Project One

Data & Text Mining

Brooke O'Donnell

9.18.20

Project1.od.html

1. Create a data frame in R to include only information on the state, year, week, week ending day, deaths of all causes, and the variable "COVID-19 (U071, Multiple Cause of Death).

```
CD2<-CD%>% select(Jurisdiction.of.Occurrence, MMWR.Year, MMWR.Week, Week.Ending.Date,
All.Cause, COVID.19..U071..Multiple.Cause.of.Death.)
```

2. Rename variables and recode some variables as you see fit.

```
CD2<-CD2%>% rename(State.location=Jurisdiction.of.Occurrence, YEAR=MMWR.Year,
WEEK=MMWR.Week, MCOD=COVID.19..U071..Multiple.Cause.of.Death.)
head(CD2)
```

```
##   State.location YEAR WEEK Week.Ending.Date All.Cause MCOD
## 1        Alabama 2019    1       1/5/2019     1077    0
## 2        Alabama 2019    2      1/12/2019     1090    0
## 3        Alabama 2019    3      1/19/2019     1114    0
## 4        Alabama 2019    4      1/26/2019     1063    0
## 5        Alabama 2019    5       2/2/2019     1095    0
## 6        Alabama 2019    6       2/9/2019     1092    0
```

3. Convert the "week ending date" variable into a "Date" data type. Check if you have succeeded in the conversion. (1 points) (Hint: the format you need to indicate as an argument in one of the variations of "mm-dd-yyyy" is not the format you intend to have but the format in the original data).

```
CD2$Week.Ending.Date <- as.Date(CD2$Week.Ending.Date,
 format = "%m/%d/%y")
```

4a. Create a subset that contains only information of 2020 and includes only variables of the state, the week, and "covid-19 Multiple Cause of Death."

```
CD.20<-CD2%>%select(State.location, YEAR, WEEK, MCOD)

CD.20<-CD.20 %>% filter(YEAR=="2020")
CD.2020<-select(CD.20, State.location, WEEK, MCOD)
summary(CD.2020)
```

```
##    State.location     WEEK          MCOD
##  Alabama  : 33   Min.   : 1.00   Min.   :   0.00
##  Alaska   : 33   1st Qu.: 9.00   1st Qu.:   0.00
##  Arizona  : 33   Median :17.00   Median :  22.00
##  Arkansas : 33   Mean   :16.98   Mean   : 214.10
##  California: 33   3rd Qu.:25.00   3rd Qu.:  92.25
```

```
##  Colorado  : 33   Max.   :33.00   Max.   :17036.00
##  (Other)   :1582                  NA's   :252
```

b. Then take a further step to only select the states (city) of Arizona, California, Florida, New Jersey, New York, New York City, and Texas. (NYC and DC are included as non-state jurisdiction)

```
CD.7<-CD.2020%>% filter(State.location=="Arizona"| State.location=="California"|
State.location=="Florida"|State.location=="New Jersey"| State.location=="New York"|
State.location=="New York City"| State.location=="Texas")
summary(CD.7)
```

```
##       State.location     WEEK        MCOD
##  Arizona     :33    Min.   : 1   Min.   :   0.0
##  California  :33    1st Qu.: 9   1st Qu.:   0.0
##  Florida     :33    Median :17   Median : 130.0
##  New Jersey  :33    Mean   :17   Mean   : 377.7
##  New York    :33    3rd Qu.:25   3rd Qu.: 442.0
##  New York City:33    Max.   :33   Max.   :4848.0
##  (Other)     :33                 NA's   :14
```

