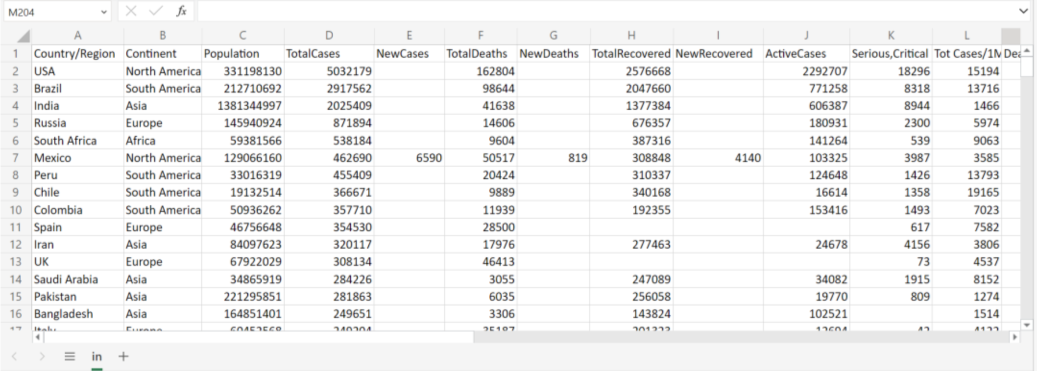
Data Science Project- Dashboard

Title of Dashboard- “The Covid-19"

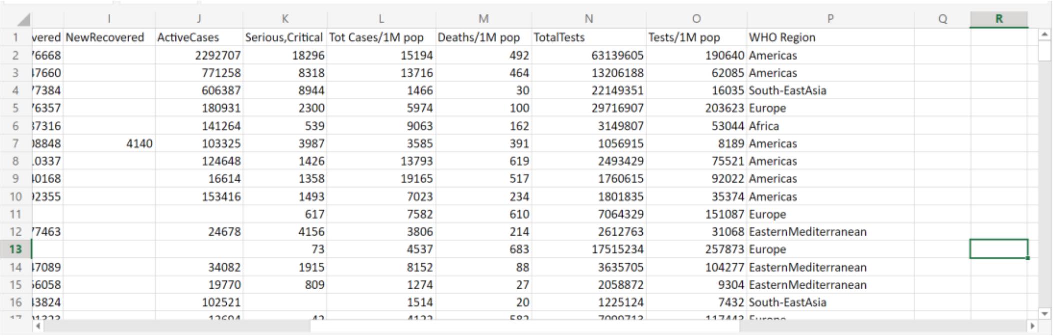
Submitted by Jasmine Kaur

Roll No- 102183047, Sub-group- 3co25

Original dataset before preprocessing:



	A	B	C	D	E	F	G	H	I	J	K	L	
1	Country/Region	Continent	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M	Dea
2	USA	North America	331198130	5032179		162804		2576668		2292707	18296	15194	
3	Brazil	South America	212710692	2917562		98644		2047660		771258	8318	13716	
4	India	Asia	1381344997	2025409		41638		1377384		606387	8944	1466	
5	Russia	Europe	145940924	871894		14606		676357		180931	2300	5974	
6	South Africa	Africa	59381566	538184		9604		387316		141264	539	9063	
7	Mexico	North America	129066160	462690	6590	50517	819	308848	4140	103325	3987	3585	
8	Peru	South America	33016319	455409		20424		310337		124648	1426	13793	
9	Chile	South America	19132514	366671		9889		340168		16614	1358	19165	
10	Colombia	South America	50936262	357710		11939		192355		153416	1493	7023	
11	Spain	Europe	46756648	354530		28500					617	7582	
12	Iran	Asia	84097623	320117		17976		277463		24678	4156	3806	
13	UK	Europe	67922029	308134		46413					73	4537	
14	Saudi Arabia	Asia	34865919	284226		3055		247089		34082	1915	8152	
15	Pakistan	Asia	221295851	281863		6035		256058		19770	809	1274	
16	Bangladesh	Asia	164851401	249651		3306		143824		102521		1514	



		I	J	K	L	M	N	O	P	Q	R
1	vered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	TotalTests	Tests/1M pop	WHO Region		
2	6668		2292707	18296	15194	492	63139605	190640	Americas		
3	7660		771258	8318	13716	464	13206188	62085	Americas		
4	7384		606387	8944	1466	30	22149351	16035	South-EastAsia		
5	6357		180931	2300	5974	100	29716907	203623	Europe		
6	7316		141264	539	9063	162	3149807	53044	Africa		
7	8848	4140	103325	3987	3585	391	1056915	8189	Americas		
8	0337		124648	1426	13793	619	2493429	75521	Americas		
9	0168		16614	1358	19165	517	1760615	92022	Americas		
10	2355		153416	1493	7023	234	1801835	35374	Americas		
11				617	7582	610	7064329	151087	Europe		
12	7463		24678	4156	3806	214	2612763	31068	EasternMediterranean		
13				73	4537	683	17515234	257873	Europe		
14	7089		34082	1915	8152	88	3635705	104277	EasternMediterranean		
15	6058		19770	809	1274	27	2058872	9304	EasternMediterranean		
16	3824		102521		1514	20	1225124	7432	South-EastAsia		

