COVID-19 Data Analysis

Question of Interest

In each state in the US there are prisons and they are located in specific counties in these states. In this analysis I want to look at the COVID-19 data in counties present in the state of Maine and compare them to COVID data in the counties where prisons are located. I am choosing Maine particularly because I have been to Maine previously and it is a very beautiful state and liked the trip very much. In brief I want to compare how the cases have manifested in the prisons compared to the outside world. I want to look if there is any Correlation between the rise in cases in the outside world and compare it to the cases inside prisons.I also want to compare the deaths in the outside world to the deaths inside the prisons.I found the prison data set to be interesting as I never thought of the fact that how COVID 19 spreads inside prisons. When we hear about COVID 19 we always assume about the cases and deaths in general public but not in seculed and gaurded places like a prison.

Source and Description of Data

The source of the data both the prison data set as well as the counties data set is from New York Times/Covid-19 data and I found the data set in github.I downloaded the datasets inot my pc. From March 2020 until the end of March 2021, The New York Times collected data about coronavirus infections, deaths and testing for state and federal prisons; immigration detention centers; juvenile detention facilities; local, regional and reservation jails; and those in the custody of the U.S. Marshals Service.The Times gathered information about infections, deaths, facility populations and tests administered to inmates and correctional officers for 2,805 facilities, and system wide totals for the same data.This population represented a particularly vulnerable group of people who were at far higher risk of coronavirus infection than members of the general public. We can use the prison data to look at the impact of the virus in prisons. The counties data set in general contains all the information related to Covid cases/deaths regarding the general public for the entire country of United States.

My Aim

I want to create two visualization graphs which shows the Covid confirmed cases and the deaths both in the general public and in prisons. A prison is already a pretty locked down area and has high security and most of the stuff bought inside the prison is thoroughly examined and let inside. I know that the prison numbers will be low as it is very secluded and many the inmates are locked separately. Even if an inmate gets COVID he/she is locked in the cell and the chance of spreading the virus is very low. If feel that prisoners are safe from the virus as they are mostly in a locked down state with minimal contact with each other. Before even I start my analysis I would assume that the deaths wont be at a high number in prisons as there is a prison doctor at the site always and I feel inmates who have contracted the virus would be quarantined in a cell anyway. The only deaths I see happening are of the prisoners and officers who already some sort of comorbidities. This is just my assumption before diving into the dataset.I cannot club both the plots into one graph as the prison cases are very low compared to the real world cases. We cannot judge properly as the real world cases are in a much higher scale compared to the prison cases. The prison cases are near the 2 digit mark in most cases and the real world cases are in 5 figure numbers so it makes no sense in putting them in the same plot.I will put the 2 plots side by side and try to compare the data.For the total confirmed cases and deaths in prisons I have taken into account both the inmate confirmed cases as well as the officer confirmed cases.

Part 1 :-Lets take a look at the confirmed cases in the general public and the prisons

```
##
## 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

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
prison <- prisons %>% group_by(`facility_state`) %>% filter(facility_state=="Maine")%>% select(facility
prison2 <- prison %>% group_by(facility_county)%>% summarize(total_inmate_cases = sum(total_inmate_case
prison2
```

