Comparative Analaysis of Covid Cases Between Three Urban Populations

In this notebook, I've reviewed COVID-19 cases over time in two counties containing cities. I wanted to see if the COVID-19 cases tended to follow the same trend between a large urban area (Cook County, Illinois, which Chicago is located in) and a smaller urban area (Onondaga County, New York, which Syracuse is located in).

Then, using that data to contruct a model, I wanted to see if a model based on these two urban counties could be applied to a third urban county (San Francisco County, California) as a way to estimate COVID-19 behavior.

Import Libraries and Data

Import Libraries

Import the following libraries: *tidyverse*: Cleaning and plotting *lubridate*: Date transformation

Import and Clean Data

U.S. COVID-19 case time series data is retrieved from https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/

Population data as of the 2020 census was retrieved from the following sources:

https://www.census.gov/quickfacts/fact/table/cookcountyillinois/PST045221 https://www.census.gov/quickfacts/onondagacountynewyork https://www.census.gov/quickfacts/sanfranciscocountycalifornia

```
url_in = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names = c("time_series_covid19_confirmed_global.csv","time_series_covid19_deaths_global.csv","time_
urls = str_c(url_in,file_names)

us_cases = read_csv(urls[3])
```

```
## Rows: 3342 Columns: 939
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (933): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cook_population = 5275541
onondaga_population = 476516
sanfransisco_population = 873965
```

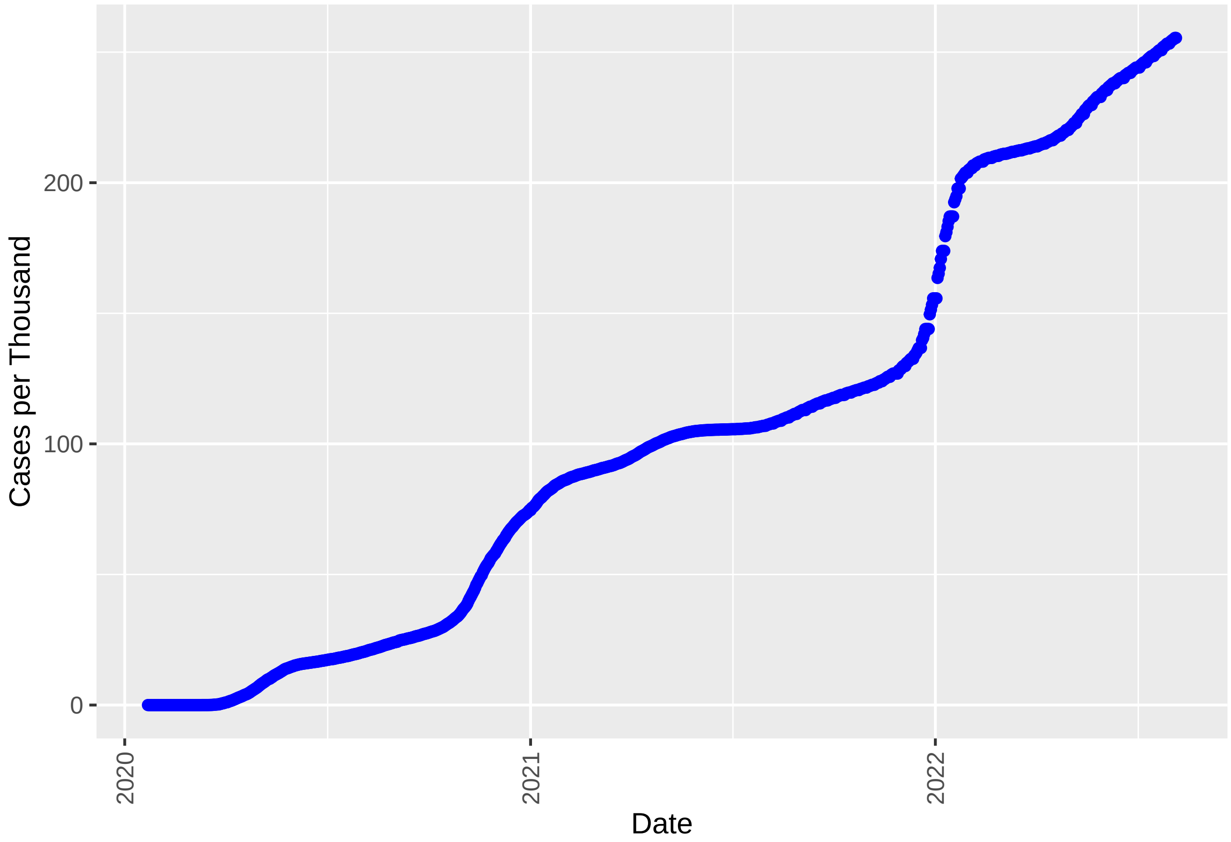
```
cook_cases = us_cases[us_cases$Admin2 == 'Cook' & us_cases$Province_State == 'Illinois',]
onondaga_cases = us_cases[us_cases$Admin2 == 'Onondaga' & us_cases$Province_State == 'New York',]

cook_cases = cook_cases %>% pivot_longer(cols=-c("UID","iso2","iso3","code3","FIPS","Admin2","Province_S
                                       names_to="date",
                                       values_to="cases"
                                       ) %>% mutate(date=mdy(date))
onondaga_cases = onondaga_cases %>% pivot_longer(cols=-c("UID","iso2","iso3","code3","FIPS","Admin2","Pr
                                       names_to="date",
                                       values_to="cases"
                                       ) %>% mutate(date=mdy(date))

cook_cases$per_thous = cook_cases$cases / (cook_population / 1000)
onondaga_cases$per_thous = onondaga_cases$cases / (onondaga_population / 1000)

cook_onondaga_cases = merge(cook_cases,onondaga_cases,by="date")
cook_onondaga_cases$average = (cook_onondaga_cases$per_thous.x + cook_onondaga_cases$per_thous.y) / 2
```