R code for Data PreProcessing:

```r
dataset<-read.csv(choose.files())
library(dplyr)
library(WriteXLS)
library('writexl')
install.packages('data.table')

install.packages('WriteXLS')
head(dataset)
ncol(dataset)
nrow(dataset)
colnames(dataset)
data.frame(dataset)
# we are beginning with data preprocessing


# 1) checking null values in the dataset

sum(is.na(dataset))
sum(is.na(dataset$NewCases))
sum(is.na(dataset$NewDeaths))
sum(is.na(dataset$NewRecovered))
#df2<-data.frame(dataset[, c("Country.Region","NewCases", "NewDeaths","NewRecovered")])
# these are features with high no of na values therefore we will split them into different dataset
```

sum(is.na(dataset$Population)) # since there is only one row with na value we can drop the row

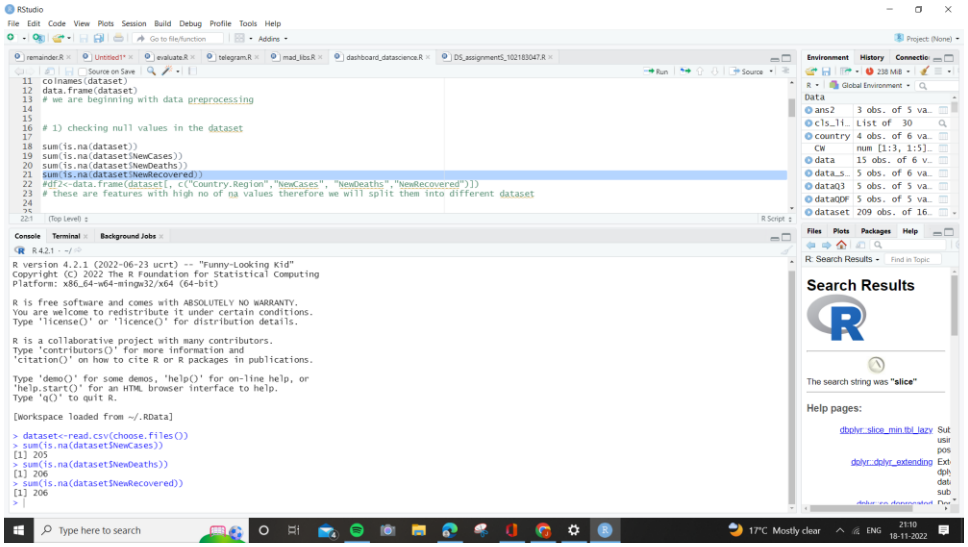
dataset[-(is.na(dataset$Population)==TRUE),]

sum(is.na(dataset$TotalRecovered)) #replacing the nas with mean value

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

sum(is.na(dataset$Deaths.1M.pop)) #replacing the nas with median value

dataset$Deaths.1M.pop[is.na(dataset$Deaths.1M.pop)] <- mean(dataset$Deaths.1M.pop, na.rm =TRUE)



```
colnames(dataset)
data.frame(dataset)
# we are beginning with data preprocessing


# 1) checking null values in the dataset

sum(is.na(dataset))
sum(is.na(dataset$NewCases))
sum(is.na(dataset$NewDeaths))
sum(is.na(dataset$NewRecovered))
#df2<-data.frame(dataset[, c("Country.Region","NewCases", "NewDeaths","NewRecovered")])
# these are features with high no of na values therefore we will split them into different dataset
```

```
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> dataset<-read.csv(choose.files())
> sum(is.na(dataset$NewCases))
[1] 205
> sum(is.na(dataset$NewDeaths))
[1] 206
> sum(is.na(dataset$NewRecovered))
[1] 206
> 
```

Data	
ans2	3 obs. of 5 va…
cls_li…	List of 30
country	4 obs. of 6 va…
CW	num [1:3, 1:5]…
data	15 obs. of 6 v…
data_s…	5 obs. of 6 va…
dataQ3	5 obs. of 5 va…
dataQDF	5 obs. of 5 va…
dataset	209 obs. of 16…

Search Results

The search string was "slice"

Help pages:

dbplyr::slice_min.tbl_lazy

dplyr::dplyr_extending

# 2) checking anomlous (inf) values in dataset

sum(is.infinite(dataset$Deaths.1M.pop))