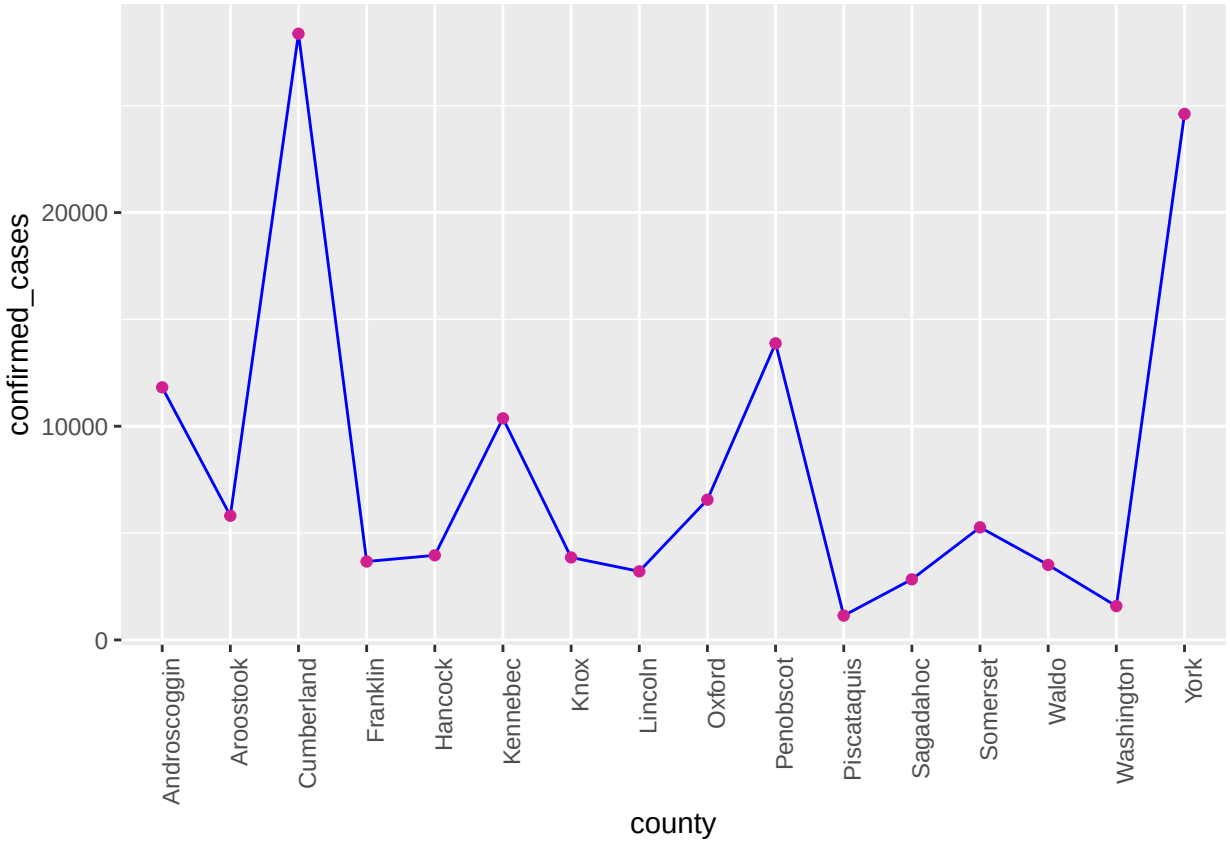
```
## # A tibble: 6 x 2
##   facility_county confirmed_cases
##   <chr>                     <int>
## 1 Cumberland                  165
## 2 Kennebec                     30
## 3 Knox                         23
## 4 Penobscot                    40
## 5 Washington                    0
## 6 York                         73
```

```
plot1 <- ggplot(data = prison2,mapping=aes(x=facility_county, y=confirmed_cases))+
  geom_line(aes(group=1),color="darkgreen") + geom_point(color="violetred")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
county <- counties %>% group_by(state) %>% filter(state =="Maine") %>% select(state,county,confirmed_cas
county1 <- county %>% group_by(county)%>% summarize(confirmed_cases = sum(confirmed_cases))
county_1 <- county1[c(1,2,3,4,5,6,7,8,9,10,11,12,13,15,16,17), ,drop=TRUE]
county_1
```

```
