Project 1: COVID Vaccinations Around the World

R Markdown

Emily Qiu, eq887]

Introduction:

The data sets chosen for this experiment are Countries of the World which consists of statistics of every country (https://www.kaggle.com/fernandol/countries-of-the-world) and COVID-19 World Vaccination Progress which consists of COVID vaccination statistics (https://www.kaggle.com/gpreda/covid-world-vaccination-progress) in order to observe vaccinations depending on each country. The first data set or 'countries.of.the.world' has twenty total variables including: Country, Region, Population,Area (sq. mi.), Pop. Density (per sq. mi.), Coastline (coast/area ratio), Net migration, Infant mortality (per 1000 births), GDP ($ per capita), and Literacy (%). The second data set or 'vaccinations' has fifteen total variables including: country, iso_code, date, total_vaccinations, people_vaccinated, people_fully_vaccinated, daily_vaccinations_raw, daily_vaccinations, total_vaccinations_per_hundred, and people_vaccinated_per_hundred. The data sets were both acquired from kaggle and I chose them because they are related to one of the pressing, present issues: COVID-19. I thought it would be interesting to see how the condition of a country can influence vaccination data which can serve as a representation of not only the country's medical care but also infection rate as well.

Tidy:

```
library(tidyverse)
library(readxl)
countries.of.the.world<-read_csv("countries of the world.csv")
vaccinations<-read_csv("country_vaccinations.csv")
# Import the data
countries<-countries.of.the.world %>% mutate_all(na_if,"")
#mutate_alll affects every variable, and na_if converts to NA
```

Both data sets were already tidy. The only changes that needed to be made was converting blank cells into NAs in the 'countries.of.the.world' data set and renaming the data set 'countries.of.the.world' to just 'countries' for easier use.

Join/Merge:

```
countryvacc<-countries%>% left_join(vaccinations, by = c("Country"="country"))
#return all rows and all columns from both data sets
countryvacc<-countryvacc%>%select(-'Net migration')%>%select(-"Literacy (%)")%
#select(-) to remove variables
```

I used the left_join() function because I wanted to merge operations between both datasets by returning all of the rows from 'countries' and any matching rows from 'vaccinations' without returning values of this dataset that did not already exist in the first dataset because I wanted to observe how the condition of a country can influence vaccination data which requires data from both 'countries' and vaccinations. I dropped many variables that I felt were either irrelevant towards the relationship between country and vaccinations such as Net migration, Literacy, Coastline, Phones, Climate, Arable, Crops, Agriculture, Service, and dates. I also dropped variables that were repetitive of another variable such as daily vaccination raw, total vaccination per hundred, people fully vaccinated per hundred, people vaccinated per hundred, and source website. The size of the cleaned, joined data set was 14 variables and 6676 obervation

Summary statistics:

```
#find date with the most date
countryvacc %>% select(date) %>% mutate(mode(date))
```

```
## # A tibble: 6,676 x 2
##    date       `mode(date)`
##    <date>     <chr>
##  1 2021-02-22 numeric
##  2 2021-02-23 numeric
##  3 2021-02-24 numeric
##  4 2021-02-25 numeric
##  5 2021-02-26 numeric
##  6 2021-02-27 numeric
##  7 2021-02-28 numeric
##  8 2021-03-01 numeric
##  9 2021-03-02 numeric
```

```
## 10 2021-03-03 numeric
## # … with 6,666 more rows
```

```
#Arrange descending order of vaccines with selected date
countryvacc%>%filter(date =="2021-02-22")%>%arrange(desc(total_vaccinations))
```

```
## # A tibble: 102 x 14
##    Country  Region Population `Area (sq. mi.)` `Pop. Density (… `Infant mor
##    <chr>    <chr>       <dbl>            <dbl> <chr>
##  1 United … NORTH…  298444215          9631420 31,0
##  2 United … WESTE…   60609153           244820 247,6
##  3 India    ASIA … 1095351995          3287590 333,2
##  4 Israel   NEAR …    6352117            20770 305,8
##  5 Brazil   LATIN…  188078227          8511965 22,1
##  6 Turkey   NEAR …   70413958           780580 90,2
##  7 United … NEAR …    2602713            82880 31,4
##  8 Germany  WESTE…   82422299           357021 230,9
##  9 France   WESTE…   60876136           547030 111,3
## 10 Italy    WESTE…   58133509           301230 193,0
## # … with 92 more rows, and 8 more variables: GDP ($ per capita) <dbl>,
## #   Birthrate <dbl>, Deathrate <dbl>, date <date>, total_vaccinations <dbl>
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>
```

```
#create a new variable that represent the proportion of the population vaccina
countryvacc<-countryvacc%>%mutate(prop_vac=(people_vaccinated/Population)*100)
```

2021-02-22 was the date wth the most data and the United states had the most Total vaccinations. A new variable was also created to represent the proportion of the population vaccinated of each Country.

```
install.packages("kableExtra")
```

```
##
## The downloaded binary packages are in
##  /var/folders/4f/t6dbjc0525l98dbtj98m9ly40000gn/T//RtmpjpbUrx/downloaded_pa
```

```
library(kableExtra)
```

```
#mean of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ mean(.x, na.rm = TRUE)))
```

```
## # A tibble: 1 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1  55384637.         1102594.                  1701.            15662.
## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#mean of numerical variables grouped by Region
countryvacc %>%group_by(Region)%>% summarize(across(where(is.numeric), ~ mean(
```

Me

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate
ASIA (EX. NEAR EAST)	319106965	1882201.95	3301.2121	7315.014	1623.9366
BALTICS	2396979	58235.02	807.9421	11318.182	934.5041
C.W. OF IND. STATES	52530469	5601054.37	3090.8370	6161.755	1040.3009
EASTERN EUROPE	11684144	113130.14	1290.3158	10122.333	917.9100
LATIN AMER. &	26865467	1029975.91	1866.5737	9057.289	1766.5090

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate
CARIB					
NEAR EAST	10762690	288328.84	1670.2419	15336.717	1810.2268
NORTHERN AFRICA	33994990	950187.78	2874.3708	4782.022	2032.9101
NORTHERN AMERICA	105047816	6639883.67	564.4650	32481.469	1026.8252
OCEANIA	9895476	3271917.15	887.0513	21055.000	1495.4051
SUB-SAHARAN AFRICA	16118729	337637.85	5297.2969	4181.771	2589.8125
WESTERN EUROPE	17247286	159948.81	438.6235	27490.187	999.5952

```
#standard deviation of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ sd(.x, na.rm = TRUE)))
```

```
## # A tibble: 1 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1 187746666.         2891781.                  2213.            11433.
## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#standard deviation of numerical variables by Region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ var(.x
```