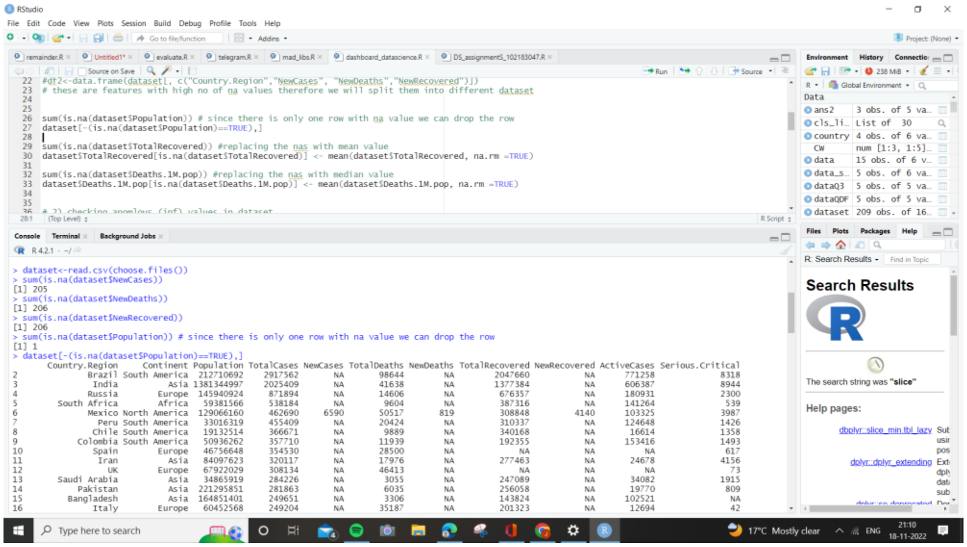
#replacing inf values with mean

dataset$Deaths.1M.pop[which(!is.finite(dataset$Deaths.1M.pop))]<-variable

#mean(dataset$Deaths...100.Recovered)

