ANA515Assignment2

Lunhan Zhang

2022-09-18

Section 1: Description of the data.

This data set contains the data a global time series of case and death data for COVID-19. It is sourced from JHU CSSE COVID-19 Data as well as The New York Times. From this data, we will look the trend of death and test-positive and also the comparation by state in US. This is data is excel csv file.

##Section 2: Reading the data into R.

```
data <- read.csv("COVID-19 Activity.csv")
```

##Section 3: Clean the data.

```
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
```

```
smallerdata <- data %>%
select(PEOPLE_POSITIVE_CASES_COUNT, CONTINENT_NAME, PROVINCE_STATE_NAME, REPORT_DATE, PEOPLE_DEATH_NEW_C
filter(CONTINENT_NAME=='America') %>%
rename("COUNTRY_NAME"="CONTINENT_NAME")%>%
filter(REPORT_DATE>= ('2022-01-01'))
```

##Section 4: Characteristics of the data.

This dataframe has 397817 rows and 7 columns. The names of the columns and a brief description of each are in the table below:

```
#this makes a new data.frame called text_tbl with two columns, Names and Description
text_tbl <- data.frame(
Names = c("PROVINCE_STATE_NAME", "PEOPLE_POSITIVE_NEW_CASES_COUNT", "PEOPLE_DEATH_NEW_COUNT"),
Description = c("State name for the data", "Number of people were tested positive each day", "Number of
)
text_tbl #prints the table
```

```
##                             Names
## 1             PROVINCE_STATE_NAME
## 2 PEOPLE_POSITIVE_NEW_CASES_COUNT
## 3          PEOPLE_DEATH_NEW_COUNT
##                                   Description
```

```
## 1                         State name for the data
## 2 Number of people were tested positive each day
## 3                Number of people death each day
```

##Section 5: Summary statistics.

```
data_pick3 <- select(smallerdata, PROVINCE_STATE_NAME, PEOPLE_DEATH_NEW_COUNT, PEOPLE_POSITIVE_NEW_CASES
```

```
Summarytable<-summary(data_pick3) #creates the summary
Summarytable #prints the summary in your output
```

```
##  PROVINCE_STATE_NAME PEOPLE_DEATH_NEW_COUNT PEOPLE_POSITIVE_NEW_CASES_COUNT
##  Length:397817       Min.   :-1330.00       Min.   :-370251.0
##  Class :character    1st Qu.:    0.00       1st Qu.:      0.0
##  Mode  :character    Median :    0.00       Median :      1.0
##                      Mean   :    0.78       Mean   :    121.8
##                      3rd Qu.:    0.00       3rd Qu.:     17.0
##                      Max.   :11447.00       Max.   : 287149.0
```