## # A tibble: 16 x 2
##    county       confirmed_cases
##    <chr>                  <int>
##  1 Androscoggin           11821
##  2 Aroostook               5812
##  3 Cumberland             28369
##  4 Franklin                3670
##  5 Hancock                 3960
##  6 Kennebec               10372
##  7 Knox                    3861
##  8 Lincoln                 3210
##  9 Oxford                  6563
## 10 Penobscot              13879
## 11 Piscataquis             1140
## 12 Sagadahoc               2836
## 13 Somerset                5269
## 14 Waldo                   3511
## 15 Washington              1583
## 16 York                   24615
```

```
ggplot(data = county_1,mapping=aes(x=county, y=confirmed_cases))+
  geom_line(aes(group=1),color="blue") + geom_point(color="violetred")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

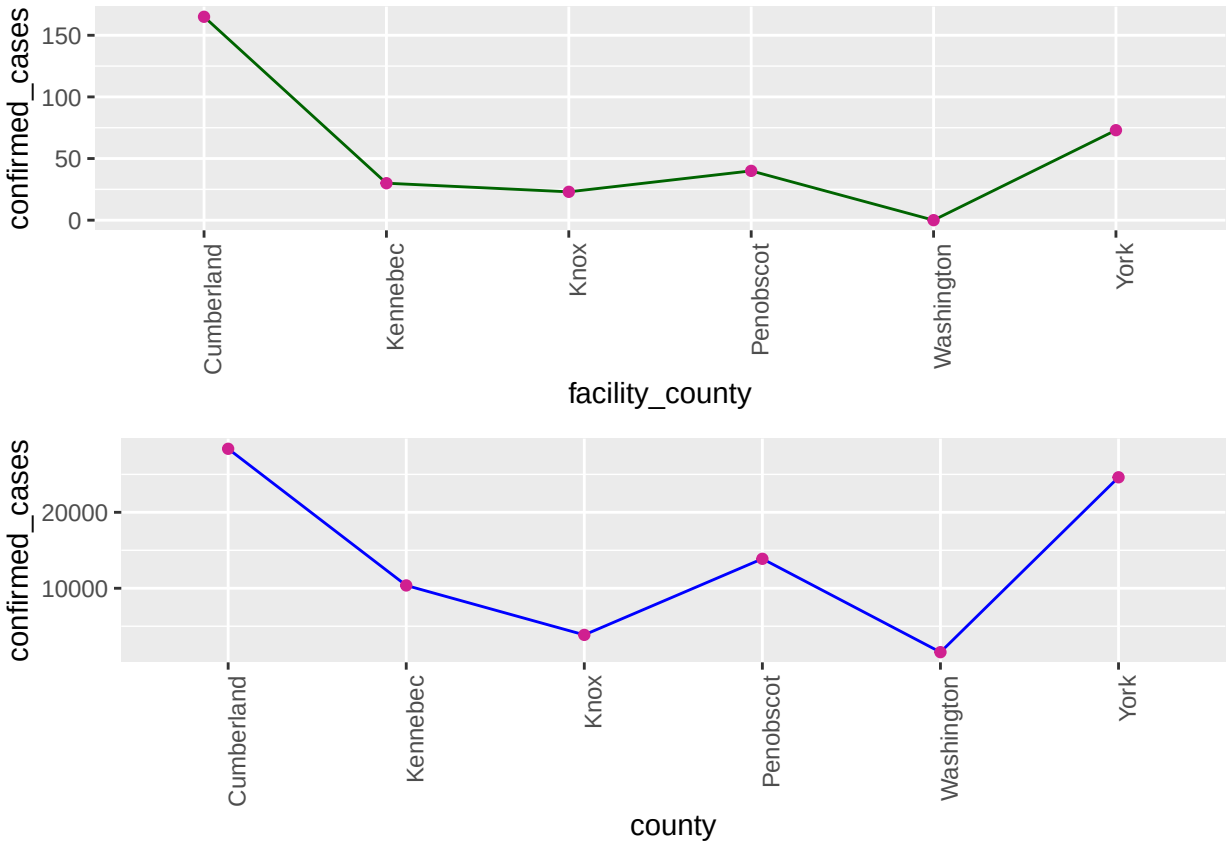


The above graph shows the cases for all the counties in the state of maine but we want the cases for only those counties which have prisons them. Hence I have filtered out only those counties which have prisons in them and I compared that with the prison data set