variable=mean(dataset$Deaths...100.Recovered) #saved 39.47385

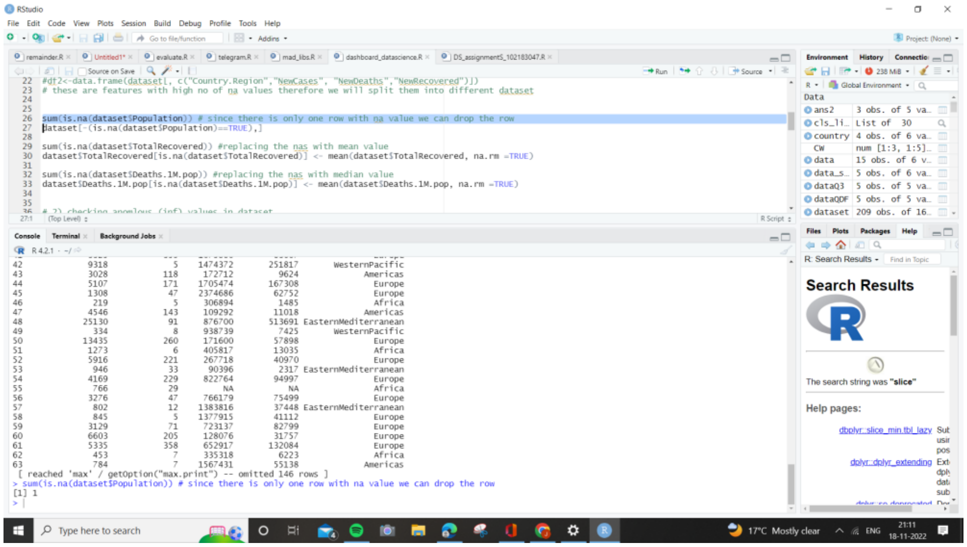
#variable=39.47385



```
#df2<-data.frame(dataset[, c("Country.Region","NewCases", "NewDeaths","NewRecovered")])
# these are features with high no of na values therefore we will split them into different dataset


sum(is.na(dataset$Population)) # since there is only one row with na value we can drop the row
dataset[-(is.na(dataset$Population)==TRUE),]

sum(is.na(dataset$TotalRecovered)) #replacing the nas with mean value
dataset$TotalRecovered[is.na(dataset$TotalRecovered)] <- mean(dataset$TotalRecovered, na.rm =TRUE)

sum(is.na(dataset$Deaths.1M.pop)) #replacing the nas with median value
dataset$Deaths.1M.pop[is.na(dataset$Deaths.1M.pop)] <- mean(dataset$Deaths.1M.pop, na.rm =TRUE)


# 2) checking anomlous (inf) values in dataset
```

```
Data
ans2       3 obs. of 5 va...
cls_li...  List of 30
country    4 obs. of 6 va...
 CW        num [1:3, 1:5]...
data       15 obs. of 6 v...
data_s...  5 obs. of 6 va...
dataQ3     5 obs. of 5 va...
dataQDF    5 obs. of 5 va...
dataset    209 obs. of 16...
```

```
> dataset<-read.csv(choose.files())
> sum(is.na(dataset$NewCases))
[1] 205
> sum(is.na(dataset$NewDeaths))
[1] 206
> sum(is.na(dataset$NewRecovered))
[1] 206
> sum(is.na(dataset$Population)) # since there is only one row with na value we can drop the row
[1] 1
> dataset[-(is.na(dataset$Population)==TRUE),]
     Country.Region     Continent Population TotalCases NewCases TotalDeaths NewDeaths TotalRecovered NewRecovered ActiveCases Serious.Critical
2            Brazil South America  212710692    2917562       NA       98644        NA        2047660           NA      771258             8318
3             India          Asia 1381344997    2025409       NA       41638        NA        1377384           NA      606387             8944
4            Russia        Europe  145940924     871894       NA       14606        NA         676357           NA      180931             2300
5      South Africa        Africa   59381566     538184       NA        9604        NA         387316           NA      141264              539
6            Mexico North America  129066160     462690     6590       50517       819         308848         4140      103325             3987
7              Peru South America   33016319     455409       NA       20424        NA         310337           NA      124648             1426
8             Chile South America   19132514     366671       NA        9889        NA         340168           NA       16614             1358
9          Colombia South America   50936262     357710       NA       11939        NA         192355           NA      153416             1493
10            Spain        Europe   46756648     354530       NA       28500        NA             NA           NA          NA              617
11             Iran          Asia   84097623     320117       NA       17976        NA         277463           NA       24678             4156
12               UK        Europe   67922029     308134       NA       46413        NA             NA           NA          NA               73
13     Saudi Arabia          Asia   34865919     284226       NA        3055        NA         247089           NA       34082             1915
14         Pakistan          Asia  221295851     281863       NA        6035        NA         256058           NA       19770              809
15       Bangladesh          Asia  164851401     249651       NA        3306        NA         143824           NA      102521               NA
16            Italy        Europe   60452568     249204       NA       35187        NA         201323           NA       12694               42
```



```
#df2<-data.frame(dataset[, c("Country.Region","NewCases", "NewDeaths","NewRecovered")])
# these are features with high no of na values therefore we will split them into different dataset


sum(is.na(dataset$Population)) # since there is only one row with na value we can drop the row
dataset[-(is.na(dataset$Population)==TRUE),]

sum(is.na(dataset$TotalRecovered)) #replacing the nas with mean value
dataset$TotalRecovered[is.na(dataset$TotalRecovered)] <- mean(dataset$TotalRecovered, na.rm =TRUE)

sum(is.na(dataset$Deaths.1M.pop)) #replacing the nas with median value
dataset$Deaths.1M.pop[is.na(dataset$Deaths.1M.pop)] <- mean(dataset$Deaths.1M.pop, na.rm =TRUE)


# 2) checking anomlous (inf) values in dataset
```

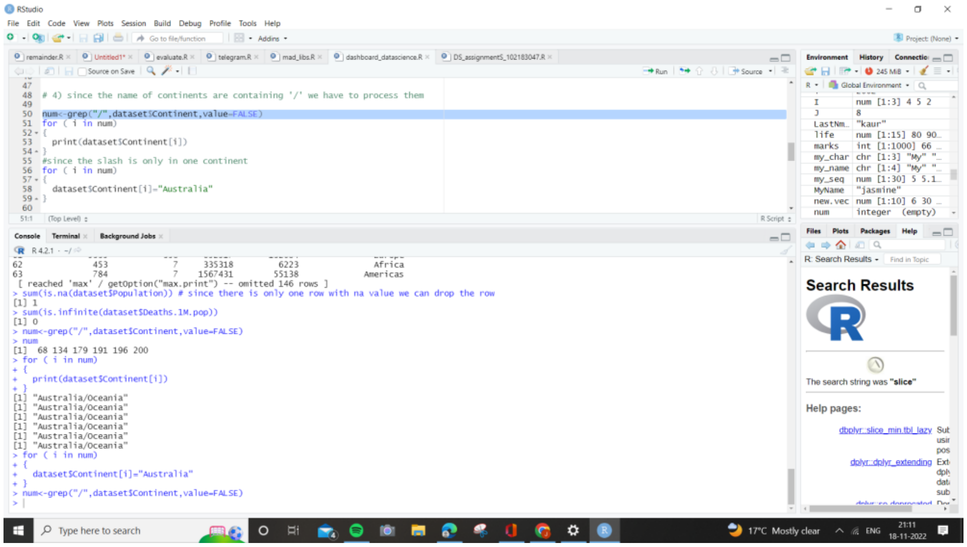
```
42            9318            5    1474372     251817        WesternPacific
43            3028          118     172712       9624              Americas
44            5107          171    1705474     167308                Europe
45            1308           47    2374686      62752                Europe
46             219            5     306894       1485                Africa
47            4546          143     109292      11018              Americas
48           25130           91     876700     513691 EasternMediterranean
49             334            8     938739       7425        WesternPacific
50           13435          260     171600      57898                Europe
51            1273            6     405817      13035                Africa
52            5916          221     267718      40970                Europe
53             946           33      90396       2317 EasternMediterranean
54            4169          229     822764      94997                Europe
55             766           29         NA         NA                Africa
56            3276           47     766179      75499                Europe
57             802           12    1383816      37448 EasternMediterranean
58             845            5    1377915      41112                Europe
59            3129           71     723137      82799                Europe
60            6603          205     128076      31757                Europe
61            5335          358     652917     132084                Europe
62             453            7     335318       6223                Africa
63             784            7    1567431      55138              Americas
 [ reached 'max' / getOption("max.print") -- omitted 146 rows ]
> sum(is.na(dataset$Population)) # since there is only one row with na value we can drop the row
[1] 1
> |
```

