

Milestones One and Two(details about the dataset).

1. Why you are interested in it?

I am interested in the dataset because I wanted to know whether the vaccination rate per country is somehow related to the population wealth as well as what countries are getting them the most.

2. Where did you find the dataset (including the link)?

I found the data at Kaggle right [here](#).

3. Who (person, government, organization, company, etc.) created the dataset?

The dataset was created by Sagar Sahni an Indian student at Madhav Institute of Technology and Science.

4. When the data is from?

The last update of the data was made on August 10, 2021.

5. What kind of information is contained in the dataset?

Here is some of the content in the dataset:

- Country
- Country ISO Code
- Date
- Total number of vaccinations

- 🍌 The total number of people vaccinated
- 🍌 The total number of people fully vaccinated
- 🍌 Daily vaccinations (raw)
- 🍌 Daily vaccinations
- 🍌 Total vaccinations per hundred

Code Milestone Three, Four, and Five.

Hello professor, here is the code for milestone three (link [here](#)) and I added the link to the website? Bewere that I have a lot of code there because there is code for milestones one and two. Link in [here](#). But if look towards the end you will see the milestone three code...

Initializing Data, Cleaning Data, and Plot Histogram Graphs.

Initial code needed to read the file later.

```
library(readr)
```

Gets me the file loaded in countries.

```
vaccinations_per_country<-read_csv("country vaccinations.csv",na="-")
```

To view the vaccinations_per_country.

```
View(vaccinations_per_country)
```

Getting rid of the empty values.

```
vaccinations_per_country_without_na<-na.omit(vaccinations_per_country)
```

View the data w/t the empty values.

```
View(vaccinations_per_country_without_na)
```

To view the countries.

```
vaccinations_per_country_without_na$country
```

To obtain the name of all the columns.

```
colnames(vaccinations_per_country_without_na)
```

Put all the columns on a new var.

```
nameOfAllColumns<-colnames(vaccinations_per_country_without_na)
```

Histogram of tot vaccination per hundred empty values.

```
hist(vaccinations_per_country$total_vaccinations_per_hundred,main="Total Vaccinations Per Hundred  
WNA",xlab="Number Of Vaccinations",ylab="Population Up To  
Date",col="green",border="blue",las=1,breaks=20)
```

Histogram of tot vaccination per hundred.

```
hist(vaccinations_per_country_without_na$total_vaccinations_per_hundred,main="Total Vaccinations Per  
Hundred",xlab="Number Of Vaccinations",ylab="Population Up To  
Date",col="green",border="blue",las=1,breaks=20)
```

Histogram of Total Number Of People Vaccinations Per Hundred with empty values.

```
hist(vaccinations_per_country$people_vaccinated_per_hundred,main="Total Number Of People Vaccinations Per  
Hundred WNA",xlab="Population Immunized",ylab="Total Population Up To  
Date",col="green",border="blue",las=1,breaks=80,xlim=c(0,100),ylim = c(0,4000))
```

Histogram of Total Number Of People Vaccinations Per Hundred.

```
hist(vaccinations_per_country_without_na$people_vaccinated_per_hundred,main="Total Number Of People
Vaccinations Per Hundred",xlab="Population Immunized",ylab="Total Population Up To
Date",col="green",border="blue",las=1,breaks=80,xlim=c(0,100),ylim = c(0,1500))
```

People Fully Vaccinated Per Hundred with n/a.

```
hist(vaccinations_per_country$people_fully_vaccinated_per_hundred,main="People Fully Vaccinated Per
Hundred",xlab="Population Fully Immunized",ylab="Total Population Up To
Date",col="green",border="blue",las=1,breaks=20)
```

People Fully Vaccinated Per Hundred.

```
hist(vaccinations_per_country_without_na$people_fully_vaccinated_per_hundred,main="People Fully Vaccinated
Per Hundred",xlab="Population Fully Immunized",ylab="Total Population Up To
Date",col="green",border="blue",las=1,breaks=20)
```

Histogram Daily Vaccinations Per Million with n/a.

```
hist(vaccinations_per_country$daily_vaccinations_per_million,main="Daily Vaccinations Per
Million",xlab="Vaccination Number",ylab="Population
Number",col="green",border="blue",las=1,breaks=200,xlim=c(0,30000))
```

Histogram Daily Vaccinations Per Million.

```
hist(vaccinations_per_country_without_na$daily_vaccinations_per_million,main="Daily Vaccinations Per
Million",xlab="Vaccination Number",ylab="Population
Number",col="green",border="blue",las=1,breaks=200,xlim=c(0,30000))
```

Finding the Mean, Median, Variance, and Standard Deviation.