Data Exploration

Below is some preliminary exploration of the case data:

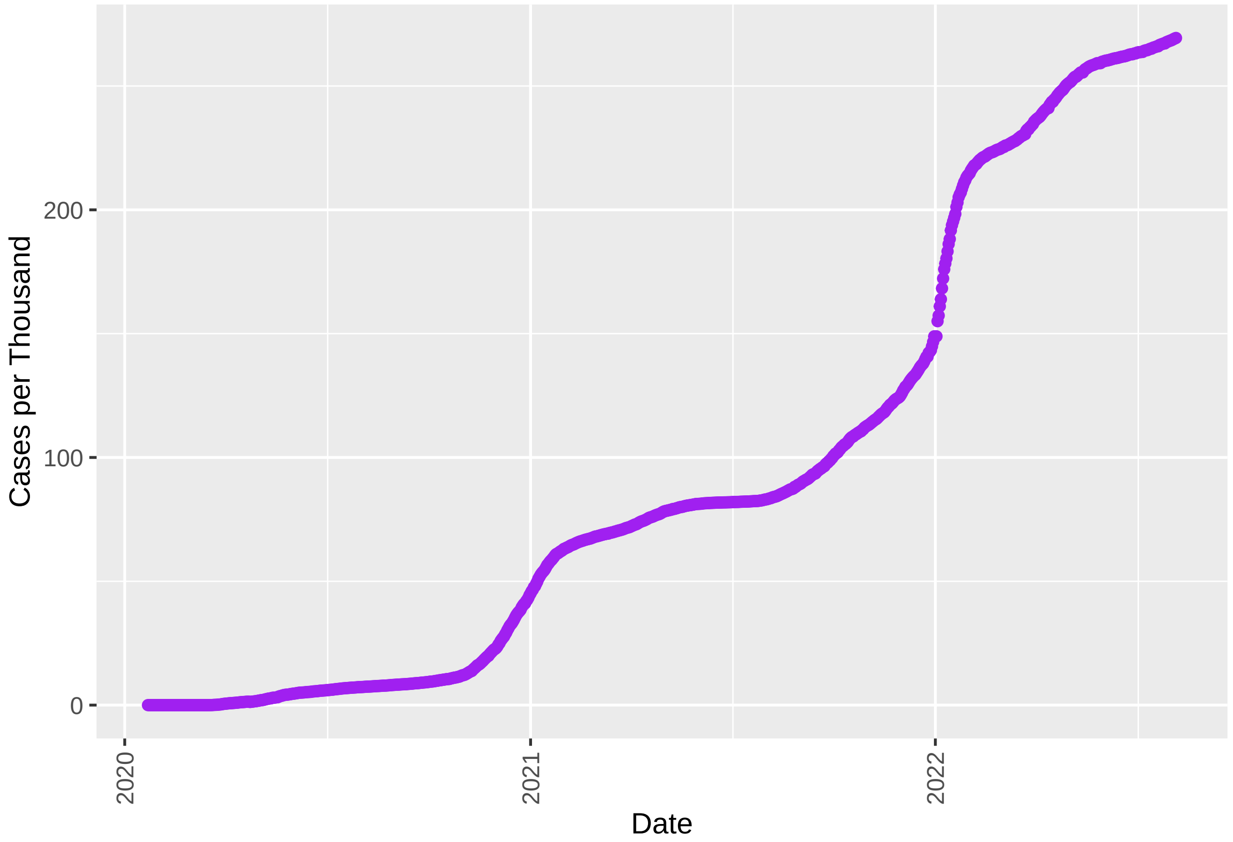
Cook County Cases Over Time

```
ggplot() + geom_point(aes(x=cook_onondaga_cases$date, y=cook_onondaga_cases$per_thous.x), color="blue")
