HW01

Michael Puchalski

2025-01-25

```
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
```

```
df<-read.csv("UScovid.csv")
dim(df)
```

```
## [1] 1384683       6
```

```
colnames(df)
```

```
## [1] "date"   "county" "state"  "fips"   "cases"  "deaths"
```

```
class(df$date)
```

```
## [1] "character"
```

```
str(date)
```

```
## function ()
```

```
typeof(df$date)
```

```
## [1] "character"
```

##Question 1a (cleaning up the data) We are interested in the data at the most recent date, June 3 2021. Create a data frame called latest that: - has only rows pertaining to data from June 3 2021, - removes rows pertaining to counties that are "Unknown", - removes the column date and fips, - is ordered by county and then state alphabetically Display the first 6 rows of the data using head().

```
latest<-df%>%
  mutate(date = as.Date(date)) %>%
  filter(date == as.Date('2021-06-03'))%>% # Filters the data frame to only include the selected date.
  filter(!is.na(county) & county!= "Unknown")%>% # This removes rows with the null value from the speci
  select(-date, -fips)%>% # Using '-' sign allows us to remove the columns that are selected
  arrange(county,state) # This allows us to sort alphabetically
head(latest, n = 6)
```

```
##      county          state cases deaths
## 1 Abbeville South Carolina  2599     41
## 2    Acadia      Louisiana  6703    195
## 3  Accomack       Virginia  2862     43
## 4       Ada          Idaho 52964    475
## 5     Adair           Iowa   873     32
## 6     Adair       Kentucky  1944     54
```

Question 1b

Calculate the case fatality rate (number of deaths divided by number of cases, and call it death.rate) for each county. Report the case fatality rate as a percent and round to two decimal places. Add death.rate as a new column to the data frame latest. Display the first 6 rows of the data frame latest.

```
latest<-latest%>%
  mutate(death.rate = round((deaths/cases) * 100, 2))
# round function allows us to define after the comma what decimal we want to round to
head(latest, n = 6)
```

```
##      county          state cases deaths death.rate
## 1 Abbeville South Carolina  2599     41       1.58
## 2    Acadia      Louisiana  6703    195       2.91
## 3  Accomack       Virginia  2862     43       1.50
## 4       Ada          Idaho 52964    475       0.90
## 5     Adair           Iowa   873     32       3.67
## 6     Adair       Kentucky  1944     54       2.78
```

Question 1c

Display the counties with the 10 largest number of cases. Be sure to also display the number of deaths and case fatality rates in these counties, as well as the state the counties belong to.

```
top_ten_cases<-latest %>%
  arrange(desc(cases))
head(top_ten_cases, n = 10)
```

```
##           county      state   cases deaths death.rate
## 1    Los Angeles California 1245127  24375       1.96
## 2   New York City   New York  949986  33257       3.50
## 3           Cook   Illinois  554390  10893       1.96
## 4       Maricopa    Arizona  551509  10084       1.83
## 5     Miami-Dade    Florida  501925   6472       1.29
```

```
## 6          Harris      Texas  401345   6462       1.61
## 7          Dallas      Texas  303533   4082       1.34
## 8       Riverside California  300879   4614       1.53
## 9  San Bernardino California  298599   4760       1.59
## 10      San Diego California  280410   3760       1.34
```

##Question 1d Display the counties with the 10 largest number of deaths. Be sure to also display the number of cases and case fatality rates in these counties, as well as the state the counties belong to

```
top_ten_deaths<-latest%>%
  arrange(desc(deaths))
head(top_ten_deaths, n = 10)
```

```
##            county      state   cases deaths death.rate
## 1   New York City   New York  949986  33257       3.50
## 2     Los Angeles California 1245127  24375       1.96
## 3            Cook   Illinois  554390  10893       1.96
## 4        Maricopa    Arizona  551509  10084       1.83
## 5      Miami-Dade    Florida  501925   6472       1.29
## 6          Harris      Texas  401345   6462       1.61
## 7          Orange California  272242   5070       1.86
## 8           Wayne   Michigan  164612   5048       3.07
## 9  San Bernardino California  298599   4760       1.59
## 10      Riverside California  300879   4614       1.53
```

##Question 1e Display the counties with the 10 highest case fatality rates. Be sure to also display the number of cases and deaths in these counties, as well as the state the counties belong to. Is there something you notice about these counties?

```
top_ten_death_rate<-latest %>%
  arrange(desc(death.rate))
head(top_ten_death_rate, n = 10)
```

```
##          county      state cases deaths death.rate
## 1         Grant   Nebraska    41      4       9.76
## 2        Sabine      Texas   524     45       8.59
## 3       Harding New Mexico    12      1       8.33
## 4     Petroleum    Montana    12      1       8.33
## 5         Foard      Texas   124     10       8.06
## 6       Hancock    Georgia   928     68       7.33
## 7      Glascock    Georgia   269     19       7.06
## 8        Motley      Texas   116      8       6.90
## 9       Candler    Georgia   978     67       6.85
## 10 Throckmorton      Texas    73      5       6.85
```

What is noticeable about the counties is that they all seem to be rural counties where the overall incedence of infection is lower, but that also means that that an instance of death is going to have a greater impact.

