AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

Assignment Title:DATA SCIENCE MID REPORT		
Assignment No:01		Date of Submission:01/03/2023
Course Title:DATA SCIENCE		
Course Code:4180		Section:B
Semester:10th Semester	2023-2024	Course Teacher: AKINUL ISLAM JONY

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	**FAIRUZ ZAHIN SNEHA**	**20-41912-1**	BSC in CSE	Sneha
2				
3				
4				
5				
6				
7				
8				
9				
10				

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	

Content

1.1 Project overview

The main purpose of doing this project is to get a proper idea about Data preprocessing. This project mainly focuses on the data pre-processing which is a very important part of Data science. Moreover, different types of pre-processing techniques are used here which also serve a purpose to get agood grip of them. so, one of its purposes can be defined by it. The data set which is given is basically contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. Also given is the percentage of the population living in urban areas. The dataset consists of the following attributes:

- Name of the States of US which sample obtained
- Arrests for Murder is measured as per 100,000 people.
- Number of Arrests for Assault are measured per 100,000 people also.
- The last column is for the Urban Population percentage those who live in those corresponded states. To build an effective analysis, first we need to prepare the dataset then we have to perform Data Pre-processing for making a cleaned dataset.

1.2 Project Solution Design

For getting the cleaned dataset by doing data analysis, I must perform the data pre-processing. Analyzing a raw dataset to detect and remove duplicates, mistakes, and extraneous data is known as data cleaning. The table contained some missing data, which I filled in using median after replacing with N/A. Then I made an effort to manage every piece of noisy data that was present in the dataset. Following that, I was unable to locate any data munging. The data set was relatively clean when the data cleaning procedure was completed. I then attempt to integrate the data by adding a new column with the name Type. So that this dataset is more consistent. For Data Pre-processing I will be using A Hands-On method to design our project. There are generally 5 steps for completing the data process. Here is how I am going to do it (data pre-processing):

1. Data cleaning:
    a. Smooth Noisy Data
    b. Handling Missing Data
    c. Data Wrangling or Munging
2. Data Integration
3. Data Transformation
4. Data Reduction
5. Data Discretization

1.3 Data Frame

A data frame is more general that contains different modes of data. As we can see dataset contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973.This dataset consists of numeric and character data. Because there are multiple modes of data. In this case, a data frame is the structure of choice.
Here is the Code to make dataset in R language:

```
dataset<-
data.frame(Name_of_states=c("Alabama","Alaska","Arizona","Arkansas","Califomia","Colorado","Connecticut","Delaware","Florida","Georgia","Hawaii","Idaho","Illinois","Indiana","Iowa","Kansas","Kentucky","Louisiana","Maine","Maryland","Massachusetts","Michigan","Minnesota","Mississippi","Missouri","Montana","Nebraska","Nevada","New-Hampshire","New-Jersey","New-Mexico","NewYork","North-Carolina","North-Dakota","Ohio","Oklahoma","Oregon","Pennsylvania","Rhode-Island","South-Carolina","South-
Dakota","Tennessee","Texas","Utah","Vermont","Virginia","Washington","WestVirginia","Wisconsin","Wyoming"),
Murder=c(13.2,10,8.1,8.8,9,7.9,3.3,5.9,15.4,17.4,5.3,2.6,10.4,7.2,2.2,6,9.7,15.4,2.1,11.3,4.4,12.1,2.7,16.1,9,6,4.3,12.2,2.1,7.4,11.4,11.1,13,0.8,7.3,6.6,4.9,6.3,3.4,14.4,3.8,13.2,12.7,3.2,2.2,8.5,4,5.7,2.6,6.8),
Assault=c(236,263,294,190,276,204,110,238,335,"NA",46,120,249,113,56,115,109,249,83,300,149,255,72,259,178,109,102,252,57,159,285,254,337,45,120,151,159,106,174,879,86,188,201,120,48,156,145,81,53,161),
Urban_population=c(58,48,80,50,91,78,77,72,80,60,83,54,83,65,570,66,52,66,51,67,85,74,66,44,70,53,62,81,56,89,70,6,45,44,75,68,67,72,87,48,45,59,80,80,32,63,73,39,66,60))
print(dataset)
```