St

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	I
ASIA (EX. NEAR	2.282301e+17	9.379370e+12	12087126.97	75685526.0	88(

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	
EAST)					
BALTICS	8.702746e+14	8.715459e+07	11962.10	734937.8	
C.W. OF IND. STATES	3.570792e+15	5.676349e+13	6489113.12	4828721.4	42
EASTERN EUROPE	1.268282e+14	9.021914e+09	556073.23	23320454.9	12
LATIN AMER. & CARIB	2.262936e+15	3.852355e+12	1524230.27	61683216.9	32
NEAR EAST	3.738240e+14	3.258929e+11	785974.02	38624308.9	80
NORTHERN AFRICA	2.317591e+14	7.203698e+11	3086695.92	1211491.3	7
NORTHERN AMERICA	1.742230e+16	1.919198e+13	224106.75	35059479.9	27
OCEANIA	7.868395e+13	1.386956e+13	1031948.72	85863519.0	27
SUB-SAHARAN AFRICA	6.289696e+14	2.018147e+11	19021984.54	13725619.8	182
WESTERN EUROPE	5.972517e+14	3.076482e+10	17226.59	59062710.5	6

```
#minimum of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ min(.x, na.rm = TRUE)))
```

```
## # A tibble: 1 x 11
```

```
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1       7026                2                     19               500

## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>


#minimum of numerical variables by Region
countryvacc  %>%group_by(Region) %>%summarize(across(where(is.numeric), ~ min(
```

Minimum

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deathrat
ASIA (EX. NEAR EAST)	359008	28	64	500	10	12
BALTICS	1324333	45226	689	10200	875	109
C.W. OF IND. STATES	2976372	29800	711	1000	16	8
EASTERN EUROPE	2010347	20273	393	2200	107	52
LATIN AMER. & CARIB	9439	102	86	1600	129	5

NEAR EAST	698585	360	615	600	178	3
NORTHERN AFRICA	273008	163610	31	4000	1552	34
NORTHERN AMERICA	7026	53	65	6900	114	7
OCEANIA	11810	21	469	800	21	6
SUB-SAHARAN AFRICA	7502	374	19	500	41	2
WESTERN EUROPE	27928	2	37	17500	93	8

```
#maximum of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ max(.x, na.rm = TRUE)))
```

```
## # A tibble: 1 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1 1313973713         17075200                  19119             55100
## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#maximum of numerical variables by Region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ max(.x
```

Maximu

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deatl
ASIA (EX. NEAR EAST)	1313973713	9596960	16307	28800	3549	

BALTICS	3585906	65200	955	12300	1004	
C.W. OF IND. STATES	142893540	17075200	1107	8900	3265	
EASTERN EUROPE	38536869	312685	2643	19000	1511	
LATIN AMER. & CARIB	188078227	8511965	7345	35000	3644	
NEAR EAST	70413958	1960582	5025	23200	4289	
NORTHERN AFRICA	78887007	2381740	4162	6900	2649	
NORTHERN AMERICA	298444215	9984670	1582	37800	1593	
OCEANIA	20264082	7686850	5516	29000	3305	
SUB-SAHARAN AFRICA	131859731	2505810	19119	11400	5073	
WESTERN EUROPE	82422299	547030	624	55100	1445	

Region Population Area (sq. mi.) Infant mortality (per 1000 births) GDP ($ per capita) Birthrate Death

```
#median of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ median(.x, na.rm = TRUE))
```

```
## # A tibble: 1 × 11
```

```
## # A tibble: 1 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                 <dbl>             <dbl>
## 1    7523934            93030                   851             12300
## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#median of numerical variables by Region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ median
```

Median

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deathra
ASIA (EX. NEAR EAST)	84402966	377835	2418	3700	1551	6
BALTICS	2274735	64589	787	11400	924	13
C.W. OF IND. STATES	15233244	603700	1539	6100	995	14
EASTERN EUROPE	7385367	88361	857	10600	965	11
LATIN AMER. & CARIB	6822378	176220	2047	6300	1828	5

NEAR EAST	3102229	20770	1727	16900	1852	4
NORTHERN AFRICA	33241259	446550	4162	4000	2198	5
NORTHERN AMERICA	33098932	9631420	475	36000	1078	7
OCEANIA	4076140	268680	585	21600	1376	7
SUB-SAHARAN AFRICA	10723106	196190	5551	2150	2801	9
WESTERN EUROPE	5450661	83870	468	27600	1038	9

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deathra

```
#range of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ range(.x, na.rm = TRUE)))
```

```
## # A tibble: 2 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1       7026                2                     19               500
## 2 1313973713         17075200                  19119             55100
## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#range of numerical variables by Region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ range(
```

Range

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deatl
ASIA (EX. NEAR	359008	28	64	500	10	

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Death
EAST)						
ASIA (EX. NEAR EAST)	13139973713	9596960	16307	28800	3549	
BALTICS	1324333	45226	689	10200	875	
BALTICS	3585906	65200	955	12300	1004	
C.W. OF IND. STATES	2976372	29800	711	1000	16	
C.W. OF IND. STATES	142893540	17075200	11076	8900	3265	
EASTERN EUROPE	2010347	20273	393	2200	107	
EASTERN EUROPE	38536869	312685	2643	19000	1511	
LATIN AMER. & CARIB	9439	102	86	1600	129	
LATIN AMER. & CARIB	188078227	8511965	7345	35000	3644	
NEAR EAST	698585	360	615	600	178	
NEAR EAST	70413958	1960582	5025	23200	4289	
NORTHERN	273008	163610	31	4000	1552	

AFRICA						
NORTHERN AMERICA	78887007	9281740	4162	6900	2649	
Region	**Population**	**Area (sq. mi.)**	**Infant mortality (per 1000 births)**	**GDP ($ per capita)**	**Birthrate**	**Death**
NORTHERN AMERICA	7026	53	65	6900	114	
NORTHERN AMERICA	298444215	9984670	1582	37800	1593	
OCEANIA	11810	21	469	800	21	
OCEANIA	20264082	7686850	5516	29000	3305	
SUB-SAHARAN AFRICA	7502	374	19	500	41	
SUB-SAHARAN AFRICA	131859731	2505810	19119	11400	5073	
WESTERN EUROPE	27928	2	37	17500	93	
WESTERN EUROPE	82422299	547030	624	55100	1445	

```
#quantile of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ quantile(.x, na.rm = TRUE
```

```
## # A tibble: 5 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1       7026                2                     19               500
## 2    2010347            20770                    471              5400
```

```
## 3    7523934           93030                 851            12300
## 4   32930091          449964                2055            26700
## 5 1313973713        17075200               19119            55100
## # … with 6 more variables: Deathrate <dbl>, total_vaccinations <dbl>,
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#quantile of numerical variables by Region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ quanti
```