```
county2 <- county1[c(3,6,7,10,16,17), , drop=TRUE]
plot2 <- ggplot(data = county2,mapping=aes(x=county, y=confirmed_cases))+
  geom_line(aes(group=1),color="blue") + geom_point(color="violetred")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
county2
```

```
## # A tibble: 6 x 2
##   county     confirmed_cases
##   <chr>                <int>
## 1 Cumberland           28369
## 2 Kennebec             10372
## 3 Knox                  3861
## 4 Penobscot            13879
## 5 Washington            1583
## 6 York                 24615
```

```
grid.arrange(plot1, plot2)
```



Summary for Part 1

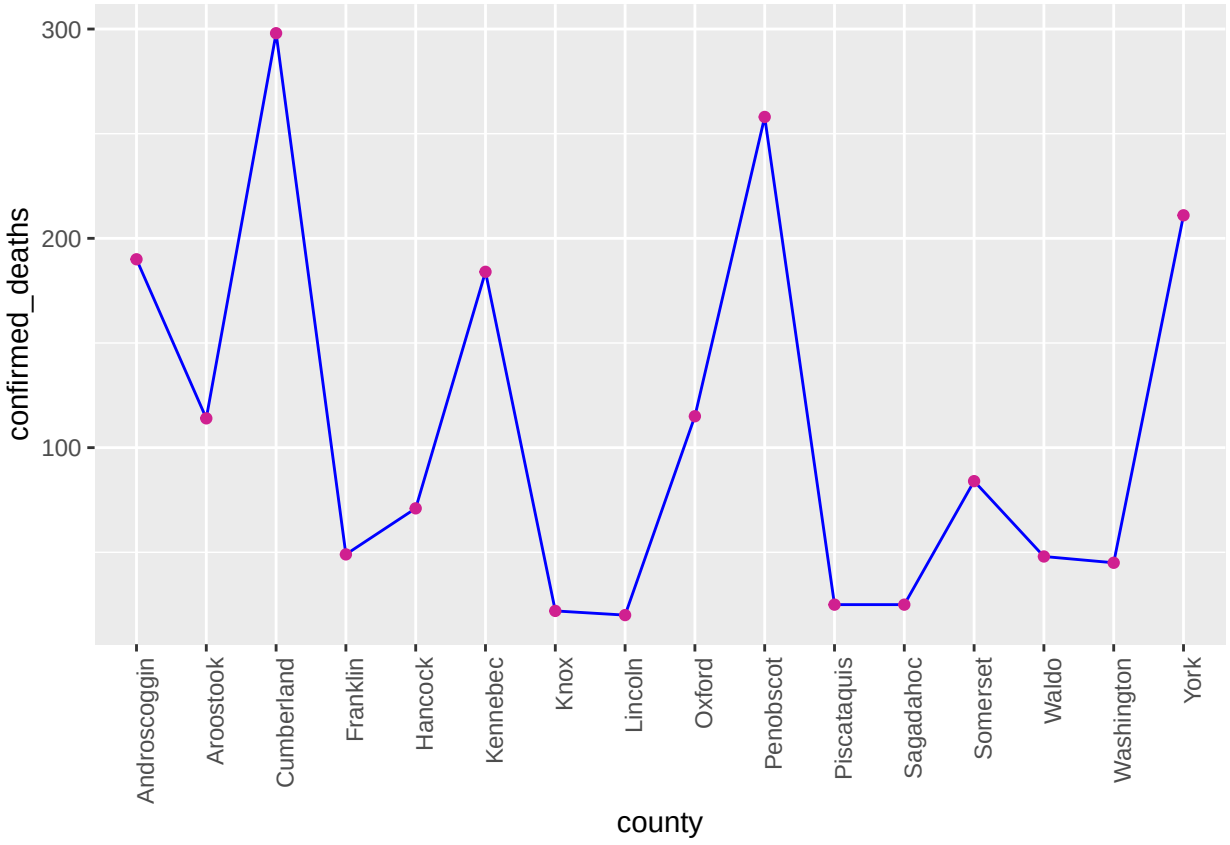
Looking at the two graphs we can see that there is a pattern in the data set. In Maine there are only 6 counties which have prisons in them. Now if we take the confirmed cases of the prisons from those 6 counties and compare them to the cases of general public in those counties we can see that the counties which have experienced a high number of confirmed cases have also experienced a surge in the cases in prisons so there is a correlation. Counties like Cumberland have 28369 confirmed cases in the general public and the prison has 165 confirmed cases. Cumberland has the most confirmed cases in all of the Maine counties and the same goes for the prisons in cumberland county too as they have the highest confirmed cases among all the prisons at 165. Now if we look at the opposite side washington county has the lowest cases of all the counties at just 1583 confirmed cases and it has no cases in the prison present in that county. This may be because the county may be following good covid protocols and the same is reflected in the confirmed cases in the prison. The same pattern follows for all the counties and their respective prisons. Counties which has experienced a low to medium amount of cases in general public have a similar amount of cases in their county prisons.