Filter

	Name_of_states	Murder	Assault	Urban_population
1	Alabama	13.2	236	58
2	Alaska	10.0	263	48
3	Arizona	8.1	294	80
4	Arkansas	8.8	190	50
5	Califomia	9.0	276	91
6	Colorado	7.9	204	78
7	Connecticut	3.3	110	77
8	Delaware	5.9	238	72
9	Florida	15.4	335	80
10	Georgia	17.4	NA	60
11	Hawaii	5.3	46	83
12	Idaho	2.6	120	54
13	Illinois	10.4	249	83
14	Indiana	7.2	113	65
15	Iowa	2.2	56	570
16	Kansas	6.0	115	66
17	Kentucky	9.7	109	52
18	Louisiana	15.4	249	66
19	Maine	2.1	83	51
20	Maryland	11.3	300	67

Filter

	Name_of_states	Murder	Assault	Urban_population
31	New-Mexico	11.4	285	70
32	NewYork	11.1	254	6
33	North-Carolina	13.0	337	45
34	North-Dakota	0.8	45	44
35	Ohio	7.3	120	75
36	Oklahoma	6.6	151	68
37	Oregon	4.9	159	67
38	Pennsylvania	6.3	106	72
39	Rhode-Island	3.4	174	87
40	South-Carolina	14.4	879	48
41	South-Dakota	3.8	86	45
42	Tennessee	13.2	188	59
43	Texas	12.7	201	80
44	Utah	3.2	120	80
45	Vermont	2.2	48	32
46	Virginia	8.5	156	63
47	Washington	4.0	145	73
48	WestVirginia	5.7	81	39
49	Wisconsin	2.6	53	66
50	Wyoming	6.8	161	60

1.4 Data Pre-processing

Data in the real world is often dirty. That is, it is in need of being cleaned up before it can be used for a desired purpose. This is often called data pre- processing. There are mainly five steps for data pre-processing. These are as follows-

1. Data cleaning:
   a. Smooth Noisy Data
   b. Handling Missing Data
   c. Data Wrangling or Munging
2. Data Integration
3. Data Transformation
4. Data Reduction
5. Data Discretization

1.Data cleaning: I will look here at three key methods (such as Data Munging, Handling Missing Data, and Smooth Noisy Data) that describe ways in which data may be cleaned or better organized, or scrubbed of potentially incorrect, incomplete, or duplicated information.

a. Smooth Noisy Data : There are times when the data is not missing, but it is corrupted for some reason. Data corruption may be a result of faulty data collection instruments, data entry problems, or technology limitations.
We can see that the following dataset contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973.Following this table there is no negative data or corrupted data here . Since the number of people who have died cannot be decimal, the data in the Murder column is noisy. So, I can round up this murder column dataset.
Here is the code:
dataset$Murder<-round(dataset$Murder)

	Name_of_states	Murder	Assault	Urban_population
1	Alabama	13	236	58
2	Alaska	10	263	48
3	Arizona	8	294	80
4	Arkansas	9	190	50
5	Califomia	9	276	91
6	Colorado	8	204	78
7	Connecticut	3	110	77
8	Delaware	6	238	72
9	Florida	15	335	80
10	Georgia	17	NA	60
11	Hawaii	5	46	83
12	Idaho	3	120	54
13	Illinois	10	249	83
14	Indiana	7	113	65
15	Iowa	2	56	570
16	Kansas	6	115	66
17	Kentucky	10	109	52
18	Louisiana	15	249	66