Quantil

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deatl
ASIA (EX. NEAR EAST)	359008	28	64.00	500	10	12
ASIA (EX. NEAR EAST)	13881427	65610	356.00	2100	934	62
ASIA (EX. NEAR EAST)	84402966	377835	2418.00	3700	1551	69
ASIA (EX. NEAR EAST)	245452739	1919440	5629.00	5000	2201	82
ASIA (EX. NEAR	1313973713	9596960	16307.00	28800	3549	203

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Death
EAST)						
BALTICS	1324333	45226	689.00	10200	875	109
BALTICS	1324333	45226	689.00	10200	875	109
BALTICS	2274735	64589	787.00	11400	924	132
BALTICS	3585906	65200	955.00	12300	1004	136
BALTICS	3585906	65200	955.00	12300	1004	136
C.W. OF IND. STATES	2976372	29800	711.00	1000	16	8
C.W. OF IND. STATES	10293011	207600	1539.00	5400	882	97
C.W. OF IND. STATES	15233244	603700	1539.00	6100	995	140
C.W. OF IND. STATES	142893540	17075200	2921.00	8900	1116	146
C.W. OF IND. STATES	142893540	17075200	11076.00	8900	3265	146
EASTERN EUROPE	2010347	20273	393.00	2200	107	52
EASTERN EUROPE	4494749	48845	741.00	7000	898	98

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Death
EASTERN EUROPE	7385367	88361	857.00	10600	965	114
EASTERN EUROPE	9981334	110910	2058.00	19000	985	131
EASTERN EUROPE	38536869	312685	2643.00	19000	1511	142
LATIN AMER. & CARIB	9439	102	86.00	1600	129	5
LATIN AMER. & CARIB	287730	21040	995.00	4800	1523	47
LATIN AMER. & CARIB	6822378	176220	2047.00	6300	1828	57
LATIN AMER. & CARIB	28302603	1138910	2563.00	9900	2069	67
LATIN AMER. & CARIB	188078227	8511965	7345.00	35000	3644	121
NEAR EAST	698585	360	615.00	600	178	3
NEAR EAST	885359	11437	995.00	11800	1556	25
NEAR EAST	3102229	20770	1727.00	16900	1852	41
NEAR EAST	6352117	212460	1861.00	19800	2194	59
NEAR EAST	70413958	1960582	5025.00	23200	4289	76

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Death
NEAR EAST	70413958	1906362	3029.00	23200	4269	7(
NORTHERN AFRICA	273008	163610	34.00	4000	1552	34
NORTHERN AFRICA	32930091	446550	246.00	4000	1714	46
NORTHERN AFRICA	33241259	446550	4162.00	4000	2198	55
NORTHERN AFRICA	33241259	1570018	4162.00	6000	2198	55
NORTHERN AFRICA	78887007	2381740	4162.00	6900	2649	55
NORTHERN AMERICA	7026	53	65.00	6900	114	7
NORTHERN AMERICA	65773	2166086	65.00	29800	1078	7
NORTHERN AMERICA	33098932	9631420	475.00	36000	1078	77
NORTHERN AMERICA	298444215	9984670	853.00	37800	1414	82
NORTHERN AMERICA	298444215	9984670	1582.00	37800	1593	82
OCEANIA	11810	21	469.00	800	21	6
OCEANIA	4076140	268680	469.00	21600	1214	75
OCEANIA	4076140	268680	585.00	21600	1376	75
OCEANIA	20264082	7686850	585.00	29000	1376	75
OCEANIA	20264082	7686850	5516.00	29000	3305	8(

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deatl
OCEANIA	26264082	7686850	5910.00	29000	3503	62
SUB-SAHARAN AFRICA	7502	374	19.00	500	41	2
SUB-SAHARAN AFRICA	81541	455	1553.00	1400	1543	62
SUB-SAHARAN AFRICA	10723106	196190	5551.00	2150	2801	94
SUB-SAHARAN AFRICA	19686505	390580	6851.25	7800	3972	156
SUB-SAHARAN AFRICA	131859731	2505810	19119.00	11400	5073	297
WESTERN EUROPE	27928	2	37.00	17500	93	8
WESTERN EUROPE	299388	2586	405.00	22000	919	71
WESTERN EUROPE	5450661	83870	468.00	27600	1038	93
WESTERN EUROPE	16491461	323802	516.00	30000	1113	102
WESTERN EUROPE	82422299	547030	624.00	55100	1445	129

```
#Median Absolute Deviation of numerical variables
countryvacc %>%summarize(across(where(is.numeric), ~ mad(.x, na.rm = TRUE)))
```

```
## # A tibble: 1 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <dbl>            <dbl>                  <dbl>             <dbl>
## 1  10607570.          137538.                   771.             12157.
## # … with 6 more variables: Deathrate <dbl>, total vaccinations <dbl>,
```

```
## #   people_vaccinated <dbl>, people_fully_vaccinated <dbl>,
## #   daily_vaccinations <dbl>, prop_vac <dbl>
```

```
#Median Absolute Deviation of numerical variables by Region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ mad(.x
```

Median A

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate
ASIA (EX. NEAR EAST)	118475776	559150.7292	3144.5946	2372.16	963.6900
BALTICS	1409066	905.8686	145.2948	1334.34	72.6474
C.W. OF IND. STATES	10780911	766652.4600	299.4852	4003.02	179.3946
EASTERN EUROPE	4285630	58586.4216	610.8312	4892.58	29.6520
LATIN AMER. & CARIB	10012398	260875.3308	1355.0964	4003.02	452.1930
NEAR EAST	3436560	29807.6730	597.4878	6819.96	438.8496
NORTHERN AFRICA	0	0.0000	0.0000	0.00	0.0000
NORTHERN AMERICA	48974962	523728.4500	560.4228	2668.68	498.1536

OCEANIA	5937367	397590.3246	171.9816	10971.24	240.1812
SUB-SAHARAN AFRICA	15777184	290196.7410	5285.1000	1260.21	1776.1548
WESTERN EUROPE	7764965	124108.4460	77.0952	4151.28	111.1950

```
#The number of unique values in a set of vector
countryvacc %>%summarize(across(where(is.numeric), ~ n_distinct(.x, na.rm = TR
```

```
## # A tibble: 1 x 11
##   Population `Area (sq. mi.)` `Infant mortality (pe… `GDP ($ per capi… Birt
##        <int>            <int>                  <int>             <int>
## 1        227              226                    218               130
## # … with 6 more variables: Deathrate <int>, total_vaccinations <int>,
## #   people_vaccinated <int>, people_fully_vaccinated <int>,
## #   daily_vaccinations <int>, prop_vac <int>
```

```
#The number of unique values in a set of vector by region
countryvacc %>%group_by(Region)%>%summarize(across(where(is.numeric), ~ n_dist
```

