Data Science - Practice 4 (R Basic III)

Make sure you not only just “write down” the R code but also “explain the answer with your own language”. All answers without explanation will not be accepted.

Problem

Import ‘covid.set.RData’ and ‘Crime.csv’. The prior is about COVID cases in US and the latter is the new DataFrame, which contains crime information in US by the Metropolitan Statistical Area regions.

< Question 1 – covid.set data>

Check first 6 rows of data to see how the data looks like. Write down the answers for the following questions. (1) number of rows, (2) number of variables, (3) number of State (ST_Name), (4) number of county (Countyname)

< Question 2 – covid.set data>

As you noticed, the “absolute” value of COVID cases strongly depends on the region’s size. To consider this so-called “size-effect”, (1) create a variable called “NewCases.pop”, which divides NewCases by population. (2) Write down a R code that can be used to compare the case between “Alabama’s Autauga” and “Alabama’s Baldwin” in 2021 (as shown below). Explain what we can be informed from NewCases.pop instead of NewCases.

Expected Result

	ST_Name	Countyname	year	Confirmed	Deaths	Population	IncidenceRate	NewCases	NewCases.pop
2	Alabama	Autauga	2021	2871865	41785	55869	5140355	6828	0.1222145
5	Alabama	Baldwin	2021	9750561	136367	223234	4367866	26310	0.1178584

< Question 3 – covid.set data>

If you check the new variable “NewCases.pop”, then you will find some strange values (NaN, Inf, -Inf). Find out why such cases happen. Fix the problem if possible, and write down the explanation.

< Question 4 – covid.set data>

Using aggregate() function, find average of NewCases and Deaths per year. The outcome should be a DataFrame as shown below.

Expected Result

	year	NewCases	Deaths
1	2020	6172.298	14299.24
2	2021	10647.574	67792.75
3	2022	7563.024	19411.11

< Question 5 – covid.set data>

Using which() function, find regions that satisfies all three following conditions: 1) indices of regions with “Confirmed” higher than average, 2) indices of regions with “Deaths” higher than average, and 3) indices of regions with “NewCases” lower than average, all in 2020. Show them in an alphabetical order.

Expected Result

```
 [1] "Acadia"      "Charles"     "Dougherty"   "Forrest"
 [5] "Glynn"       "Iberia"      "Jones"       "Kansas City"
 [9] "Monroe"      "Newton"      "Orangeburg"  "Sumter"
[13] "Troup"       "Ulster"
```

< Question 6 – Crime data>

Load ‘Crime.csv’ and assign the data set to variable name called “Crime”. Crime data set contains total 12 variables including "MSA", "ViolentCrime", "Murder", "Rape", "Robbery", "AggravatedAssault", "PropertyCrime", "Burglary", "Theft", "MotorVehicleTheft", "State", and "City". Check first 6 rows and write down the list of things that you find strange.

Expected Result

```
> head(Crime)
  X1     Abilene..TX.M.S.A. X412.5 X5.3  X56 X78.4 X272.8 X3.609.00     X852 X2.493.60 X263.4 TX     Abilene
1  2       Akron, OH M.S.A.  238.4  5.1 38.2  75.2  119.8  2,552.40    575.3  1,853.00  124.1 OH       Akron
2  3      Albany, GA M.S.A.  667.9  7.8 30.4 157.9  471.8  3,894.10 1,099.60  2,652.80  141.7 GA      Albany
3  4      Albany, OR M.S.A.  114.3  2.5 28.2  20.7   63.0  3,208.40    484.6  2,476.10  247.7 OR      Albany
4  5 Albuquerque, NM M.S.A.  792.6  6.1 63.8 206.7  516.0  4,607.80    883.4  3,047.60  676.9 NM Albuquerque
5  6  Alexandria, LA M.S.A.  936.4  4.5 35.5 120.1  776.3  4,565.90 1,167.00  3,083.70  315.2 LA  Alexandria
6  7     Altoona, PA M.S.A.  216.5  0.8 28.7  25.5  161.6  1,430.20    218.9  1,163.60   47.8 PA     Altoona
```

< Question 7 – Crime data>

Referring to the problems that you pointed out from above, fix the problem. Updated Crime data set as shown below.