```
county <- counties %>% group_by(state) %>% filter(state =="Maine") %>% select(state,county,confirmed_cas
county1 <- county %>% group_by(county)%>% summarize(confirmed_deaths = sum(confirmed_deaths))
county_1 <- county1[c(1,2,3,4,5,6,7,8,9,10,11,12,13,15,16,17), ,drop=TRUE]
county_1
```

Part 2:- Lets take a look at the deaths in the general public and the prisons.

```
## # A tibble: 16 x 2
##    county       confirmed_deaths
##    <chr>                   <int>
##  1 Androscoggin              190
##  2 Aroostook                 114
##  3 Cumberland                298
##  4 Franklin                   49
##  5 Hancock                    71
##  6 Kennebec                  184
##  7 Knox                       22
##  8 Lincoln                    20
##  9 Oxford                    115
## 10 Penobscot                 258
## 11 Piscataquis                25
## 12 Sagadahoc                  25
## 13 Somerset                   84
## 14 Waldo                      48
## 15 Washington                 45
## 16 York                      211
```

```
ggplot(data = county_1,mapping=aes(x=county, y=confirmed_deaths))+
  geom_line(aes(group=1),color="blue") + geom_point(color="violetred")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The above graph shows the deaths for all the counties in the state of Maine but we want the deats for only those counties which have prisons them. Hence I have filtered out only those counties which have prisons in

them and I compared that with the prison data set

```
prison <- prisons %>% group_by(`facility_state`) %>% filter(facility_state=="Maine")%>% select(facility
prison2 <- prison %>% group_by(facility_county)%>% summarize(total_inmate_deaths = sum(total_inmate_dea
prison2
```