```



Onondaga County Cases Over Time

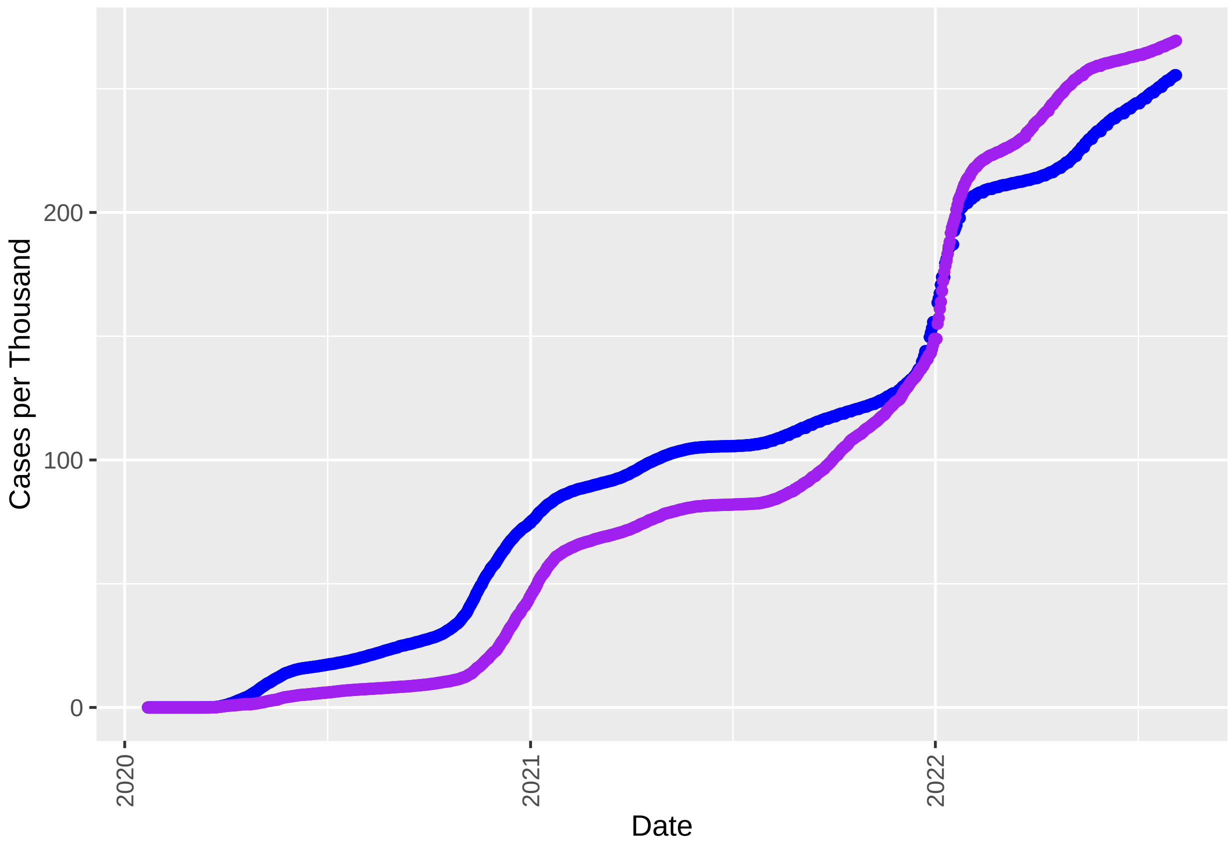
```
ggplot() + geom_point(aes(x=cook_onondaga_cases$date, y=cook_onondaga_cases$per_thous.y), color="purple
```