The number o

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deathrate
ASIA (EX. NEAR EAST)	28	28	28	25	28	28
BALTICS	3	3	3	3	3	3
C.W. OF IND. STATES	12	12	12	12	12	12
EASTERN EUROPE	12	12	12	12	11	11

Region	Population	Area (sq. mi.)	Infant mortality (per 1000 births)	GDP ($ per capita)	Birthrate	Deathrate
LATIN AMER. & CARIB	45	44	45	40	43	44
NEAR EAST	16	16	16	15	16	16
NORTHERN AFRICA	6	6	5	4	5	5
NORTHERN AMERICA	5	5	5	5	5	5
OCEANIA	21	21	18	19	20	19
SUB-SAHARAN AFRICA	51	51	51	29	51	50
WESTERN EUROPE	28	28	28	25	28	28

```
install.packages("psych")
```

```
##
## The downloaded binary packages are in
##  /var/folders/4f/t6dbjc0525l98dbtj98m9ly40000gn/T//RtmpjpbUrx/downloaded_pa
```

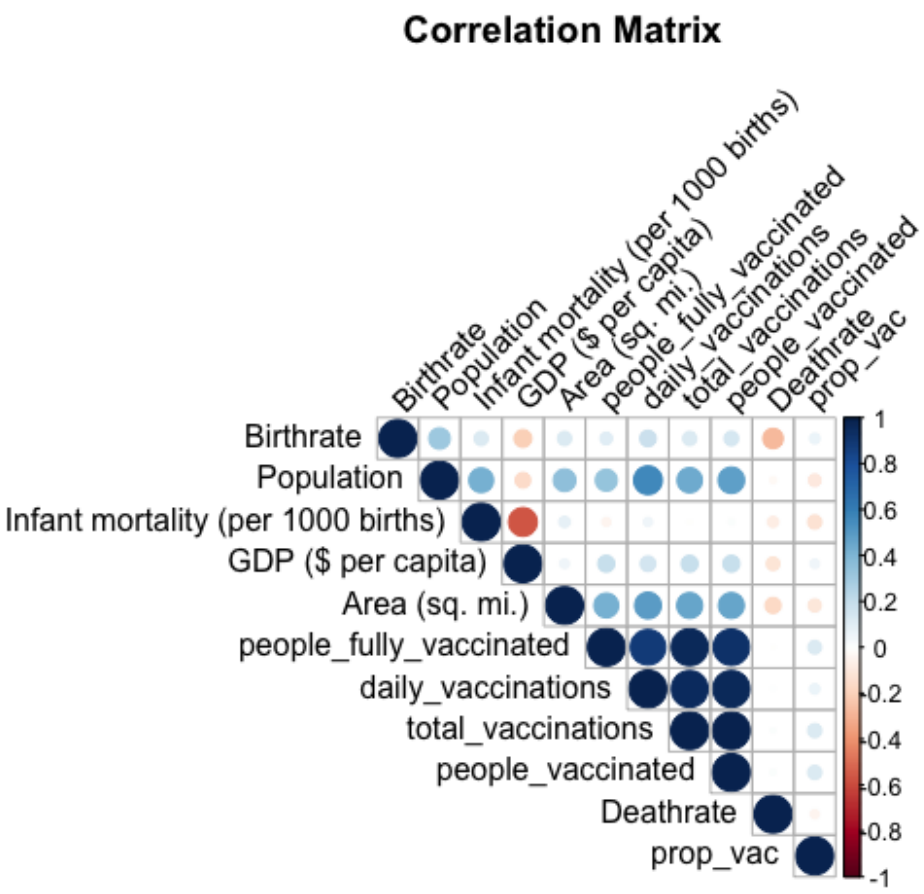
```
library(psych)
```

```
#select_if was used to select only numerical variables and kbl() was used to c
countryvacc %>% select_if(is.numeric)%>% describe()%>%kbl(caption = "Summary s
```

	vars	n	mean	sd	
Population	1	6676	5.538464e+07	1.877467e+08	7.523
Area (sq. mi.)	2	6676	1.102594e+06	2.891781e+06	9.303
Infant mortality (per 1000 births)	3	6673	1.700980e+03	2.213270e+03	8.510

	vars	n	mean	sd	
GDP ($ per capita)	4	6675	1.566199e+04	1.143262e+04	1.2300
Birthrate	5	6604	1.419740e+03	8.137397e+02	1.1940
Deathrate	6	6603	7.721914e+02	3.846155e+02	7.6800
total_vaccinations	7	4090	2.368479e+06	8.727029e+06	2.2373
people_vaccinated	8	3596	1.884505e+06	6.412606e+06	2.0225
people_fully_vaccinated	9	2396	8.922204e+05	3.449809e+06	1.0896
daily_vaccinations	10	6412	6.042178e+04	2.081636e+05	5.8115
prop_vac	11	3596	7.786520e+00	1.458094e+01	2.9405

```
countryvacc1<-countryvacc %>% select_if(is.numeric) %>%na.omit
#select numerical and remove NA
countryvacc2<-signif(cor(countryvacc1),2)
#returnns integer values
library(corrplot)
#corrplot is used to create a correlation plot
corrplot(countryvacc2, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, title="Correlation Matrix", mar=c(0,0,
```



From the summary statistics it can be seen that the region Asia has the highest mean, maximum and median population but Northern America has the highest mean and largest range of total vaccinations as well as people vaccinated which suggests that population size isn't a sole reason for the number of vaccines a country has. Both regions have the similar standard deviations while Baltics has the highest with the lowest population mean. NORTHERN AMERICA has the lowest minimum population while Commonwealth of Independent States has the highest minimum. Even though the Northern America region has the most vaccines it does not have the highest proportion of their population, NEAR EAST has the highest proportion of their population vaccinated while OCEANIA has the lowest. From the graphical display of a correlation matrix population has the highest positive correlation in daily vaccinations and Area shows a slight positive correlation in vaccinations while birthrate, infant mortality, GDP and death rate show very little correlation.

