```
,name,gender,year,month,count
0,YUSUF,E,1949,1,57
1,ILKER,E,1952,1,52
2,MUZEYYEN,K,1953,1,53
3,KADRIYE,K,1954,1,52
4,MIKAIL,E,1955,1,51
5,OZNUR,K,1956,1,113
6,MURAT,E,1959,1,63
7,DURSUN,E,1960,1,82
8,COSKUN,E,1961,1,52
9,HAVA,K,1962,1,66
10,SEVIM,K,1962,3,61
```

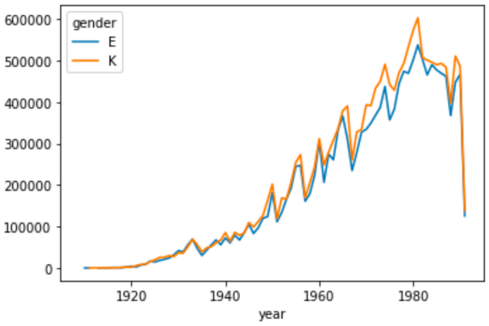
a) Plot the total number of males and females as a function of year. (10 Points)

b) Find top-10 most frequent names for males and females. Give the result in a table format as shown below. Note that the table is in descending sorted order. For example, the most frequent name for males is ALI. (10 Points)



```
problem5("final_data/names-month-random.csv")
```

	male	female
1	ALI	FATMA
2	HASAN	AYSE
3	MEHMET	EMINE
4	OSMAN	HATICE
5	MUSTAFA	ZEHRA
6	ISMAIL	SERIFE
7	AHMET	ZEYNEP
8	HUSEYIN	MERYEM
9	ABDULLAH	HANIFE
10	MAHMUT	CEMILE


gender
E
K
year