COVID Dataset

2023-05-02

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

For this project, I will be attempting to build a model that can estimate how many deaths have occurred in the US, by state, from COVID-19, based off of the number of cases. I am curious to see if there is a linear relationship between the two variables.

Setup

```
{r load_data} library(tidyverse) conf_us <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/
master/csse_covid_19_data/csse_covid_19_time_series/
time_series_covid19_confirmed_US.csv") conf_glob <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/
master/csse_covid_19_data/csse_covid_19_time_series/
time_series_covid19_confirmed_global.csv") deaths_us <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/
master/csse_covid_19_data/csse_covid_19_time_series/
time_series_covid19_deaths_US.csv") deaths_glob <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/
master/csse_covid_19_data/csse_covid_19_time_series/
time_series_covid19_deaths_global.csv")
```

First, I have loaded the data sets into R - I have both US and global data, but I will only be dealing with US data for the model. This data is from the John Hopkins COVID-19 dashboard, found here on github: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data It was finally retired and data stopped being collected and updated on March 9, 2023.

```{r summary} conf_glob <- conf_glob %>% pivot_longer(cols=-c('Province/State', 'Country/Region', 'Lat', 'Long'), names_to='date', values_to="cases") %>% select(-c(Lat,Long))

deaths_glob <- deaths_glob %>% pivot_longer(cols=-c('Province/State', 'Country/Region', 'Lat', 'Long'), names_to='date', values_to="deaths") %>% select(-c(Lat,Long))

conf_us <- conf_us %>% pivot_longer(cols=-(UID:Combined_Key), names_to='date', values_to='cases') %>% select(Admin2:cases) %>% mutate(date=mdy(date)) %>% select(-c(Lat, Long_))

deaths_us <- deaths_us %>% pivot_longer(cols=-(UID:Population), names_to='date', values_to='deaths') %>% select(Admin2:deaths) %>% mutate(date=mdy(date)) %>% select(-c(Lat, Long_))
```

In this step, I cleaned the data to get it into a more usable and understandable format. I wanted to ensure variables were clearly named and I could easily identify what I was working with.

```{r combine}
global <- conf_glob %>%
  full_join(deaths_glob) %>%
  rename(Country_Region='Country/Region',
         Province_State='Province/State') %>%
  mutate(date=mdy(date))
global <- global %>% filter(cases > 0)

global <- global %>%
  unite('Combined_Key',
  c(Province_State, Country_Region),
  sep=", ",
  na.rm=TRUE,
  remove = FALSE)

uid_lookup_url <-
"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/
csse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv"
uid <- read.csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
global <- global %>%
  left_join(uid, by=c("Province_State", "Country_Region")) %>%
              select(-c(UID, FIPS)) %>%
              select(Province_State, Country_Region, date, cases,
deaths, Population, Combined_Key)


us <- conf_us %>%
  full_join(deaths_us)



summary(us)
summary(global)
```

In this step, I integrated population data into the global dataset, as there was a population variable in the US one and I thought that would be useful.

```{r us_by_state} us_by_state <- us %>% group_by(Province_State, Country_Region, date) %>% summarize(cases=sum(cases), deaths=sum(deaths), Population = sum(Population)) %>% mutate(deaths_per_mill=deaths*1000000 / Population) %>% select(Province_State, Country_Region, date, cases, deaths, deaths_per_mill, Population) %>% ungroup()
```

us_totals <- us_by_state %>% group_by(Country_Region, date) %>% summarize(cases=sum(cases), deaths=sum(deaths), Population = sum(Population)) %>% mutate(deaths_per_mill=deaths*1000000 / Population) %>% select(Country_Region, date, cases, deaths, deaths_per_mill, Population) %>% ungroup()

tail(us_totals)

Here, I create two dataframes:
1. us_by_state reflects the cases and deaths in the US split by day and also by state
2. us_totals reflects the total number of cases and deaths in the US per day

```{r us}
us_totals %>%
  filter(cases>0) %>%
  ggplot(aes(x=date, y=cases)) +
  geom_line(aes(color="cases")) +
  geom_point(aes(color="cases")) +
  geom_line(aes(y=deaths, color="deaths")) +
  geom_point(aes(y=deaths, color="deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  labs(title = "COVID-19 in the US", y=NULL)
```

In this first visualization, we can see the change in cases and deaths in the US over time. It may appear to be common sense, but it is important to note that cases are always higher than deaths, since people do survive COVID-19 when they get infected by it. We can see some form of seasonality in the graph too, when people got more or less reckless depending on the time of year and media narratives.

{r cali} state <- "California" us_by_state %>% filter(Province_State == state) %>% filter(cases>0) %>% ggplot(aes(x=date, y=cases)) + geom_line(aes(color="cases")) + geom_point(aes(color="cases")) + geom_line(aes(y=deaths, color="deaths")) + geom_point(aes(y=deaths, color="deaths")) + scale_y_log10() + theme(legend.position="bottom",   axis.text.x = element_text(angle=90)) + labs(title = str_c("COVID-19 in ", state), y=NULL)

{r tex} state <- "Texas" us_by_state %>% filter(Province_State == state) %>% filter(cases>0) %>% ggplot(aes(x=date, y=cases)) + geom_line(aes(color="cases")) + geom_point(aes(color="cases")) + geom_line(aes(y=deaths, color="deaths")) + geom_point(aes(y=deaths, color="deaths")) + scale_y_log10() + theme(legend.position="bottom",   axis.text.x = element_text(angle=90)) + labs(title = str_c("COVID-19 in ", state), y=NULL)

In these two visualizations, I look at the data for California and Texas. I chose these as I lived in both of these states at various stages of the pandemic. It is interesting to see how they are similar and differ in their case rates.

```{r model} mod <- lm(deaths ~ cases, data=us_by_state) summary(mod)

us_by_state_w_pred <- us_by_state %>% mutate(pred=predict(mod))
tail(us_by_state_w_pred)

us_by_state_w_pred %>% ggplot() + geom_point(aes(x=cases, y=deaths), color="purple") + geom_point(aes(x=cases, y=pred), color="green")

```

Finally. here is the model that I built that uses cases in a given state to predict the number of deaths by day. I found that initially this model worked very well with its predictions, but as time went on, its predictions got significantly less accurate. This is due to the fact that currently, COVID is not growing as quickly as it did in the prime of the pandemic. We did not have vaccines, and didn't even fully understand the extent of coronavirus until more than a year into it. Thus, we see that the model overestimates the number of deaths closer to the present day.