c. As a last step, "spread" the data frame by week (making "week" into columns) on the values of the variable "covid-19 Multiple Cause of Death".

```
spread(CD.7, key="WEEK", value="MCOD")
```

```
##   State.location 1 2  3 4 5  6 7  8  9 10 11  12  13   14   15   16   17
## 1       Arizona 0 0  0 0 0  0 0  0  0  0  0  NA  31   58   91   96  106
## 2    California 0 0 NA 0 0 NA 0 NA  0 NA NA  44 130  289  381  533  526
## 3       Florida 0 0  0 0 0  0 0  0  0 NA NA  17  92  217  302  309  321
## 4    New Jersey 0 0  0 0 0  0 0  0  0  0 NA  36 284 1232 2346 2480 2031
## 5      New York 0 0  0 0 0  0 0 NA  0  0 NA  41 292 1120 2002 2002 1590
## 6  New York City 0 0  0 0 0  0 0  0  0  0 NA 124 979 3305 4848 3909 2620
## 7        Texas 0 0  0 0 0  0 0  0 NA NA  0  10  44  127  205  236  242
##    18  19  20  21  22  23  24  25  26  27   28   29   30   31  32  33
## 1 104  127 117  95 113 132 170 203 252 365  478  534  454  376 266 162
## 2 542  550 499 484 457 442 441 451 481 613  705  847  948  901 798 409
## 3 306  289 257 232 230 217 192 217 312 440  738 1091 1222 1160 913 465
## 4 1547 1114 820 549 394 308 219 160 138  99   85   59   63   58  35  31
## 5 1134  877 669 449 328 262 180 126  74  61   64   58   42   38  30  37
## 6 1723 1050 615 413 280 196 139  88  71  49   40   50   44   32  19  24
## 7  235  227 249 170 214 179 223 296 427 776 1192 1575 1609 1262 726 224
```

d. At which week did covid-19 deaths started to occur in most of the 6 states and NYC?

COVID-19 cases started to occur in most of the 6 states and NYC around week 12/13.

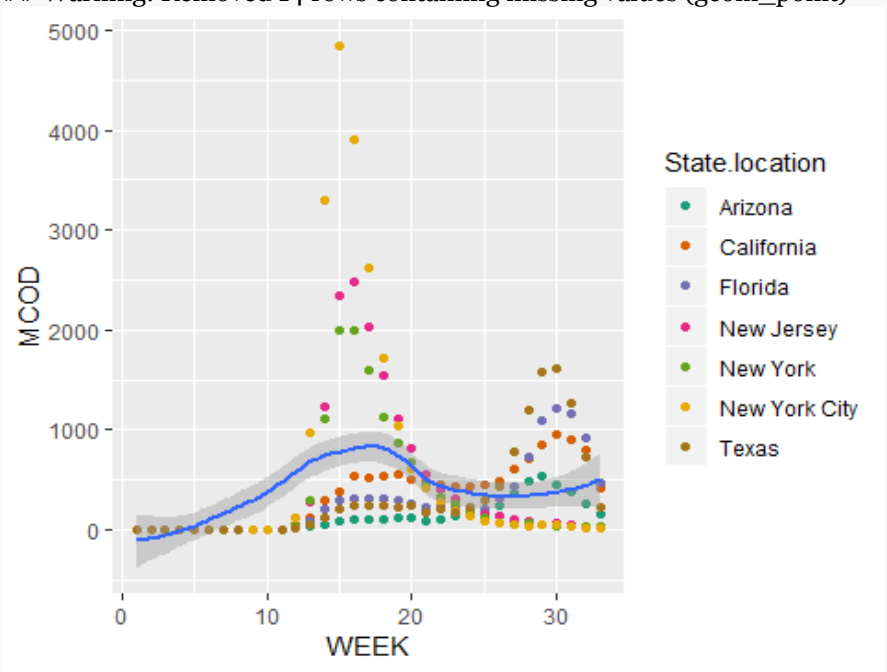
5a. Use the data frame from step 4 (the one that contains only 2020 information) to plot the number of covid-19 deaths (Multiple Cause of Death) by the week; and add a dimension of "state" by color.

```
ggplot(data= CD.7, mapping=aes(x=WEEK,
y=MCOD))+geom_point(mapping=aes(color=State.location))+geom_smooth()+scale_color_bre
wer(palette="Dark2")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 14 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 14 rows containing missing values (geom_point)
```



b. Which state (or city) saw deaths peak the earliest? At which week?

Deaths due to the COVID-19 pandemic started to spike in the 13/14th week in New York City, New York State, and New Jersey.

c.What state(s) (or city) saw a second wave of death? During which weeks?