Visualizations

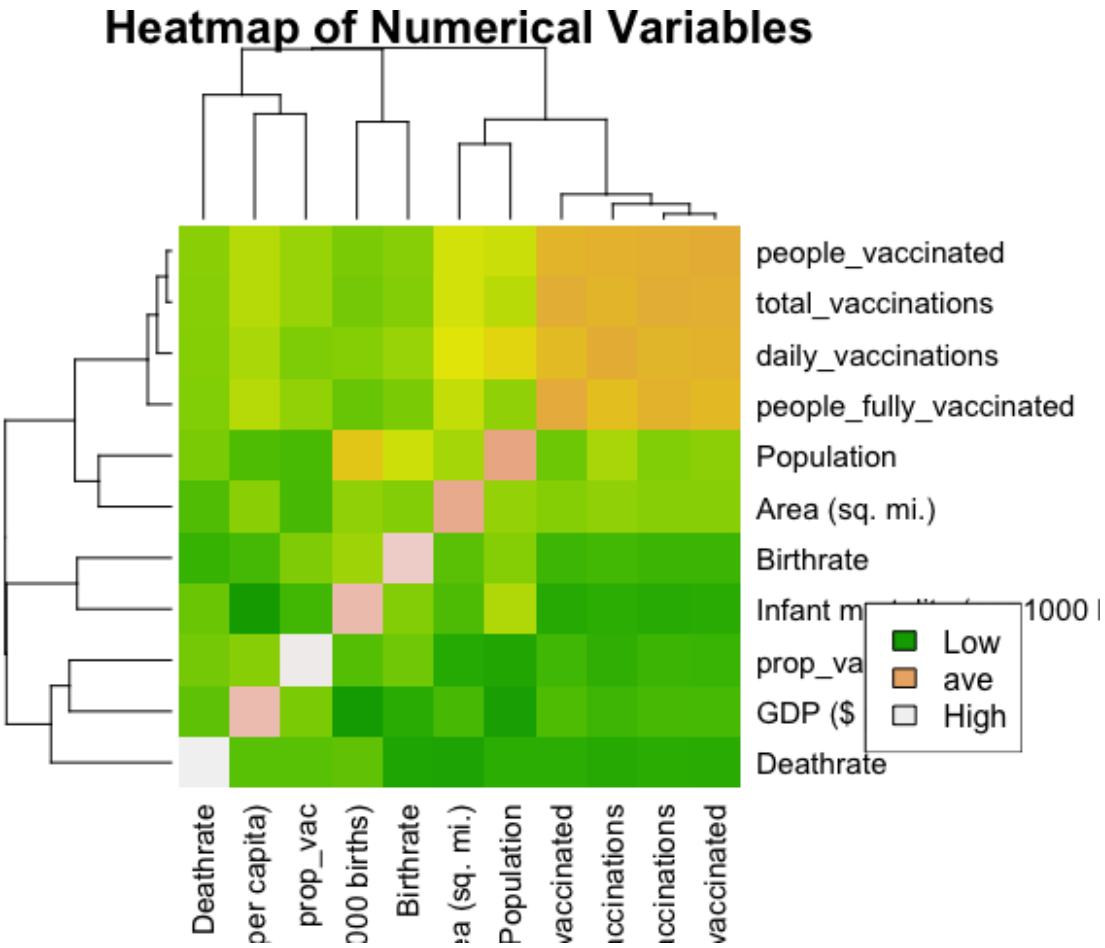
```
install.packages("ggplot2")
```

```
##
```

```
## The downloaded binary packages are in
##  /var/folders/4f/t6dbjc0525l98dbtj98m9ly40000gn/T//RtmpjpbUrx/downloaded_pa
```

```
library(ggplot2)
```

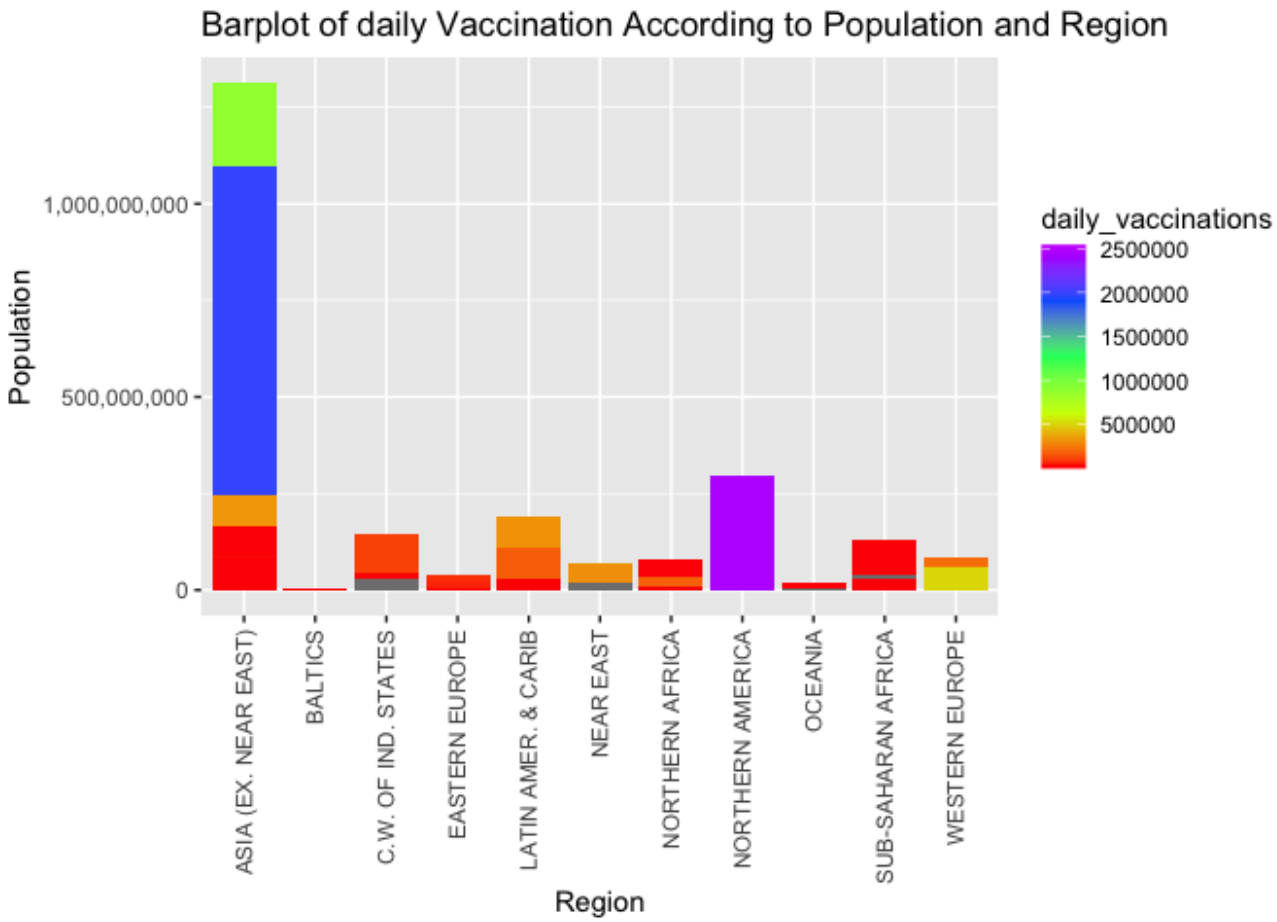
```
heatmap(countryvacc2, scale="column", col = terrain.colors(256), main="Heatmap
#heatmap produces high quality matrix and offers statistical tools to normaliz
legend(x="bottomright", legend=c("Low", "ave", "High"),
     fill=terrain.colors(3))
```




```
#creates key
```