# 4) since the name of continents are containing '/' we have to process them

num<-grep("/",dataset$Continent,value=FALSE)

for (i in num)

```
{
  print(dataset$Continent[i])
}
#since the slash is only in one continent
for ( i in num)
{ dataset$Continent[i]="Australia"
}
```



```
# 4) since the name of continents are containing '/' we have to process them

num<-grep("/",dataset$Continent,value=FALSE)
for ( i in num)
{
  print(dataset$Continent[i])
}
#since the slash is only in one continent
for ( i in num)
{
  dataset$Continent[i]="Australia"
}

```

```
62           453         7     335318       6223          Africa
63           784         7    1567431      55138        Americas
 [ reached 'max' / getOption("max.print") -- omitted 146 rows ]
> sum(is.na(dataset$Population)) # since there is only one row with na value we can drop the row
[1] 1
> sum(is.infinite(dataset$Deaths.1M.pop))
[1] 0
> num<-grep("/",dataset$Continent,value=FALSE)
> num
[1]  68 134 179 191 196 200
> for ( i in num)
+ {
+   print(dataset$Continent[i])
+ }
[1] "Australia/Oceania"
[1] "Australia/Oceania"
[1] "Australia/Oceania"
[1] "Australia/Oceania"
[1] "Australia/Oceania"
[1] "Australia/Oceania"
> for ( i in num)
+ {
+   dataset$Continent[i]="Australia"
+ }
> num<-grep("/",dataset$Continent,value=FALSE)
> |
```

```
I         num [1:3] 4 5 2
J         8
LastNm..  "kaur"
life      num [1:15] 80 90…
marks     int [1:1000] 66 …
my_char   chr [1:3] "My" "…
my_name   chr [1:4] "My" "…
my_seq    num [1:30] 5 5.1…
MyName    "jasmine"
new.vec   num [1:10] 6 30 …
num       integer  (empty)
```

Search Results

The search string was **"slice"**

Help pages:

dbplyr::slice_min.tbl_lazy

dplyr::dplyr_extending

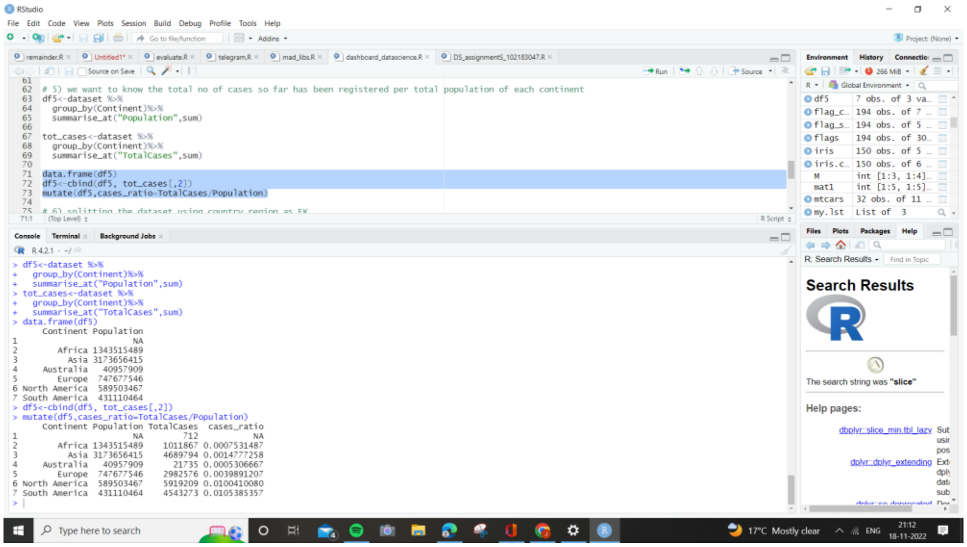
# 5) we want to know the total no of cases so far has been registered per total population of each continent

```
df5<-dataset %>%
  group_by(Continent)%>%
  summarise_at("Population",sum)


tot_cases<-dataset %>%
  group_by(Continent)%>%
  summarise_at("TotalCases",sum)
```