Texas saw the highest second wave starting at about week 27/28 hitting its peak around week 30 then started to decline. Florida California and Arizona also all saw a second wave.

6a. (Use the data frame from step 5) Use the "summarise" function in "dplyr" to create a data set that contains only "the total number of death of covid (with multiple causes)" and the state (city) name. Then "merge" the new data frame with "population_by_state_2019.csv" (available on Canvas) to combine the death toll with information on state population. You should end up with a data frame containing 3 variables: state, total covid death, and state population.

```
POP19<-read.csv("population.state19.csv", header=TRUE, sep=",", fileEncoding="UTF-8-BOM")

POP19<-POP19%>%rename(State.location="state")

sum(is.na(CD.7))

## [1] 14

CD.7<-CD.7%>%drop_na()

table1<-CD.7%>%
group_by(State.location) %>%
summarise(statetotal = sum(MCOD, na.rm = TRUE))

POP19<-POP19%>% filter(State.location=="Arizona"| State.location=="California"|
State.location=="Florida"|State.location=="New Jersey"| State.location=="New York"|
State.location=="New York City"| State.location=="Texas")
```

```
table2<-merge(table1, POP19, by.x=c("State.location"))
summary(table2)

##       State.location   statetotal      pop2019
##  Arizona     :1     Min.   : 4330   Min.   : 7278717
##  California  :1     1st Qu.: 9994   1st Qu.: 8609504
##  Florida     :1     Median :11471   Median :19453561
##  New Jersey  :1     Mean   :11710   Mean   :19133875
##  New York    :1     3rd Qu.:12782   3rd Qu.:25236809
##  New York City:1    Max.   :20618   Max.   :39512223
##  (Other)     :1
```

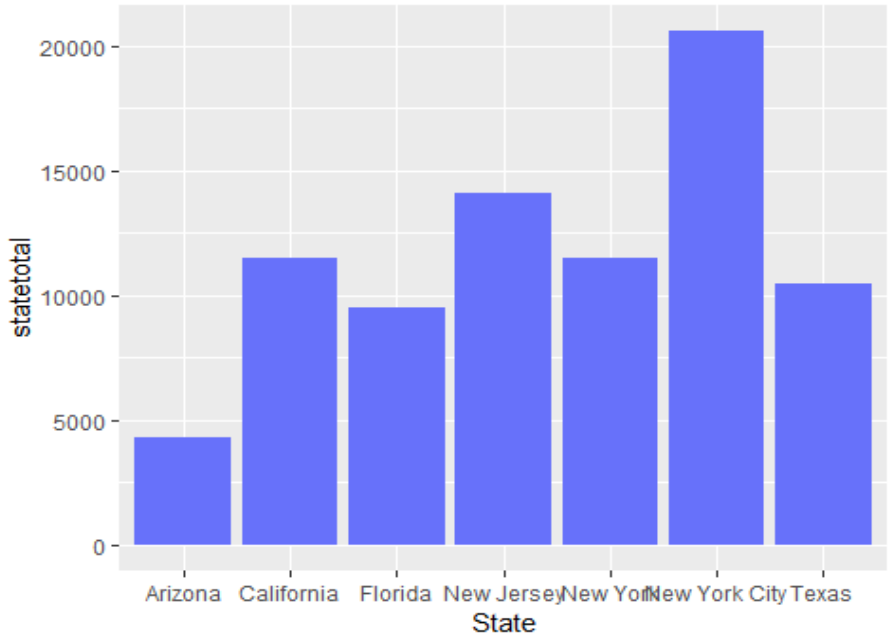
b. Obtain a bar-chart to see which state has the highest covid death toll.

```
table2<-table2%>%rename(., State=State.location)

table2<-as.data.frame(table2)
```

New York City has the highest COVID-19 death toll. New Jersey is the state with the highest COVID-19 death toll.

```
ggplot(data=table2)+geom_bar(mapping=aes(x=State, y=statetotal), stat="identity",fill="#6771FA")
```