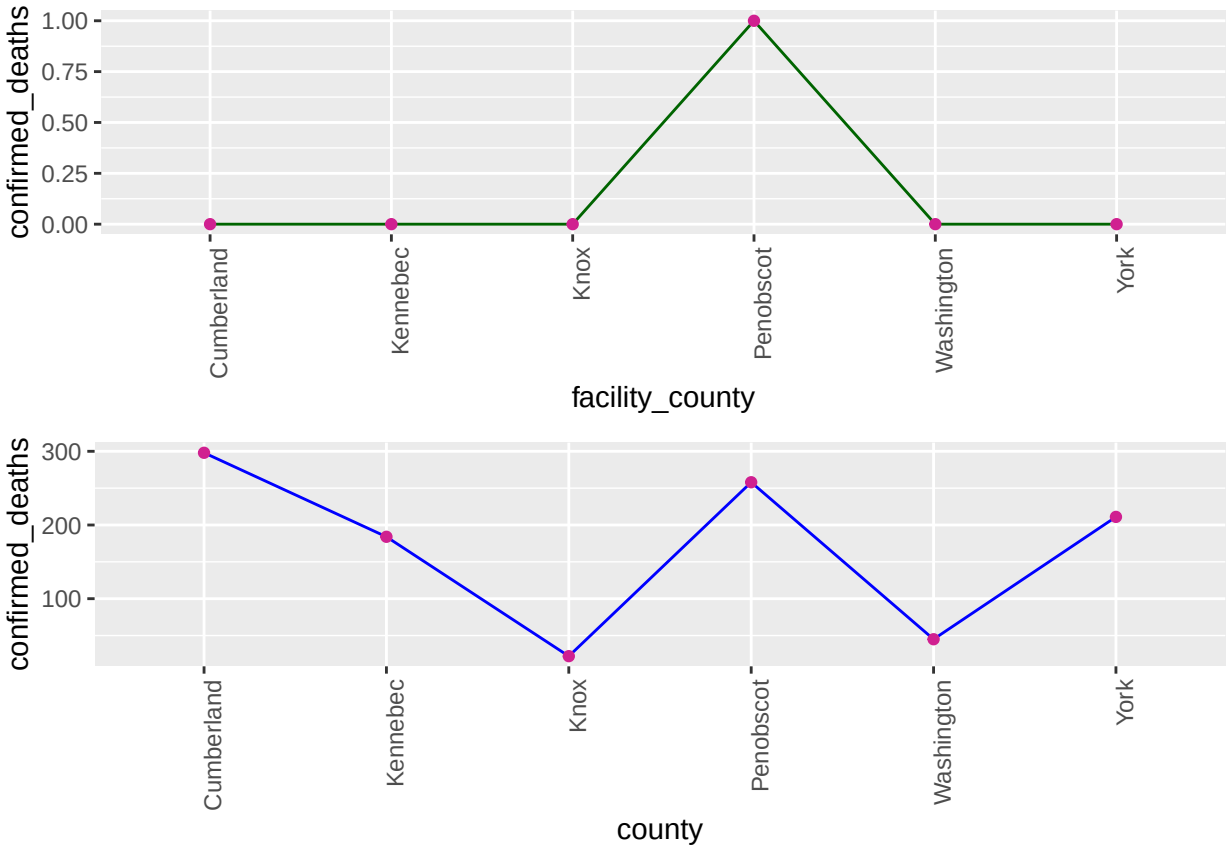
```
## # A tibble: 6 x 2
##   facility_county confirmed_deaths
##   <chr>                      <int>
## 1 Cumberland                     0
## 2 Kennebec                       0
## 3 Knox                           0
## 4 Penobscot                      1
## 5 Washington                     0
## 6 York                           0
```

```
plot3 <- ggplot(data = prison2,mapping=aes(x=facility_county, y=confirmed_deaths))+
  geom_line(aes(group=1),color="darkgreen") + geom_point(color="violetred")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
county2 <- county1[c(3,6,7,10,16,17), , drop=TRUE]
plot4 <- ggplot(data = county2,mapping=aes(x=county, y=confirmed_deaths))+
  geom_line(aes(group=1),color="blue") + geom_point(color="violetred")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
county2
```

```
## # A tibble: 6 x 2
##   county     confirmed_deaths
##   <chr>                 <int>
## 1 Cumberland              298
## 2 Kennebec                184
## 3 Knox                     22
## 4 Penobscot               258
## 5 Washington               45
## 6 York                    211
```

```
grid.arrange(plot3, plot4)
```



Summary for Part 2

If we compare the two graphs it is clearly evident that all almost 99% of the confirmed cases in prison have recovered. This may be because prisoners are less exposed to the outside world and prisons are usually devoid of any pollution. Another reason might be that prisoners are fit enough to overcome the virus as they might have at least 1 or 2 hours of good physical activity. When we see the general public deaths we can see that counties with most number of cases have the most deaths. This is trend we have been observing world wide across all the countries. There is no pattern here between the two graphs and it is clearly evident that prisoners have not succumbed to novel corona virus as compared to the general public.

Conclusion

Out of the two graphs which were plotted for the cases and the deaths we can see that deaths does not follow a pattern where as the confirmed cases follow a pattern. We can see that the confirmed cases as well as the death rate in prisoners is very less when compared to the general public and this may be due to a myriad of conditions ranging from physical fitness, location of prison, to lack of social interaction. The death rate in prisoners is almost negligible when compared to the general public and the death rate is below 1%. I could have further expanded this into a detailed analysis if I had age and gender of the prisoners and I could have faceted the graphs accordingly and compared them to the general public. Overall, from this analysis I can conclude that prisoners have a good chance of managing and navigating through the pandemic over the coming years without giving too much effort.