data.frame(df5)

df5<-cbind(df5, tot_cases[,2])

mutate(df5,cases_ratio=TotalCases/Population)



```
# 5) we want to know the total no of cases so far has been registered per total population of each continent
df5<-dataset %>%
  group_by(Continent)%>%
  summarise_at("Population",sum)

tot_cases<-dataset %>%
  group_by(Continent)%>%
  summarise_at("TotalCases",sum)

data.frame(df5)
df5<-cbind(df5, tot_cases[,2])
mutate(df5,cases_ratio=TotalCases/Population)

# 6) splitting the dataset using country.region as FK
```

```
> df5<-dataset %>%
+   group_by(Continent)%>%
+   summarise_at("Population",sum)
> tot_cases<-dataset %>%
+   group_by(Continent)%>%
+   summarise_at("TotalCases",sum)
> data.frame(df5)
      Continent Population
1                       NA
2        Africa 1343515489
3          Asia 3173656415
4     Australia   40957909
5        Europe  747677546
6 North America  589503467
7 South America  431110464
> df5<-cbind(df5, tot_cases[,2])
> mutate(df5,cases_ratio=TotalCases/Population)
      Continent Population TotalCases  cases_ratio
1                       NA        712           NA
2        Africa 1343515489    1011867 0.0007531487
3          Asia 3173656415    4689794 0.0014777258
4     Australia   40957909      21735 0.0005306667
5        Europe  747677546    2982576 0.0039891207
6 North America  589503467    5919209 0.0100410080
7 South America  431110464    4543273 0.0105385357
> |
```

df5	7 obs. of 3 va…
flag_c…	194 obs. of 7 …
flag_s…	194 obs. of 5 …
flags	194 obs. of 30…
iris	150 obs. of 5 …
iris.c…	150 obs. of 6 …
M	int [1:3, 1:4]…
mat1	int [1:5, 1:5]…
mtcars	32 obs. of 11 …
my.lst	List of 3

Search Results

The search string was **"slice"**

Help pages:

dbplyr::slice_min.tbl_lazy Sub usir pos

dplyr::dplyr_extending Ext dply dat sub

6) splitting the dataset using country.region as FK

df1<-data.frame(dataset[, c("Country.Region","TotalCases","TotalDeaths","TotalRecovered","ActiveCases")])

df1

df2<-data.frame(dataset[, c("Country.Region","NewCases", "NewDeaths","NewRecovered")])

df3<-data.frame(dataset[,c("Country.Region","Tot.Cases.1M.pop","Deaths.1M.pop","Tests.1M.pop")])

df4<-df5

#7) preprocessing of df2

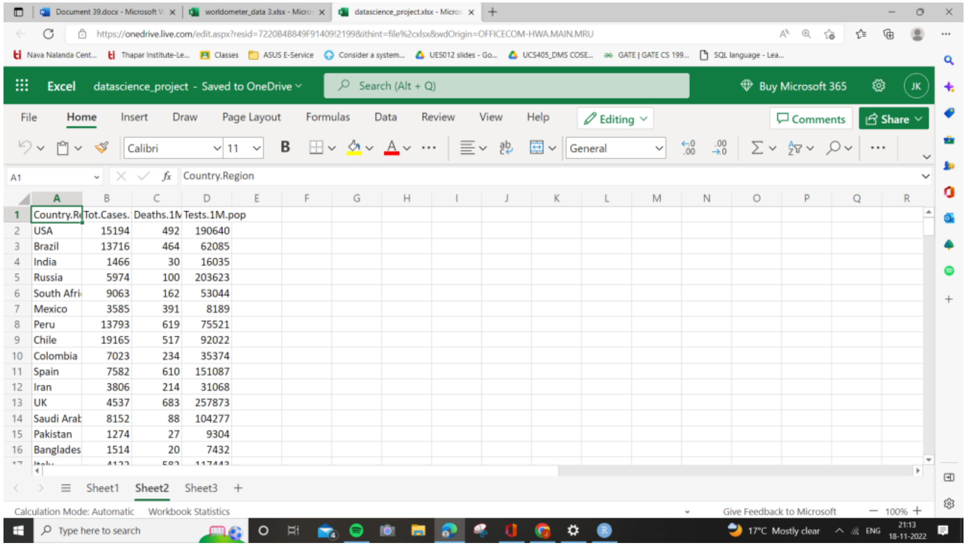
df2

sum(is.na(df2)) #dropping

8) now are four datasets are preprocessed for tableau representation, so we will convert them into xlsx file