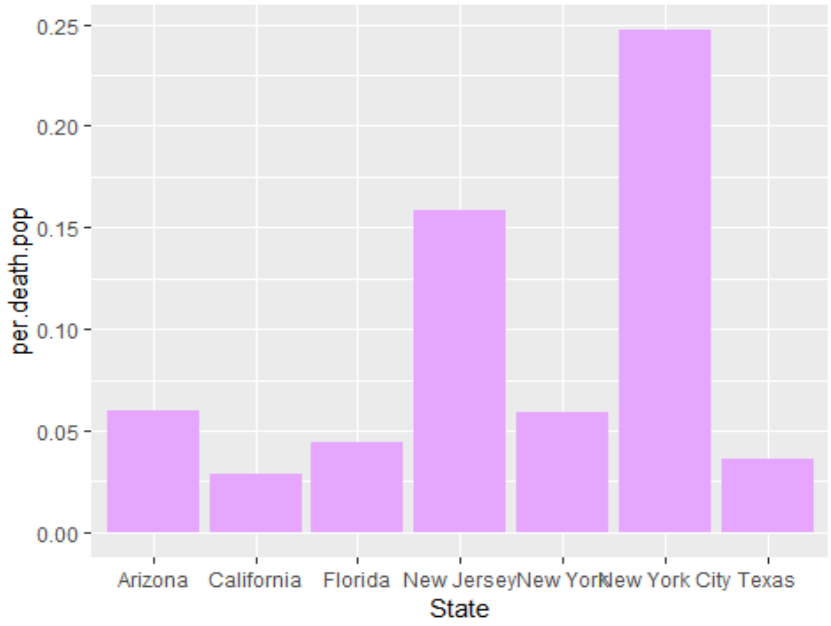
c. Calculate a new variable of "percentage of covid death in the total population" and save it to the new (small) data frame (Hint: you can use either the mutate function or simple base R function).

```
table2$per.death.pop<-table2$statetotal/table2$pop2019*100
summary(table2)

##       State     statetotal      pop2019       per.death.pop
##  Arizona     :1   Min.   : 4330   Min.   : 7278717   Min.   :0.02903
##  California  :1   1st Qu.: 9994   1st Qu.: 8609504   1st Qu.:0.04022
##  Florida     :1   Median :11471   Median :19453561   Median :0.05899
##  New Jersey  :1   Mean   :11710   Mean   :19133875   Mean   :0.09055
##  New York    :1   3rd Qu.:12782   3rd Qu.:25236809   3rd Qu.:0.10905
##  New York City:1   Max.   :20618   Max.   :39512223   Max.   :0.24731
##  (Other)     :1
```

d. Please plot the percentage you obtain in the previous step by the state. Which state has the highest percentage of covid death in the total population?

```
ggplot(data=table2)+geom_bar(mapping=aes(x=State, y=per.death.pop),
stat="identity",fill="#E6A6FD")
```



New York City has the highest percentage of COVID-19 deaths in the total population. A researcher can say that almost .25% of deaths in the total population are due to COVID-19 in NYC.

7. a. Formulate a data frame for 2020 that contains only 1. the state, 2. covid death (Multiple Cause of Death), 3. and deaths of all causes (basically the number of all people who died in the first 33 weeks in 2020)

```
CD.2020<-CD2%>%select(State.location, YEAR, WEEK, MCOD, All.Cause)

CD.2020<-CD.2020 %>% filter(YEAR=="2020")

STATE.MCOD.ALL<-CD.2020%>%select(State.location, MCOD, All.Cause)
head(STATE.MCOD.ALL)

##   State.location MCOD All.Cause
## 1        Alabama    0      1077
## 2        Alabama    0      1124
## 3        Alabama    0      1038
## 4        Alabama    0      1053
## 5        Alabama    0      1023
## 6        Alabama    0      1115
```

b. Then use the "summarise" function to create a smaller data frame with aggregated information that contains 1. total covid deaths and 2. total deaths of all causes

```
small.df<-STATE.MCOD.ALL %>% group_by(State.location) %>% summarise(allcausetotal =
sum(All.Cause, na.rm = TRUE), COV.total = sum(MCOD, na.rm = TRUE))
summary(small.df)
```

```
##    State.location allcausetotal      COV.total
##  Alabama  : 1    Min.   :  2719   Min.   :    0.0
##  Alaska   : 1    1st Qu.:  11495   1st Qu.:   366.2
##  Arizona  : 1    Median :  29445   Median :  1389.5
##  Arkansas : 1    Mean   :  74473   Mean   :  6058.3
##  California: 1    3rd Qu.:  49878   3rd Qu.:  4080.0
##  Colorado : 1    Max.   :2002338   Max.   :163902.0
##  (Other)  :48
```