	Name_of_states	Murder	Assault	Urban_population
33	North-Carolina	13	337	45
34	North-Dakota	1	45	44
35	Ohio	7	120	75
36	Oklahoma	7	151	68
37	Oregon	5	159	67
38	Pennsylvania	6	106	72
39	Rhode-Island	3	174	87
40	South-Carolina	14	879	48
41	South-Dakota	4	86	45
42	Tennessee	13	188	59
43	Texas	13	201	80
44	Utah	3	120	80
45	Vermont	2	48	32
46	Virginia	8	156	63
47	Washington	4	145	73
48	WestVirginia	6	81	39
49	Wisconsin	3	53	66
50	Wyoming	7	161	60

b. Handling Missing Data : Sometimes data may be in the right format, but some of the values are missing. As we can see in the dataset, we have missing values (represented by NA – available) of the number of assault for Georgia. A simple workaround for this is to replace all the NAs with some common values, such as zero or average of all the values for that attribute. Here, I am going to use the average of the attribute for handling the missing values. So, for Georgia, I will use the value of 182.

Here is the code:

```
dataset[is.na(dataset)] <-mean(dataset$Assault,na.rm=TRUE)
dataset$Assault<-round(dataset$Assault)
```

	Name_of_states	Murder	Assault	Urban_population
1	Alabama	13	236	58
2	Alaska	10	263	48
3	Arizona	8	294	80
4	Arkansas	9	190	50
5	Califomia	9	276	91
6	Colorado	8	204	78
7	Connecticut	3	110	77
8	Delaware	6	238	72
9	Florida	15	335	80
10	Georgia	17	182	60
11	Hawaii	5	46	83
12	Idaho	3	120	54
13	Illinois	10	249	83
14	Indiana	7	113	65
15	Iowa	2	56	570
16	Kansas	6	115	66
17	Kentucky	10	109	52
18	Louisiana	15	249	66
19	Maine	2	83	51
20	Maryland	11	300	67

c. Data Wrangling or Munging: Often, the data is not in a format that is easy to work with. For example, it may be stored or presented in a way that is hard to process. As previously discussed, data wrangling is the process of manually converting or mapping data from one raw form into another format. But, fortunately for us, this dataset does not involve any data wrangling steps.

So, at the end of Data Cleaning stage the dataset would look like-

	Name_of_states	Murder	Assault	Urban_population
1	Alabama	13	236	58
2	Alaska	10	263	48
3	Arizona	8	294	80
4	Arkansas	9	190	50
5	Califomia	9	276	91
6	Colorado	8	204	78
7	Connecticut	3	110	77
8	Delaware	6	238	72
9	Florida	15	335	80
10	Georgia	17	182	60
11	Hawaii	5	46	83
12	Idaho	3	120	54
13	Illinois	10	249	83
14	Indiana	7	113	65
15	Iowa	2	56	570
16	Kansas	6	115	66
17	Kentucky	10	109	52
18	Louisiana	15	249	66
19	Maine	2	83	51
20	Maryland	11	300	67

	Name_of_states	Murder	Assault	Urban_population
31	New-Mexico	11	285	70
32	NewYork	11	254	6
33	North-Carolina	13	337	45
34	North-Dakota	1	45	44
35	Ohio	7	120	75
36	Oklahoma	7	151	68
37	Oregon	5	159	67
38	Pennsylvania	6	106	72
39	Rhode-Island	3	174	87
40	South-Carolina	14	879	48
41	South-Dakota	4	86	45
42	Tennessee	13	188	59
43	Texas	13	201	80
44	Utah	3	120	80
45	Vermont	2	48	32
46	Virginia	8	156	63
47	Washington	4	145	73
48	WestVirginia	6	81	39
49	Wisconsin	3	53	66
50	Wyoming	7	161	60

2. Data Integration: To be as efficient and effective for various data analyses as possible, data from various sources commonly needs to be integrated. As we can see the following table, there is no extra dataset or table which can be integrated.
So, at the end of Data Integration stage the dataset would look same as the data cleaning.