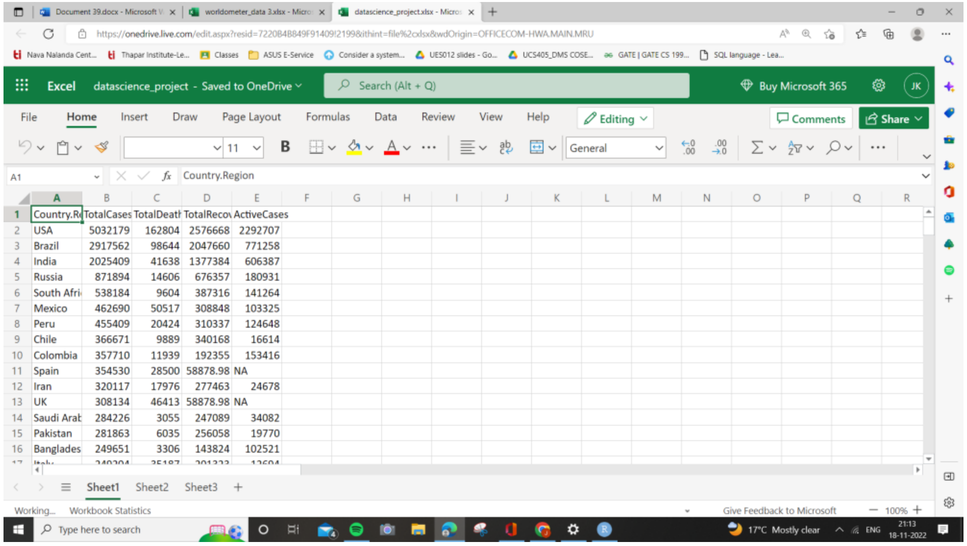
```
write.csv(df4, file="C:\\Users\\jaskirat singh\\Downloads\\datascienceproj\\dataset4.csv", row.names = FALSE)

write.csv(df1, file="C:\\Users\\jaskirat singh\\Downloads\\datascienceproj\\dataset1.csv", row.names = FALSE)

write.csv(df3, file="C:\\Users\\jaskirat singh\\Downloads\\datascienceproj\\dataset3.csv", row.names = FALSE)
```

The final preprocessed data is:



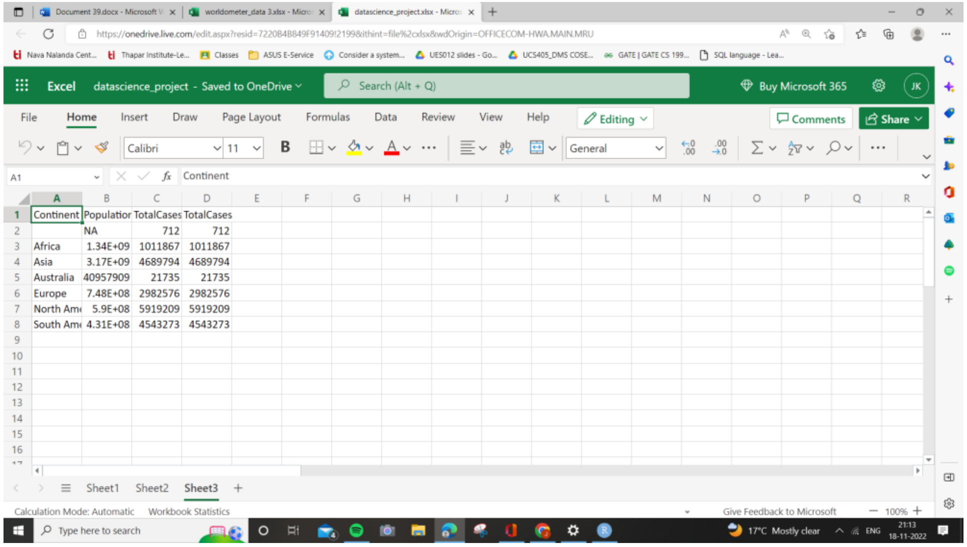
A1: Country.Region

Country.Re	Tot.Cases.	Deaths.1M	Tests.1M.pop
USA	15194	492	190640
Brazil	13716	464	62085
India	1466	30	16035
Russia	5974	100	203623
South Afri	9063	162	53044
Mexico	3585	391	8189
Peru	13793	619	75521
Chile	19165	517	92022
Colombia	7023	234	35374
Spain	7582	610	151087
Iran	3806	214	31068
UK	4537	683	257873
Saudi Arab	8152	88	104277
Pakistan	1274	27	9304
Banglades	1514	20	7432