c. Next, reduce the data frame to include only the 6 states above and NYC, as we did in the previous questions.

```
small.df<-small.df%>% filter(State.location=="Arizona"| State.location=="California"|
State.location=="Florida"|State.location=="New Jersey"| State.location=="New York"|
State.location=="New York City"| State.location=="Texas")
summary(small.df)
```

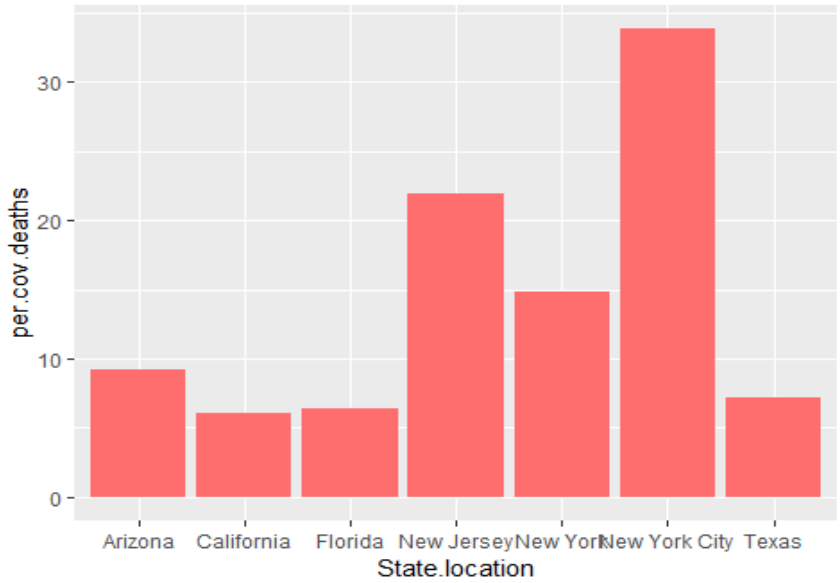
```
##       State.location allcausetotal      COV.total
##  Arizona      :1     Min.   : 47097   Min.   : 4330
##  California   :1     1st Qu.: 62561   1st Qu.: 9994
##  Florida      :1     Median : 77241   Median :11471
##  New Jersey   :1     Mean   :104559   Mean   :11710
##  New York     :1     3rd Qu.:146884   3rd Qu.:12782
##  New York City:1     Max.   :188687   Max.   :20618
##  (Other)      :1
```

d. Then create a variable: the percentage of covid deaths in total deaths; plot this percentage by state. Which state (city) has the highest percentage of covid deaths of all deaths?

```
small.df$per.cov.deaths<-small.df$COV.total/small.df$allcausetotal*100
summary(small.df)
```

```
##       State.location allcausetotal      COV.total     per.cov.deaths
##  Arizona      :1     Min.   : 47097   Min.   : 4330   Min.   : 6.079
##  California   :1     1st Qu.: 62561   1st Qu.: 9994   1st Qu.: 6.810
##  Florida      :1     Median : 77241   Median :11471   Median : 9.194
##  New Jersey   :1     Mean   :104559   Mean   :11710   Mean   :14.220
##  New York     :1     3rd Qu.:146884   3rd Qu.:12782   3rd Qu.:18.398
##  New York City:1     Max.   :188687   Max.   :20618   Max.   :33.853
##  (Other)      :1
```

```
ggplot(data=small.df)+geom_bar(mapping=aes(x=State.location, y=per.cov.deaths),
stat="identity",fill="#FF6E6E")
```



New York City has the highest percentage of COVID-19 deaths of all deaths. A researcher could state that out of all deaths in NYC COVID-19 accounted for about 34% of deaths.