3. Data Transformation: Data must be transformed so it is consistent and readable.The data transformation process involves one or more of smoothing, removing noise from data, summarization, generalization, and normalization. For this example, I will apply smoothing, which is simpler than summarization and normalization.
As we can see, in our data, the percentage of urban population of New-York is 6 which is very low comparing with others. So, I will replace the value or percentage with 74 in place of 6(by doing averaging).
Here is the code-

```
avg<-mean(dataset$Urban_population)
dataset$Urban_population[32]<-avg
dataset$Urban_population<-round(dataset$Urban_population)
```

So, at the end of Data Transformation stage the dataset would look like-

	Name_of_states	Murder	Assault	Urban_population
25	Missouri	9	178	70
26	Montana	6	109	53
27	Nebraska	4	102	62
28	Nevada	12	252	81
29	New-Hampshire	2	57	56
30	New-Jersey	7	159	89
31	New-Mexico	11	285	70
32	NewYork	11	254	74
33	North-Carolina	13	337	45
34	North-Dakota	1	45	44
35	Ohio	7	120	75
36	Oklahoma	7	151	68
37	Oregon	5	159	67
38	Pennsylvania	6	106	72
39	Rhode-Island	3	174	87
40	South-Carolina	14	879	48
41	South-Dakota	4	86	45
42	Tennessee	13	188	59

4. Data Reduction: Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained. As. we can see the urban population of Iowa is 570% which is too much high comparing with other US states. So, I will reduce the percentage and now which is 57%.
Here is the code-

```
reduced_data<-570
dataset[dataset == reduced_data] <- 57
dataset
```

So, at the end of Data Reduction stage the dataset would look like-

	Name_of_states	Murder	Assault	Urban_population
1	Alabama	13	236	58
2	Alaska	10	263	48
3	Arizona	8	294	80
4	Arkansas	9	190	50
5	Califomia	9	276	91
6	Colorado	8	204	78
7	Connecticut	3	110	77
8	Delaware	6	238	72
9	Florida	15	335	80
10	Georgia	17	182	60
11	Hawaii	5	46	83
12	Idaho	3	120	54
13	Illinois	10	249	83
14	Indiana	7	113	65
15	Iowa	2	56	57
16	Kansas	6	115	66
17	Kentucky	10	109	52
18	Louisiana	15	249	66

5. Data Discretization: We are often dealing with data that are collected from processes that are continuous, such as temperature, ambient light, and a company's stock price. But sometimes we need to convert these continuous values into more manageable parts. This mapping is called discretization. Here, I will discretize the number of murders of the US states in four categories. These are as follows-

• less than or equal to 4 (represented by 0),
• more than 4 but less than or equal to 10 (represented by 1),
• more than 10 but less than or equal to 15 (represented by 2), and
• more than 15 (represented by 3)

Here is the Code to make dataset in R language:

```
dataset$Murder<- as.factor( ifelse(dataset$Murder<=4, 0,
                                   ifelse(dataset$ Murder<=10, 1,
                                          ifelse(dataset$ Murder<=15, 2,
                                                 ifelse(dataset$ Murder >15, 3,'none')

                                          )


                                   )
)
)
```

So, at the end of Data Reduction stage the dataset would look like-

	Name_of_states	Murder	Assault	Urban_population
1	Alabama	2	236	58
2	Alaska	1	263	48
3	Arizona	1	294	80
4	Arkansas	1	190	50
5	Califomia	1	276	91
6	Colorado	1	204	78
7	Connecticut	0	110	77
8	Delaware	1	238	72
9	Florida	2	335	80
10	Georgia	3	182	60
11	Hawaii	1	46	83
12	Idaho	0	120	54
13	Illinois	1	249	83
14	Indiana	1	113	65
15	Iowa	0	56	57
16	Kansas	1	115	66
17	Kentucky	1	109	52
18	Louisiana	2	249	66
19	Maine	0	83	51
20	Maryland	2	300	67
21	Massachusetts	0	149	85