I have chosen the following variables, which are my columns: total_vaccinations_per_hundred (this is for Total Number of People Vaccinated Per Hundred) and people_vaccinated_per_hundred_hist (this one is for Total Vaccinations per Hundred.)

For Total Number of People Vaccinated Per Hundred (total_vaccinations_per_hundred)

Storing the numeric data in a variable.

```
numeric_values_of_total_vaccinations_per_hundred<-vaccinations_per_country_without_na$total_vaccinations_per_hundred
```

Calculate the mean.

```
mean(numeric_values_of_total_vaccinations_per_hundred)
```

```
mean_of_total_vaccinations_per_hundred<-mean(numeric_values_of_total_vaccinations_per_hundred)
```

```
mean_of_total_vaccinations_per_hundred
```

```
>>> 39.99296
```

Calculate the median

```
median(numeric_values_of_total_vaccinations_per_hundred)
```

```
median_of_total_vaccinations_per_hundred<-median(numeric_values_of_total_vaccinations_per_hundred)
```

```
median_of_total_vaccinations_per_hundred
```

```
>>> 25.775
```

Calculating the Variance

```
var(numeric_values_of_total_vaccinations_per_hundred)
```

```
variance_of_total_vaccinations_per_hundred<-var(numeric_values_of_total_vaccinations_per_hundred)
```

```
variance_of_total_vaccinations_per_hundred
```

```
>>> 1509.202
```

Calculating the Standard Deviation.

```
sd(numeric_values_of_total_vaccinations_per_hundred)
```

```
standard_deviation_of_total_vaccinations_per_hundred<-sd(numeric_values_of_total_vaccinations_per_hundred)
```

```
standard_deviation_of_total_vaccinations_per_hundred
```

```
>>> 38.84845
```

Mean = 39.99296, Median = 25.775, Variance = 1509.202, Standar Deviation = 38.84845.

For Total Vaccinations per Hundred(people_vaccinated_per_hundred).

Storing the numeric data in a variable.

```
hundred<-vaccinations_per_country_without_na$people_vaccinated_per_hundred
```

Calculate the mean.

```
mean(numeric_values_of_total_number_of_people_vaccinated_per_hundred)
```

```
mean_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred<-mean(numeric_values_of_total_n  
umber_of_people_vaccinated_per_hundred)
```

```
mean_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred
```

```
>>> 25.09553
```

Calculate the median

```
median(numeric_values_of_total_number_of_people_vaccinated_per_hundred)
```

```
median_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred<-median(numeric_values_of_tot  
al_number_of_people_vaccinated_per_hundred)
```

```
median_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred
```

```
>>> 17.75
```

Calculating the Variance

```
var(numeric_values_of_total_number_of_people_vaccinated_per_hundred)
```

```
variance_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred<-var(numeric_values_of_total_  
number_of_people_vaccinated_per_hundred)
```

```
variance_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred
```

```
>>> 487.9403
```

Calculating the Standard Deviation.

```
sd(numeric_values_of_total_number_of_people_vaccinated_per_hundred)

standard_deviation_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred<-sd(numeric_values_
of_total_number_of_people_vaccinated_per_hundred)

standard_deviation_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred

>>> 22.08937
```

Mean = 25.09553, Median = 17.75, Variance = 487.9403, Standar Deviation = 22.08937.

Observations of The Findings.

Mean: As expected the mean on the second graph is bigger than the one on the left, which, should be the case; because, by definition, the mean represents the expectation in this case we can that there is nothing exceptional or strange. Also, remember that in this case the more people we have the higher the vaccinated, which is the case of the right-hand-side graph.

Median: Same thing here, the larger the data set the higher this value is supposed to be; that is because the median is the number in the middle so we expect that number to be larger if values are > 0 .

Variance: The variance tells us how much spread is the data. A claim that I will represent you later with a graph.

Standard Deviation: Tell us the relation of the mean with the data at hand. A low std mean data is clustered around the man, spread away otherwise. Therefore we can say that,.....

