Assignment2

Qingyuan Li

2022-09-19

Section 1: Description of Data

The dataset I am using is the 2020 year’s documented daily COVID cases and deaths data collected and published by New York Times. The primary data published are the daily cumulative number of cases and deaths reported in each county and state across the U.S. since the beginning of the pandemic, and I picked only the 2020 csv file. The data is the product of dozens of journalists working across several time zones to monitor news conferences, analyze data releases and seek clarification from public officials on how they categorize cases.This data is very helpful to study the most impactful COVID-19 cases/deaths surges and also the geographic time series of how it spreaded all over the US. The format is csv (a flat file). It is delimited by comma. It is not binary.

Section 2: Reading the Data into R

Section 3: Clean the Data

Section 4: Characteristics of the Data

This dataframe has 884737 rows and 6 columns. The names of the columns and a brief description of each are in the table below:

Table1

| Name of columns | Descriptions |
| --- | --- |
| date | Date of the record was documented |
| county | The county’s name |
| state | The county’s State |
| fips | The FIPS code for the county |
| cases | The accumulative number of cases reported eversince in this county |
| deaths | The accumulative number of deaths reported eversince in this county |

Section 5: Summary Statistics

Table2

## covid.cases covid.deaths covid.fips   
## Min. : 0 Min. : 0.0 Min. : 1001   
## 1st Qu.: 36 1st Qu.: 0.0 1st Qu.:18183   
## Median : 228 Median : 4.0 Median :29215   
## Mean : 1952 Mean : 53.6 Mean :31262   
## 3rd Qu.: 993 3rd Qu.: 21.0 3rd Qu.:46099   
## Max. :770915 Max. :25144.0 Max. :78030   
## NA's :18761 NA's :8266