Cook and Onondaga County Cases Over Time, Compared

Cook County is plotted in blue, while Onondaga County is plotted in purple. Cases are plotted per thousand people, in order to compare the two counties equally.

```
ggplot() + geom_point(aes(x=cook_onondaga_cases$date, y=cook_onondaga_cases$per_thous.x), color="blue")
```



Modelling and Analysis

Using the average of the cases per day in both Cook and Onondaga Counties, I constructed a model of COVID-19 cases per day.

As can be seen, both counties follow a very similar pattern, and over time become almost identical. It is very interesting that, for cities with very different populations and in very different areas of the country, that the COVID-19 case behavior is very similar. This leads to a question: Can we use a model generated from these two counties to predict the behavior of a third?

Cook County is plotted in blue, Onondaga County is plotted in purple, and the model predictions are plotted in yellow.

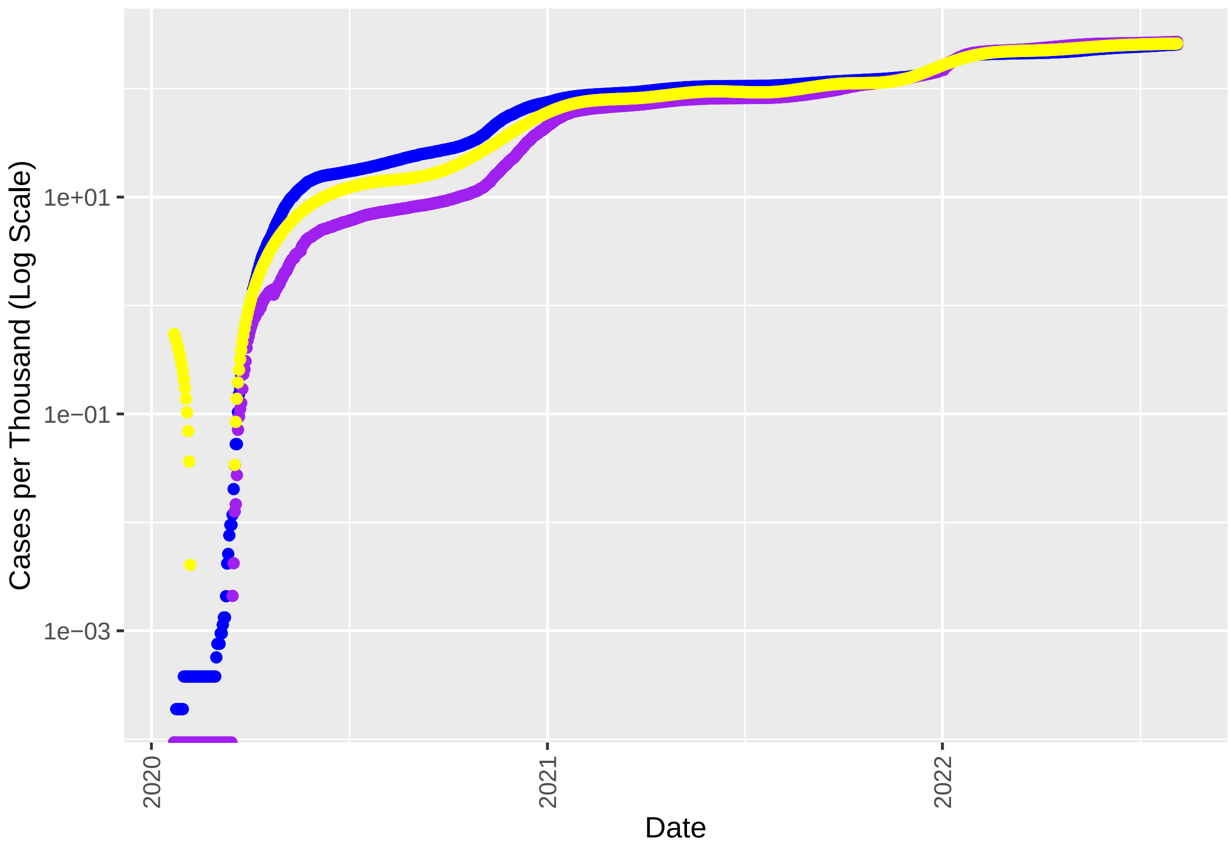
```
library(splines)
model <- lm(average ~ ns(date, 15), data = cook_onondaga_cases)

cook_onondaga_cases = cook_onondaga_cases %>% mutate(combinedpred=predict(model))

ggplot() + geom_point(aes(x=cook_onondaga_cases$date, y=cook_onondaga_cases$per_thous.x), color="blue")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```

```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 40 rows containing missing values (geom_point).
```


Cases per Thousand (Log Scale)
1e+01
1e−01
1e−03
2020
2021
2022
Date

Predicting Cases in San Francisco County

I selected San Francisco County as a third county that I wanted to see if I could predict cases for based on the model achieved from Cook and Onondaga Counties.

Below, I've plotted the San Fransisco cases in red and the model predictions in yellow.

As is evident, while the model does closely follow the same shape as the actual San Fransisco data, the actual data tends lower than the model's predictions.

```
sanfransisco_cases = us_cases[us_cases$Admin2 == 'San Francisco' & us_cases$Province_State == 'Californi
sanfransisco_cases = sanfransisco_cases %>% pivot_longer(cols=-c("UID","iso2","iso3","code3","FIPS","Adm
                                       names_to="date",
                                       values_to="cases"
                                       ) %>% mutate(date=mdy(date))
sanfransisco_cases$per_thous = sanfransisco_cases$cases / (sanfransisco_population / 1000)
```

```
ggplot() + geom_point(aes(x=sanfransisco_cases$date, y=sanfransisco_cases$per_thous), color="red") + geo
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