From this heatmap it can be seen that that most of the correlations between the ‘countries’ dataset and ‘vaccines’ dataset are pretty low and the only signs of average correlation are amongst the vaccinations. Out of all the variables in the ‘countries’ data set Area and Populations seem to have the highest correlations to vaccinations compared to the rest with daily vaccinations being the highest with Population.

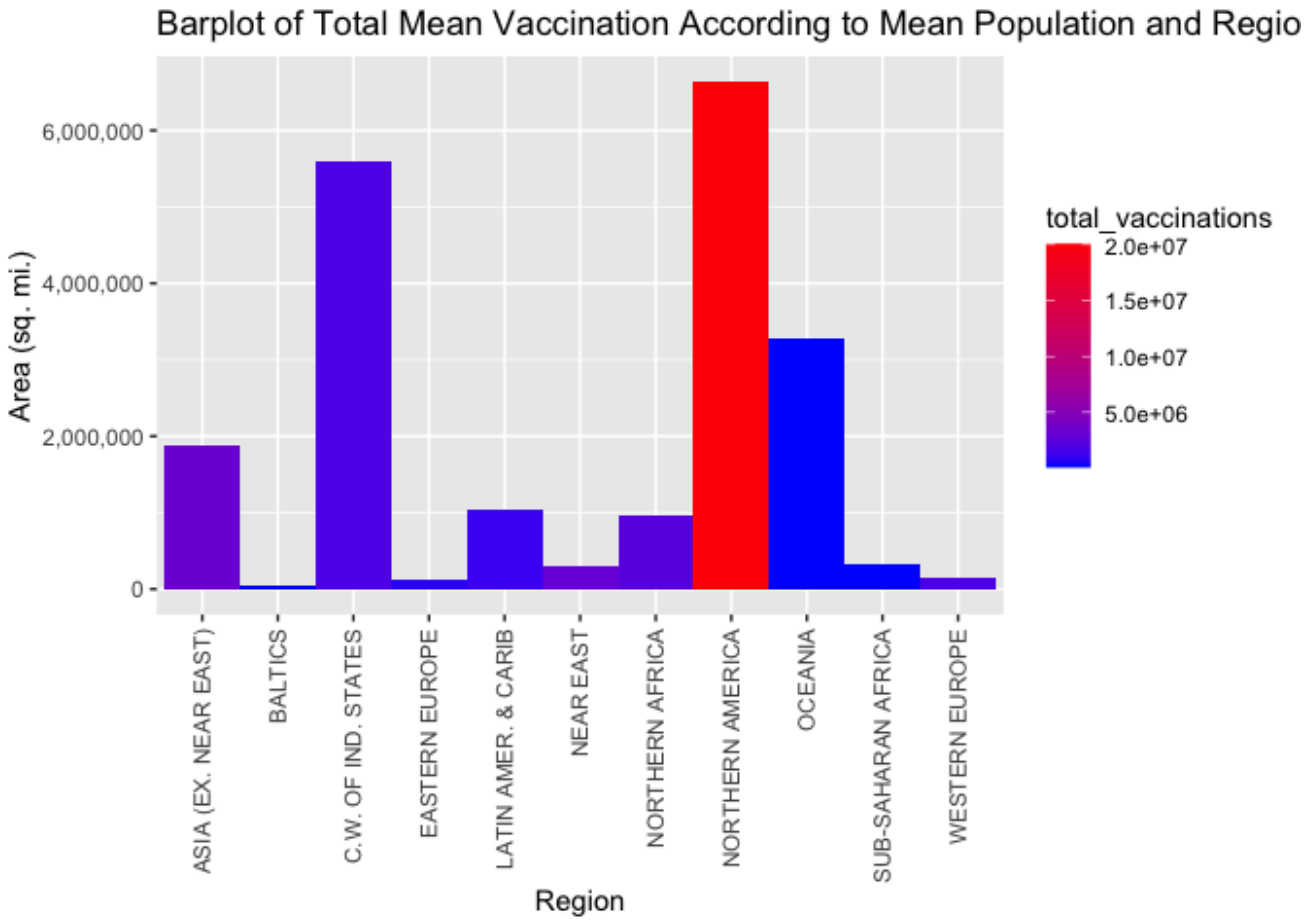
```
#create bar plot with three variables
countryvacc%>%ggplot(aes(x = Region, y = Population, fill =daily_vaccinations)
               geom_bar(stat = "identity", fun.y = "mean", na.rm = TRUE,
       position = position_dodge(width = 0.9)) +
               labs(list(x = "x", y = "count",fill = "group"))+ theme(axis.te
```




```
#scale_fill_gradientn adds color, scale_y_continuous adjusts Y-values from sci
```

This barplot shows that while Asia has the highest Population with high daily vaccinations, Northern America has by far the highest daily vaccination even though it has a smaller population. Sub-Saharian Africa and Northern Africa have the least.

```
#create bar plot with three variables
countryvaccmean<-countryvacc %>%group_by(Region)%>% summarize(across(where(is.
#displaying statistics (using a stat="summary" function)
countryvaccmean%>%ggplot(aes(x = Region, y = `Area (sq. mi.)`, fill =total_vac
               geom_bar(stat = "summary", fun.y = "mean", width = 1, positior
               labs(list(x = "x", y = "count",fill = "group"))+ theme(axis.te
```

