Large Data Programming Challenge

P. polytes team

9/14/2021

For this challenge, we evaluated the correlation between new COVID cases and new tests performed in Mexico.First, we built a scatter plot to watch if there a was some kind of trend between new tests/day and new cases/dat in Mexico. We did see a possible correlation depicted in the plot. Then, we carried out a correlation test between the number of cases detected and the number of tests performed each day. We observed a positive correlation (coefficient:0.6809503 with p-value $<$ 2.2e-16) which is a no brainer during a pandemic, and even more so in a country where the pandemic has been so mishandled and transmission has always been out of control. By increasing the number of tests performed each day, the power to detect more cases is also increased (again: a no brainer!).

```
library(tidyverse)
#read the Dataset sheet into "R". The dataset will be called "data".
data <- read.csv("https://covid.ourworldindata.org/data/owid-covid-data.csv",
     na.strings = "",header=T)


# We subset the data from Mexico.
mex_rows <- which(data$location %in% "Mexico")
data2 <- data[mex_rows,]
data2 <- data2 %>% select("iso_code","date","new_cases", "new_tests")

### Scatter Plot of New Covid cases vs New Covid Tests in Mexico
p <- ggplot(data=data2, aes(x=new_cases,y=new_tests)) +geom_point(color="red",alpha=0.3)+ labs(x="New Ca

print(p)
```

```
## Warning: Removed 59 rows containing missing values (geom_point).
```


```
# We run a Pearson correlation test
pearson_cor <- cor.test(data2$new_cases,data2$new_tests)

pearson_cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$new_cases and data2$new_tests
## t = 22.024, df = 561, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6339903 0.7229070
## sample estimates:
##       cor
## 0.6809503
```