Scatter Plot Graph and Corretaltion.

Suppose the following are true: z-value=5%, that means lower and upper bolund are -2.5% and 2.5% respectively.

Drawing the Scatter Plot.

```
plot(numeric_values_of_total_vaccinations_per_hundred,
numeric_values_of_total_number_of_people_vaccinated_per_hundred, main = "TVPH vs TNPVPH", xlab = "Total
Vaccinations Per Hundred", ylab = "Total Number of People Vaccinations Per Hundred", pch = 19, frame = FALSE)
```

Drwaing the Regression Line.

```
lines(lowess(numeric_values_of_total_vaccinations_per_hundred,  
numeric_values_of_total_number_of_people_vaccinated_per_hundred), col = "red")
```

Finding the Correlation

```
cor(numeric_values_of_total_vaccinations_per_hundred,  
numeric_values_of_total_number_of_people_vaccinated_per_hundred)  
>>> 0.9820427
```

Calculating the Confidence Interval of Total Vaccinations Per Hundred

```
table_of_numeric_values_of_total_vaccinations_per_hundred=table(numeric_values_of_total_vaccinations_per_hu  
ndred)
```

Get size of the matrix

```
n_for_numeric_values_of_total_vaccinations_per_hundred=nrow(table(numeric_values_of_total_vaccinations_per_  
hundred))
```

See what number you get

```
n_for_numeric_values_of_total_vaccinations_per_hundred  
>>> 7308
```

Calculating the Confidence Interval of Total Vaccinations Per Hundred

```
c(mean_of_total_vaccinations_per_hundred - qt(0.975, df =  
n_for_numeric_values_of_total_vaccinations_per_hundred-1) *  
standard_deviation_of_total_vaccinations_per_hundred /  
sqrt(n_for_numeric_values_of_total_vaccinations_per_hundred), mean_of_total_vaccinations_per_hundred +  
qt(0.975, df = n_for_numeric_values_of_total_vaccinations_per_hundred-1) *  
standard_deviation_of_total_vaccinations_per_hundred /  
sqrt(n_for_numeric_values_of_total_vaccinations_per_hundred))  
>>> 39.10213 40.88378
```


Transform the matrix back for Total Number of People Vaccinated Per Hundred

```
table_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred=table(numeric_values_of_total_number_of_people_vaccinated_per_hundred)
```

Get size of the matrix

```
n_for_numeric_values_of_total_number_of_people_vaccinated_per_hundred=nrow(table_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred)
```

See what number you get

```
N_for_numeric_values_of_total_number_of_people_vaccinated_per_hundred
```

```
>>> 5670
```

Calculating the Confidence Interval of Total Number of People Vaccinated Per Hundred

```
c(mean_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred - qt(0.975, df =  
n_for_numeric_values_of_total_number_of_people_vaccinated_per_hundred-1) *  
standard_deviation_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred /  
sqrt(n_for_numeric_values_of_total_number_of_people_vaccinated_per_hundred),  
mean_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred + qt(0.975, df =  
n_for_numeric_values_of_total_number_of_people_vaccinated_per_hundred-1) *  
standard_deviation_of_numeric_values_of_total_number_of_people_vaccinated_per_hundred /  
sqrt(n_for_numeric_values_of_total_number_of_people_vaccinated_per_hundred))
```

```
>>> 24.52045 25.67062
```

Linear Model (milestone six).

Create the linear model.

```
my_linear_model<-lm(numeric_values_of_total_vaccinations_per_hundred~  
numeric_values_of_total_number_of_people_vaccinated_per_hundred, vaccinations_per_country_without_na)
```

Check on the coefficient of the linear model.

```
coef(my_linear_model)
```

```
(Intercept) -3.349864, 1.7
```

Draw a scatter plot diagram of the linear regression.

```
plot(numeric_values_of_total_vaccinations_per_hundred~numeric_values_of_total_number_of_people_vaccinated_per_hundred,xlab="LowerStats",ylab="MedianValue", main="Linear Model Of TVPH vs TNPVPH")
```

```
abline(my_linear_model,col = "red")
```

Predict.

```
predict(my_linear_model, data.frame(numeric_values_of_total_number_of_people_vaccinated_per_hundred = 5))
```