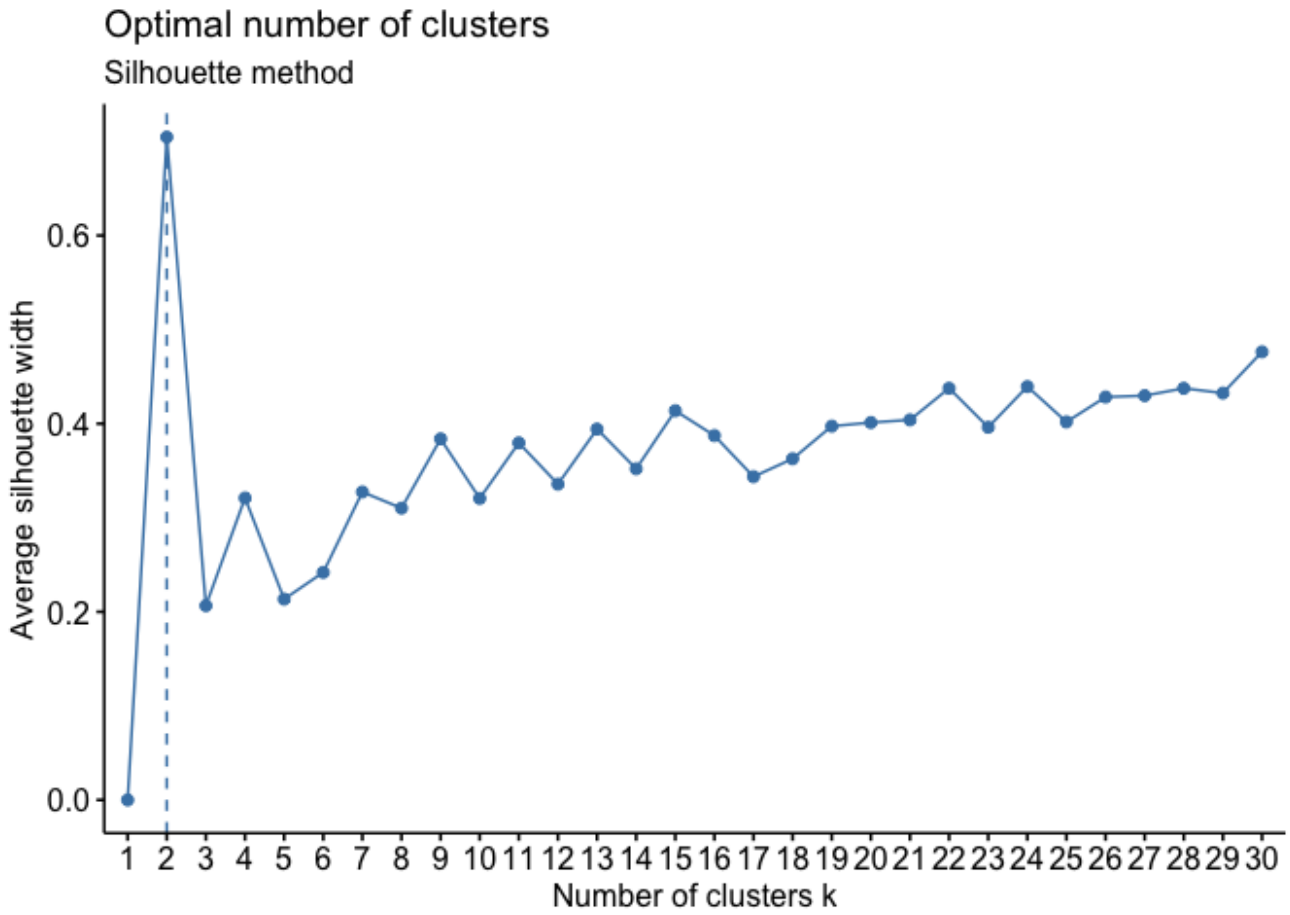



```
#theme rotates text
```

This barplot shows that northern America has the highest mean Area (sq.mi) and also has the highest total vaccinations. However Commonwealth of Independent States has the second highest mean Area and appears to possess a lower amount of mean total vaccinations than Asia and Near East which suggests little relationship.

Dimensionality Reduction

```
library(cluster)
library(ggpubr)
library(factoextra)
countryvac<-countryvacc1%>%mutate_if(is.numeric, scale)%>%na.omit
#remove NA and scale
fviz_nbclust(countryvac,kmeans, method = "silhouette",k.max = 30)+
  labs(subtitle = "Silhouette method")
```




```
#Dertemines and visualize the optimal number of clusters
```

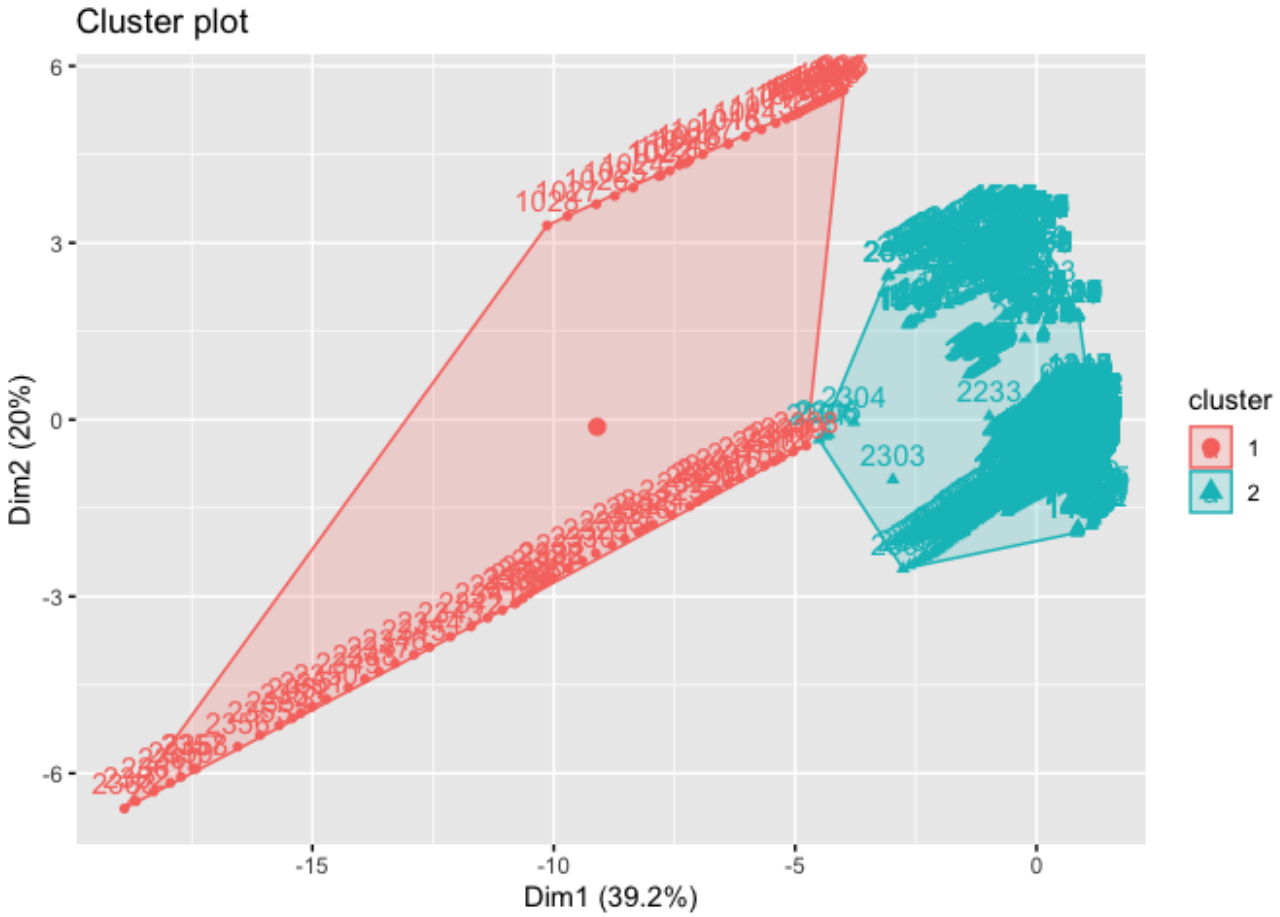
The Silhouette method shows that the optimal numbers of clusters is 2.

```
pam<-countryvac%>%pam(k=2)
countryvaccpam<-countryvac%>%mutate(cluster=pam$clustering)%>%relocate(cluster
#set 2 clusters and create a cluster variable
km    <- kmeans(countryvaccpam,2)
str(km)
```