	Name_of_states	Murder	Assault	Urban_population
30	New-Jersey	1	159	89
31	New-Mexico	2	285	70
32	NewYork	2	254	74
33	North-Carolina	2	337	45
34	North-Dakota	0	45	44
35	Ohio	1	120	75
36	Oklahoma	1	151	68
37	Oregon	1	159	67
38	Pennsylvania	1	106	72
39	Rhode-Island	0	174	87
40	South-Carolina	2	879	48
41	South-Dakota	0	86	45
42	Tennessee	2	188	59
43	Texas	2	201	80
44	Utah	0	120	80
45	Vermont	0	48	32
46	Virginia	1	156	63
47	Washington	0	145	73
48	WestVirginia	1	81	39
49	Wisconsin	0	53	66
50	Wyoming	1	161	60

1.5 New variable integration

After completing the data pre-processing, I will add a new column (named population level) in the data frame based on the urban population variable. I convert the urban population percentage into level, for example, small (<50%),medium(>=50% to <60%),large(>=60% to <70%),extra-large(>=70%).As the newly added variable (population level) are numerical, so I also convert the population level variable into an ordered factor variable (named ordered factor population) like small = 1, medium = 2, large = 3, and extra-large = 4, add it to the data frame. So, in total the data frame will contain 6 variables.
Here is the code-

```
dataset$population_level <- as.factor( ifelse(dataset$Urban_population<50, 'small',
                                  ifelse(dataset$Urban_population<60, 'medium',
                                      ifelse(dataset$Urban_population<70, 'large',
                                          ifelse(dataset$Urban_population<Inf, 'extra large','none')
                                      )
                                  )
)
)
dataset$ordered_factor_population <- as.factor( ifelse(dataset$population_level=='small', 1,
                                      ifelse(dataset$population_level=='medium', 2,
                                          ifelse(dataset$population_level=='large',3,
ifelse(dataset$population_level=='extra large', 4,'none')
                                      )
                                  )
)
)
```

So, at the end of New variable integration stage the dataset would look like-

	Name_of_states	Murder	Assault	Urban_population	population_level	ordered_factor_population
1	Alabama	2	236	58	medium	2
2	Alaska	1	263	48	small	1
3	Arizona	1	294	80	extra large	4
4	Arkansas	1	190	50	medium	2
5	Califomia	1	276	91	extra large	4
6	Colorado	1	204	78	extra large	4
7	Connecticut	0	110	77	extra large	4
8	Delaware	1	238	72	extra large	4
9	Florida	2	335	80	extra large	4
10	Georgia	3	182	60	large	3
11	Hawaii	1	46	83	extra large	4
12	Idaho	0	120	54	medium	2
13	Illinois	1	249	83	extra large	4
14	Indiana	1	113	65	large	3
15	Iowa	0	56	57	medium	2
16	Kansas	1	115	66	large	3
17	Kentucky	1	109	52	medium	2
18	Louisiana	2	249	66	large	3
19	Maine	0	83	51	medium	2
20	Maryland	2	300	67	large	3
21	Massachusetts	0	149	85	extra large	4

	Name_of_states	Murder	Assault	Urban_population	population_level	ordered_factor_population
30	New-Jersey	1	159	89	extra large	4
31	New-Mexico	2	285	70	extra large	4
32	NewYork	2	254	74	extra large	4
33	North-Carolina	2	337	45	small	1
34	North-Dakota	0	45	44	small	1
35	Ohio	1	120	75	extra large	4
36	Oklahoma	1	151	68	large	3
37	Oregon	1	159	67	large	3
38	Pennsylvania	1	106	72	extra large	4
39	Rhode-Island	0	174	87	extra large	4
40	South-Carolina	2	879	48	small	1
41	South-Dakota	0	86	45	small	1
42	Tennessee	2	188	59	medium	2
43	Texas	2	201	80	extra large	4
44	Utah	0	120	80	extra large	4
45	Vermont	0	48	32	small	1
46	Virginia	1	156	63	large	3
47	Washington	0	145	73	extra large	4
48	WestVirginia	1	81	39	small	1
49	Wisconsin	0	53	66	large	3
50	Wyoming	1	161	60	large	3