Expected Result

```
> head(Crime)
                  MSA ViolentCrime Murder Rape Robbery AggravatedAssault PropertyCrime Burglary    Theft
1     Abilene, TX M.S.A.        412.5    5.3 56.0    78.4            272.8      3,609.00      852 2,493.60
2       Akron, OH M.S.A.        238.4    5.1 38.2    75.2            119.8      2,552.40    575.3 1,853.00
3      Albany, GA M.S.A.        667.9    7.8 30.4   157.9            471.8      3,894.10 1,099.60 2,652.80
4      Albany, OR M.S.A.        114.3    2.5 28.2    20.7             63.0      3,208.40    484.6 2,476.10
5 Albuquerque, NM M.S.A.        792.6    6.1 63.8   206.7            516.0      4,607.80    883.4 3,047.60
6  Alexandria, LA M.S.A.        936.4    4.5 35.5   120.1            776.3      4,565.90 1,167.00 3,083.70
  MotorVehicleTheft State        City
1             263.4    TX     Abilene
2             124.1    OH      Akron
3             141.7    GA      Albany
4             247.7    OR      Albany
5             676.9    NM Albuquerque
6             315.2    LA  Alexandria
```

< Question 8 – Crime data>

What are the Top 6 regions in Murder? Write down a R code that rearranges data set by the “Murder” with descending order and shows the first 6 rows of it.

Expected Result

	MSA	ViolentCrime	Murder	Rape	Robbery	AggravatedAssault	PropertyCrime
95	Detroit-Dearborn-Livonia, MI M.D.	901.5	19.3	57.2	247.0	577.9	2,935.80
238	New Orleans-Metairie, LA M.S.A.	534.4	16.3	47.7	168.0	302.5	3,030.20
226	Monroe, LA M.S.A.2	1,160.00	15.1	41.3	122.2	981.3	4,701.90
306	Savannah, GA M.S.A.	400.5	15.0	22.4	157.8	205.2	3,313.80
261	Philadelphia, PA M.D.	861.1	14.8	66.5	351.9	427.9	2,827.00
233	Myrtle Beach-Conway-North Myrtle Beach, SC-NC M.S.A.	390.2	14.2	62.7	75.3	237.9	3,794.80

	Burglary	Theft	MotorVehicleTheft	State	City
95	700.9	1,769.30	465.7	MI	Detroit
238	532.2	2,196.40	301.6	LA	New Orleans
226	1,179.50	3,356.00	166.3	LA	Monroe
306	684.2	2,283.90	345.7	GA	Savannah
261	450.5	2,113.20	263.4	PA	Philadelphia
233	773.4	2,745.70	275.7	SC	Myrtle Beach

< Question 9 – Crime data>

(1) Set random number generator value as 2022 and create two data sets by randomly selecting 100 samples (Crime1 & Crime2). (2) Compare average value of "ViolentCrime" of two sets and that of Crime set. Are they similar? Explain the result with your own language. Ignore NA if there's any.

< Question 10 – Crime data>

(1) Create a data subset called "PropertyCrime" which only includes regions with the 0.75 quantile in PropertyCrime. As shown below, show first 6 rows by rearranging them with PropertyCrime (descending order). (2) From this, which state(s) is (are) most frequently appeared ones? Write down the name of states, number of cases, and relevant codes to produce the result shown below. Ignore NA if there's any.

Expected Result

(1)

	MSA	ViolentCrime	Murder	Rape	Robbery	AggravatedAssault
272	Pueblo, CO M.S.A.	615.7	8.0	104.9	114.7	388.2
140	Hammond, LA M.S.A.	769.4	3.9	43.7	141.9	580.0
339	Tucson, AZ M.S.A.	421.4	4.5	52.5	123.7	240.8
318	Spokane-Spokane Valley, WA M.S.A.	304.9	3.8	44.9	85.5	170.7
225	Monroe, LA M.S.A.2	1,160.00	15.1	41.3	122.2	981.3
4	Albuquerque, NM M.S.A.	792.6	6.1	63.8	206.7	516.0

	PropertyCrime	Burglary	Theft	MotorVehicleTheft	State	City
272	5190.6	1,392.10	3,174.80	623.7	CO	Pueblo
140	4770.9	1,374.30	3,086.20	310.3	LA	Hammond
339	4748.5	555.5	3,936.40	256.6	AZ	Tucson
318	4745.0	922.7	3,350.00	472.3	WA	Spokane
225	4701.9	1,179.50	3,356.00	166.3	LA	Monroe
4	4607.8	883.4	3,047.60	676.9	NM	Albuquerque

(2)

CA	TX	GA	LA	WA	AL	AR	FL	MO	OH	SC	TN	AK	KY	MT	NC	NJ	OR	AZ	CO	HI
11	11	7	7	7	5	4	4	3	3	3	3	2	2	2	2	2	2	1	1	1

IL	IN	KS	MI	MS	NM	OK	WV
1	1	1	1	1	1	1	1

< Question 11 – Crime data>

Which state is the worst in the overall number of violent crime? Write down a R code to answer this question and explanation.