KalaivaniKalyanFinalProjectStep2

Kalaivani Kalyanasundaram

8/8/2021

Why is Vaccination important?

How to import and clean my data

```
library(ggplot2)
library(readxl)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

setwd("/Users/kalya/Documents/Kalai/dsc520")
vaccines_df <- read.csv("data/us_state_vaccinations.csv")
vaccines_ne <- vaccines_df[ which( vaccines_df$location == "Nebraska"), ]
covid_df <- read.csv("data/nytimes/covid-19-data/us-states.csv")
nebraska_df <- covid_df[ which( covid_df$state == "Nebraska"), ]
final_data <- merge(vaccines_ne,nebraska_df)
```

What does the final data set look like?

```
## [1] 279

## [1] 14

##            date location total_vaccinations total_distributed
people_vaccinated
## 7316 2021-01-12 Nebraska              74439            177375
65924
## 7317 2021-01-13 Nebraska              79699            199800
69113
## 7318 2021-01-14 Nebraska              84920            199800
72079
## 7319 2021-01-15 Nebraska              91195            211500
77734
## 7320 2021-01-16 Nebraska                 NA                NA
```

```
NA
## 7321 2021-01-17 Nebraska                 NA                NA
NA
##      people_fully_vaccinated_per_hundred total_vaccinations_per_hundred
## 7316                                0.44                           3.85
## 7317                                0.54                           4.12
## 7318                                  NA                           4.39
## 7319                                0.67                           4.71
## 7320                                  NA                             NA
## 7321                                  NA                             NA
##      people_fully_vaccinated people_vaccinated_per_hundred
## 7316                    8459                          3.41
## 7317                   10529                          3.57
## 7318                      NA                          3.73
## 7319                   13050                          4.02
## 7320                      NA                            NA
## 7321                      NA                            NA
##      distributed_per_hundred daily_vaccinations_raw daily_vaccinations
## 7316                    9.17                     NA                 NA
## 7317                   10.33                 5260.0               5260
## 7318                   10.33                 5221.0               5240
## 7319                   10.93                 6275.0               5585
## 7320                      NA                 3123.5               4970
## 7321                      NA                 3123.5               4601
##      daily_vaccinations_per_million share_doses_used
## 7316                             NA            0.420
## 7317                           2719            0.399
## 7318                           2709            0.425
## 7319                           2887            0.431
## 7320                           2569               NA
## 7321                           2379               NA
```

The final dataset after merging has 279 rows and 14 different columns we can use to answer our question in this study.

Questions for future steps.

What information is not self-evident?

The effect of vaccination is not really self-evident by just looking at the data. We will need a visualization to show it

What are different ways you could look at this data?

I can view the relationship between different variables. Most specifically, I will check out the relationship between covid vaccinations and covid deaths in the similar time frame. This relationship will show us the effect of vaccinations and address our main question.

How do you plan to slice and dice the data?

I plan on selecting the main parts of the data using the dlpyr package and using ggplot to represent the data. I will slice and dice the data as to my convenience to show the relationship. For instance, I will select the deaths and vaccinations columns and plot a grap to represent the data.

How could you summarize your data to answer key questions?

I will use the summary() function to very simply show the data but I will be using graphs as the main summary for my data because that will show readers the actual effect compared with this raw data.

What types of plots and tables will help you to illustrate the findings to your questions?

I will use tables and plots to show my data. Specifically, I will be using histograms, bar charts and tables to precisely show the data. This will help visualise the data and help answer my questions related to vaccinations in the United States.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I do not plan on using any ML techniques yet. I will just be doing a lot of data analysis with this data set and plot various graphs and tables to aid my understanding. In the future however, I can see myself using a ML to analyse the data for me and provide clustering for me to better understand this data.

Questions for future steps.

- Does vaccinations truly help?
- What is the difference of deaths between vaccinated and unvaccinated populations?
- Will proper representation of the data aid my understanding to help solve this question?
- Most importantly, are vaccines important?