1.6 The Cleaned dataset

This is our cleaned dataset at the end of data pre-processing. Now, I have a cleaned and nice-looking dataset which is ready for processing. It could appear that I didn't actually process or analyze the data. However, using our pre-processing methods, I was able to create a far more useful and superior dataset. However, the practical exercise I conducted here has provided some insights into what must take place before one can obtain attractive data for processing. so finally the cleaned dataset is as follows-

	Name_of_states	Murder	Assault	Urban_population	population_level	ordered_factor_population
1	Alabama	2	236	58	medium	2
2	Alaska	1	263	48	small	1
3	Arizona	1	294	80	extra large	4
4	Arkansas	1	190	50	medium	2
5	California	1	276	91	extra large	4
6	Colorado	1	204	78	extra large	4
7	Connecticut	0	110	77	extra large	4
8	Delaware	1	238	72	extra large	4
9	Florida	2	335	80	extra large	4
10	Georgia	3	182	60	large	3
11	Hawaii	1	46	83	extra large	4
12	Idaho	0	120	54	medium	2
13	Illinois	1	249	83	extra large	4
14	Indiana	1	113	65	large	3
15	Iowa	0	56	57	medium	2
16	Kansas	1	115	66	large	3
17	Kentucky	1	109	52	medium	2
18	Louisiana	2	249	66	large	3
19	Maine	0	83	51	medium	2
20	Maryland	2	300	67	large	3
21	Massachusetts	0	149	85	extra large	4

	Name_of_states	Murder	Assault	Urban_population	population_level	ordered_factor_population
21	Massachusetts	0	149	85	extra large	4
22	Michigan	2	255	74	extra large	4
23	Minnesota	0	72	66	large	3
24	Mississippi	3	259	44	small	1
25	Missouri	1	178	70	extra large	4
26	Montana	1	109	53	medium	2
27	Nebraska	0	102	62	large	3
28	Nevada	2	252	81	extra large	4
29	New-Hampshire	0	57	56	medium	2
30	New-Jersey	1	159	89	extra large	4
31	New-Mexico	2	285	70	extra large	4
32	NewYork	2	254	74	extra large	4
33	North-Carolina	2	337	45	small	1
34	North-Dakota	0	45	44	small	1
35	Ohio	1	120	75	extra large	4
36	Oklahoma	1	151	68	large	3
37	Oregon	1	159	67	large	3
38	Pennsylvania	1	106	72	extra large	4
39	Rhode-Island	0	174	87	extra large	4
40	South-Carolina	2	879	48	small	1
41	South-Dakota	0	86	45	small	1
42	Tennessee	2	188	59	medium	2

	Name_of_states	Murder	Assault	Urban_population	population_level	ordered_factor_population
30	New-Jersey	1	159	89	extra large	4
31	New-Mexico	2	285	70	extra large	4
32	NewYork	2	254	74	extra large	4
33	North-Carolina	2	337	45	small	1
34	North-Dakota	0	45	44	small	1
35	Ohio	1	120	75	extra large	4
36	Oklahoma	1	151	68	large	3
37	Oregon	1	159	67	large	3
38	Pennsylvania	1	106	72	extra large	4
39	Rhode-Island	0	174	87	extra large	4
40	South-Carolina	2	879	48	small	1
41	South-Dakota	0	86	45	small	1
42	Tennessee	2	188	59	medium	2
43	Texas	2	201	80	extra large	4
44	Utah	0	120	80	extra large	4
45	Vermont	0	48	32	small	1
46	Virginia	1	156	63	large	3
47	Washington	0	145	73	extra large	4
48	WestVirginia	1	81	39	small	1
49	Wisconsin	0	53	66	large	3
50	Wyoming	1	161	60	large	3