A1 fx Country.Region

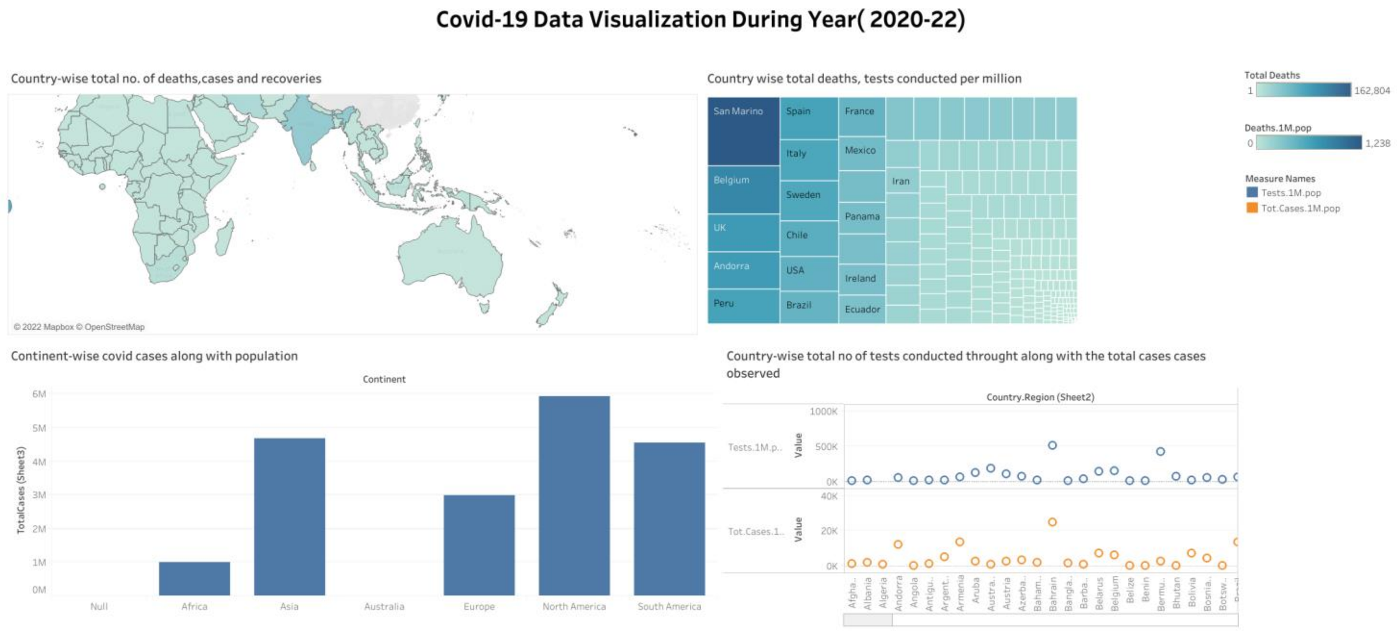
	A	B	C	D	E
1	Country.Re	TotalCases	TotalDeath	TotalRecov	ActiveCases
2	USA	5032179	162804	2576668	2292707
3	Brazil	2917562	98644	2047660	771258
4	India	2025409	41638	1377384	606387
5	Russia	871894	14606	676357	180931
6	South Afri	538184	9604	387316	141264
7	Mexico	462690	50517	308848	103325
8	Peru	455409	20424	310337	124648
9	Chile	366671	9889	340168	16614
10	Colombia	357710	11939	192355	153416
11	Spain	354530	28500	58878.98	NA
12	Iran	320117	17976	277463	24678
13	UK	308134	46413	58878.98	NA
14	Saudi Arab	284226	3055	247089	34082
15	Pakistan	281863	6035	256058	19770
16	Banglades	249651	3306	143824	102521



A1 fx Continent

	A	B	C	D
1	Continent	Populatior	TotalCases	TotalCases
2		NA	712	712
3	Africa	1.34E+09	1011867	1011867
4	Asia	3.17E+09	4689794	4689794
5	Australia	40957909	21735	21735
6	Europe	7.48E+08	2982576	2982576
7	North Ame	5.9E+08	5919209	5919209
8	South Ame	4.31E+08	4543273	4543273

The dashboard:



Covid-19 Data Visualization During Year(2020-22)

Country-wise total no. of deaths,cases and recoveries

Country wise total deaths, tests conducted per million

Continent-wise covid cases along with population

Country-wise total no of tests conducted throught along with the total cases cases observed