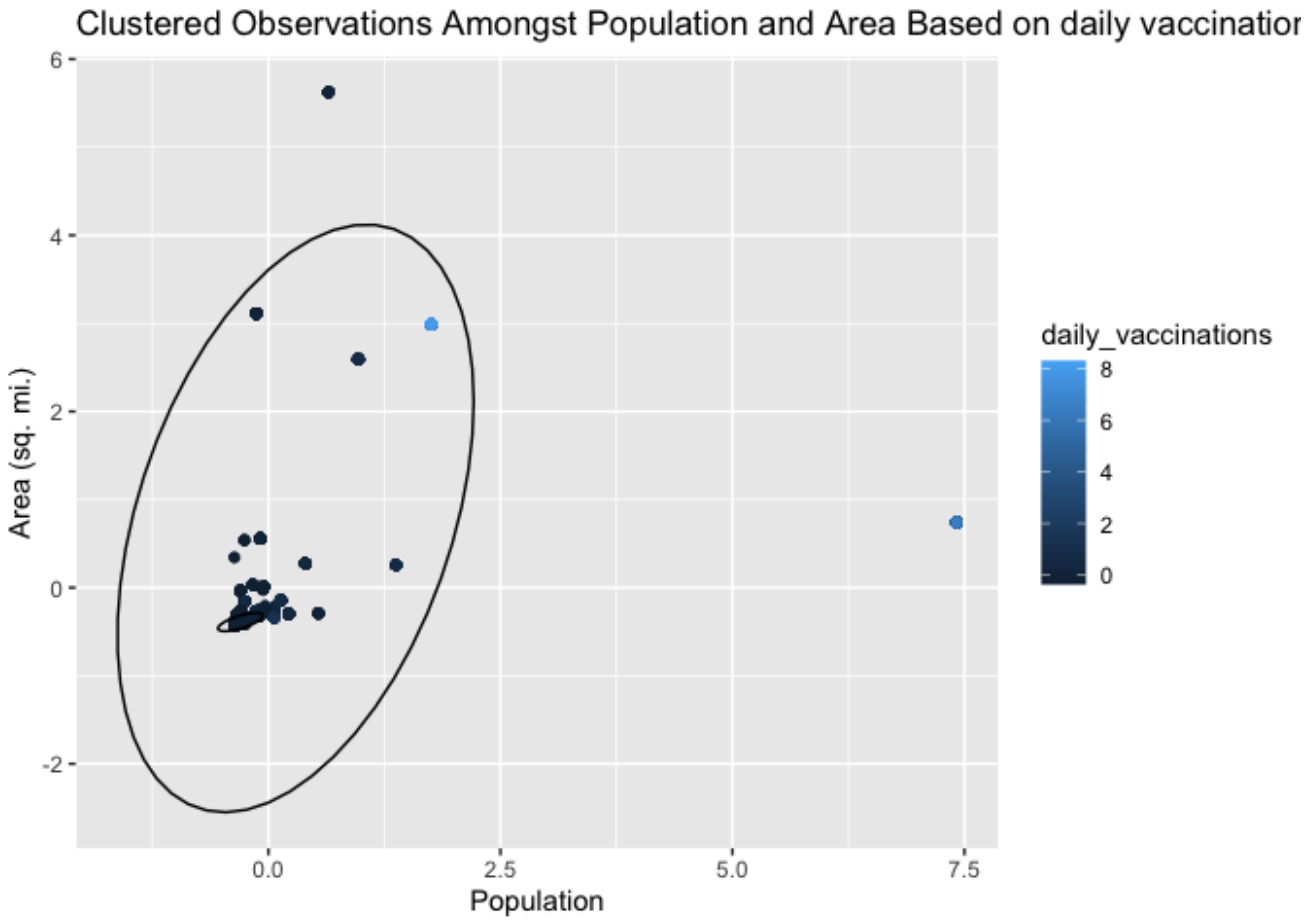
```
## List of 9
##  $ cluster     : int [1:2363] 2 2 2 2 2 2 2 2 2 2 ...
##  $ centers     : num [1:2, 1:12] 1 1.787 3.893 -0.154 2.139 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:12] "cluster" "Population" "Area (sq. mi.)" "Infant morta
##  $ totss       : num 26417
##  $ withinss    : num [1:2] 4330 14303
##  $ tot.withinss: num 18633
##  $ betweenss   : num 7784
##  $ size        : int [1:2] 90 2273
##  $ iter        : int 1
##  $ ifault      : int 0
```

```
##  - attr(*, "class")= chr "kmeans"
```

```
fviz_cluster(km, data = countryvaccpam)
```




```
 #kmeans analysis and graph
countryvaccpam%>%ggplot(aes(x=Population, y = `Area (sq. mi.)`, color = daily_
```




```
#create a scatter plot using 3 numerical variables
```

The cluster plot shows that cluster one is significantly larger than cluster two. From the scatter plot it can be see that the Observations among the Population and area are mostly in the same area with a consistent levels of daily vaccinations. There are only two noticeable points that would be considered to have high levels of daily vaccination suggesting that population and area don’t have a very strong relationship with the variable.