##Question 1f Display the counties with the 10 highest case fatality rates among counties with at least 100,000 cases. Be sure to also display the number of cases and deaths in these counties, as well as the state the counties belong to.

```
larger_set_top_ten<-latest%>%
  filter(cases>= 100000)%>%
  arrange(desc(death.rate))
head(larger_set_top_ten, n = 10)
```

```
##           county         state  cases deaths death.rate
## 1  New York City      New York 949986  33257       3.50
## 2          Wayne      Michigan 164612   5048       3.07
## 3      Middlesex Massachusetts 134980   3761       2.79
## 4         Bergen    New Jersey 104301   2868       2.75
## 5         Macomb      Michigan 100190   2441       2.44
## 6   Philadelphia  Pennsylvania 153521   3692       2.40
## 7      St. Louis      Missouri 100195   2249       2.24
## 8      Fairfield   Connecticut 100093   2198       2.20
## 9           Pima       Arizona 116997   2406       2.06
## 10       Oakland      Michigan 118035   2368       2.01
```

##Question 1g Display the number of cases, deaths, case fatality rates for the following counties: i. Albemarle, Virginia ii. Charlottesville city, Virginia

```
local_rate<-latest %>%
  filter(state == "Virginia")%>%
  filter(county == "Albemarle" | county == "Charlottesville city")
local_rate
```

```
##                 county    state cases deaths death.rate
## 1            Albemarle Virginia  5801     83       1.43
## 2 Charlottesville city Virginia  4014     57       1.42
```

#Question 2 For this question, we focus on data at the state level. Note that the dataset has data on the 50 states, plus DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands. For the purpose of this question, we will consider DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands, as "states" as well. ##Question 2a We are interested in the data at the most recent date, June 3 2021. Create a data frame called state.level that: - has 55 rows: 1 for each state, DC, and territory - has 3 columns: name of the state, number of cases, number of deaths - is ordered alphabetically by name of the state Display the first 6 rows of the data frame state.level.

```
state.level<-df%>%
  group_by(state)%>%
  summarize(
    cases = sum(cases, na.rm = TRUE),
    deaths = sum(deaths, na.rm = TRUE) # inserted in na.rm = TRUE because without it, Puerto Rico has n
  ) %>%
  arrange(state)
head(state.level, n = 6)
```

```
## # A tibble: 6 x 3
##   state          cases   deaths
##   <chr>          <int>    <int>
## 1 Alabama    108217598  1986203
## 2 Alaska      12176769    57974
```

```
## 3 Arizona    170946356  3429632
## 4 Arkansas    67032054  1101842
## 5 California 714568374 11112159
## 6 Colorado    93331929  1477994
```

##Question 2b Calculate and add the state.rate case fatality rate

```
state.level<-state.level%>%
  mutate(state.rate = round((deaths/cases) * 100, 2))
head(state.level, n = 6)
```

```
## # A tibble: 6 x 4
##   state          cases   deaths state.rate
##   <chr>          <int>    <int>      <dbl>
## 1 Alabama    108217598  1986203       1.84
## 2 Alaska      12176769    57974       0.48
## 3 Arizona    170946356  3429632       2.01
## 4 Arkansas    67032054  1101842       1.64
## 5 California 714568374 11112159       1.56
## 6 Colorado    93331929  1477994       1.58
```

##Question 2c and 2d What is the fatility rate in Virginia and Puerto Rico?

```
filter(state.level, state == "Virginia" | state == "Puerto Rico")
```

```
## # A tibble: 2 x 4
##   state           cases  deaths state.rate
##   <chr>           <int>   <int>      <dbl>
## 1 Puerto Rico  31720631  453629       1.43
## 2 Virginia    122074227 2044362       1.67
```

According to my output, the death rate for Virginia is 1.67 and 1.43 for Puerto Rico. This was only attainable after going back and making the summary statistics ignore NA values during summation.

##Question 2e Which states have the 10 highest case fatality rates?

```
top_fatality<-state.level%>%
  arrange(desc(state.rate))
head(top_fatality, n = 10)
```

```
## # A tibble: 10 x 4
##    state                   cases   deaths state.rate
##    <chr>                   <int>    <int>      <dbl>
##  1 New York            391662873 15657152       4
##  2 New Jersey          186855114  7453154       3.99
##  3 Connecticut          62301738  2290608       3.68
##  4 Massachusetts       129843236  4719141       3.63
##  5 District of Columbia  10123274   302969       2.99
##  6 Pennsylvania        199718611  5686194       2.85
##  7 Michigan            157373540  4466740       2.84
##  8 Louisiana           101013970  2670041       2.64
##  9 Mississippi          65526776  1595828       2.44
## 10 Maryland             90183940  2140500       2.37
```

In order: New York, New Jersey, Conneticut, DC, Pennsylvania, Michigan, Louisiana, Mississippi, Maryland
##Question 2f Which states have the 10 lowest fatality rates?

```
low_fatality<-state.level%>%
  arrange(state.rate)
head(low_fatality, n = 10)
```

```
## # A tibble: 10 x 4
##    state              cases  deaths state.rate
##    <chr>              <int>   <int>      <dbl>
##  1 Alaska          12176769   57974       0.48
##  2 Utah            77145081  412498       0.53
##  3 Nebraska        44705960  471247       1.05
##  4 Idaho           37540727  399664       1.06
##  5 Wyoming         11041401  121919       1.1
##  6 Wisconsin      134079973 1489246       1.11
##  7 Virgin Islands    644217    7419       1.15
##  8 Oklahoma        84840984 1012422       1.19
##  9 Montana         20625079  273504       1.33
## 10 Kentucky        81490302 1108570       1.36
```

In order: Alaska, Utah, Nebraska, Idaho, Wyoming, Wisconsin, Virgin Islands, Oklahoma, Montana, Kentucky

##Question 2g Export this dataset as a .csv file named stateCovid.csv. We will be using this file for the next homework.

```
write.csv(state.level, "stateCovid.csv", row.names = FALSE)
```