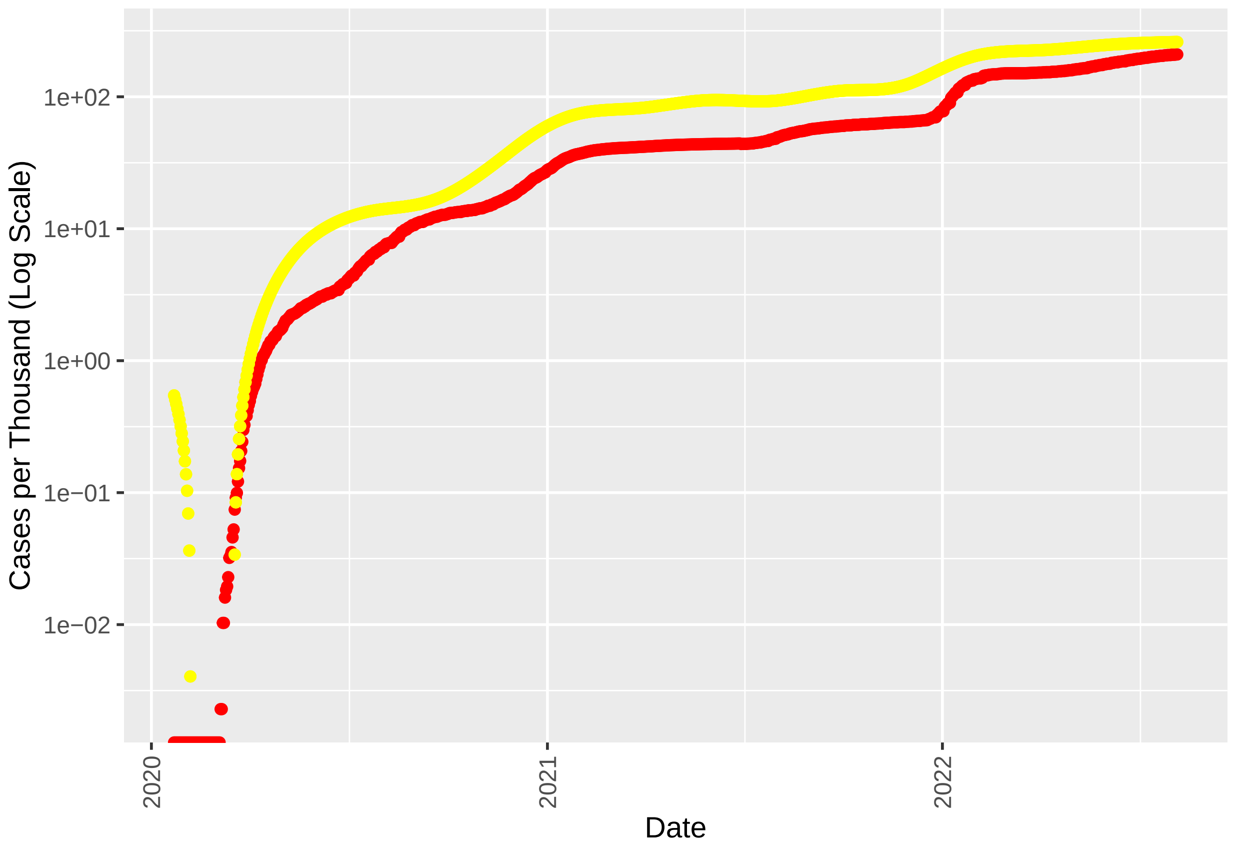
```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 40 rows containing missing values (geom_point).
## Removed 40 rows containing missing values (geom_point).
```


1e+02
1e+01
1e+00
1e−01
1e−02
Cases per Thousand (Log Scale)
2020
2021
2022
Date

Bias and Future Analysis

Bias sources for this analysis may be personal (I've lived in both Chicago and Syracuse), and so have a personal connection to this data, however I don't have strong feelings about my results in this case.

Differences in governmental policies and social norms (mask mandates, vaccine rolleout, vaccine adoption, etc) in all three areas may hvae an impact on these results. I've also chosen three areas that were very proactive in terms of COVID-19 containment policies. I suspect that further analysis with other areas of the country may produce different results. For example, counties that did not adopt mask mandate policies or that have low social acceptance of those policies may follow different models than those produced above.