# John Hopkins Covid

JS

2022-10-08

## John Hopkins Covid Data Analyis

Early on when looking at COVID data I knew I wanted to explore the three prairie provinces in Canada - Alberta, Saskatchewan and Manitoba - to see which produced the best results in cases and deaths. I spent the pandemic era in between those 3 areas and the general populations attitude towards lock downs and anecdotally views on restrictions were keenly different in all three. Alberta - very anti restrictions, Saskatchewan a mix, Manitoba very pro restrictions.

Initially we will load the libraries and setup knit

```
# include libraries required and setup knit

knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(lubridate)
```

Next we will import the data set. Since I am only going to be using data from the three provinces I will only be including global cases and deaths. Additionally will need province population data for future use and will import that now as well from STAT Canada. A link to the data sources are available at:

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

https://www150.statcan.gc.ca/t1/tbl1/en/cv!recreate.action?pid=1710000501&selectedNodeIds=1D8,1D9,1D10,3D1&checkedLevels=1D1&refPeriods=20200101,20200101&dimensionLayouts=layout2,layout3,layout3,layout3&vectorDisplay=false

```
# import data from source

url1 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid

url2 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid

url3 <- "https://www150.statcan.gc.ca/t1/tbl1/en/dtl!downloadDbLoadingData.action?pid=1710000501&latest

covid_case_data <- read_csv(url1)
```

```
## Rows: 289 Columns: 996
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (994): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...
```

```
##
## 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.
```

```
covid_death_data <- read_csv(url2)
```

```
## Rows: 289 Columns: 996
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (994): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...
##
## 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.
```

```
province_pop <- read_csv(url3)
```

```
## Rows: 3 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): GEO, DGUID, Sex, Age group, UOM, SCALAR_FACTOR, VECTOR, COORDINATE
## dbl (5): REF_DATE, UOM_ID, SCALAR_ID, VALUE, DECIMALS
## lgl (3): STATUS, SYMBOL, TERMINATED
##
## 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.
```

```
view(covid_case_data)
view(covid_death_data)
view(province_pop)
```

We'll do a quick check to see what kind of information we have available. We can already see we have a time series by day and a province data source, but am looking to see if have a population column for future use and analysis.

Turns out, we do not. As such we're going to jump back up to "r import" and import a data set with provincial population data to use and then check headers again.

```
head(covid_death_data)
```

```
## # A tibble: 6 x 996
##   Province~1 Count~2   Lat  Long 1/22/~3 1/23/~4 1/24/~5 1/25/~6 1/26/~7 1/27/~8
##   <chr>      <chr>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 <NA>       Afghan~  33.9 67.7        0       0       0       0       0       0
## 2 <NA>       Albania  41.2 20.2        0       0       0       0       0       0
## 3 <NA>       Algeria  28.0  1.66       0       0       0       0       0       0
## 4 <NA>       Andorra  42.5  1.52       0       0       0       0       0       0
## 5 <NA>       Angola  -11.2 17.9        0       0       0       0       0       0
## 6 <NA>       Antarc~ -71.9 23.3        0       0       0       0       0       0
## # ... with 986 more variables: ‘1/28/20‘ <dbl>, ‘1/29/20‘ <dbl>,
## #   ‘1/30/20‘ <dbl>, ‘1/31/20‘ <dbl>, ‘2/1/20‘ <dbl>, ‘2/2/20‘ <dbl>,
## #   ‘2/3/20‘ <dbl>, ‘2/4/20‘ <dbl>, ‘2/5/20‘ <dbl>, ‘2/6/20‘ <dbl>,
## #   ‘2/7/20‘ <dbl>, ‘2/8/20‘ <dbl>, ‘2/9/20‘ <dbl>, ‘2/10/20‘ <dbl>,
## #   ‘2/11/20‘ <dbl>, ‘2/12/20‘ <dbl>, ‘2/13/20‘ <dbl>, ‘2/14/20‘ <dbl>,
## #   ‘2/15/20‘ <dbl>, ‘2/16/20‘ <dbl>, ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>,
## #   ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ‘2/21/20‘ <dbl>, ‘2/22/20‘ <dbl>, ...
```

```
head(covid_case_data)
```

```
## # A tibble: 6 x 996
##   Province~1 Count~2   Lat  Long 1/22/~3 1/23/~4 1/24/~5 1/25/~6 1/26/~7 1/27/~8
##   <chr>      <chr>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 <NA>       Afghan~  33.9 67.7        0       0       0       0       0       0
## 2 <NA>       Albania  41.2 20.2        0       0       0       0       0       0
## 3 <NA>       Algeria  28.0  1.66       0       0       0       0       0       0
## 4 <NA>       Andorra  42.5  1.52       0       0       0       0       0       0
## 5 <NA>       Angola  -11.2 17.9        0       0       0       0       0       0
## 6 <NA>       Antarc~ -71.9 23.3        0       0       0       0       0       0
## # ... with 986 more variables: ‘1/28/20‘ <dbl>, ‘1/29/20‘ <dbl>,
## #   ‘1/30/20‘ <dbl>, ‘1/31/20‘ <dbl>, ‘2/1/20‘ <dbl>, ‘2/2/20‘ <dbl>,
## #   ‘2/3/20‘ <dbl>, ‘2/4/20‘ <dbl>, ‘2/5/20‘ <dbl>, ‘2/6/20‘ <dbl>,
## #   ‘2/7/20‘ <dbl>, ‘2/8/20‘ <dbl>, ‘2/9/20‘ <dbl>, ‘2/10/20‘ <dbl>,
## #   ‘2/11/20‘ <dbl>, ‘2/12/20‘ <dbl>, ‘2/13/20‘ <dbl>, ‘2/14/20‘ <dbl>,
## #   ‘2/15/20‘ <dbl>, ‘2/16/20‘ <dbl>, ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>,
## #   ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ‘2/21/20‘ <dbl>, ‘2/22/20‘ <dbl>, ...
```

```
head(province_pop)
```

```
## # A tibble: 3 x 16
##   REF_DATE GEO   DGUID Sex   Age g~1 UOM   UOM_ID SCALA~2 SCALA~3 VECTOR COORD~4
##      <dbl> <chr> <chr> <chr> <chr>   <chr>  <dbl> <chr>     <dbl> <chr>  <chr>
## 1     2020 Mani~ 2016~ Both~ All ag~ Pers~    249 units         0 v4688~ 8.1.1
## 2     2020 Sask~ 2016~ Both~ All ag~ Pers~    249 units         0 v4691~ 9.1.1
## 3     2020 Albe~ 2016~ Both~ All ag~ Pers~    249 units         0 v4695~ 10.1.1
## # ... with 5 more variables: VALUE <dbl>, STATUS <lgl>, SYMBOL <lgl>,
## #   TERMINATED <lgl>, DECIMALS <dbl>, and abbreviated variable names
## #   1: ‘Age group‘, 2: SCALAR_FACTOR, 3: SCALAR_ID, 4: COORDINATE
```

###Cleaning

We now have all the info we need to proceed, but will first have to clean up a bit. We will be pivoting to run the dates down as rows, changing date to a date format, changing to a factor and then checking everything is in the formats we desire.

```
# select only columns required
# filter out unknown age groups

covid_death_clean <- covid_death_data %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
                 names_to = "date",
                 values_to = "deaths") %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long)) %>%
  rename(Province = 'Province/State',
         Country = 'Country/Region')
```

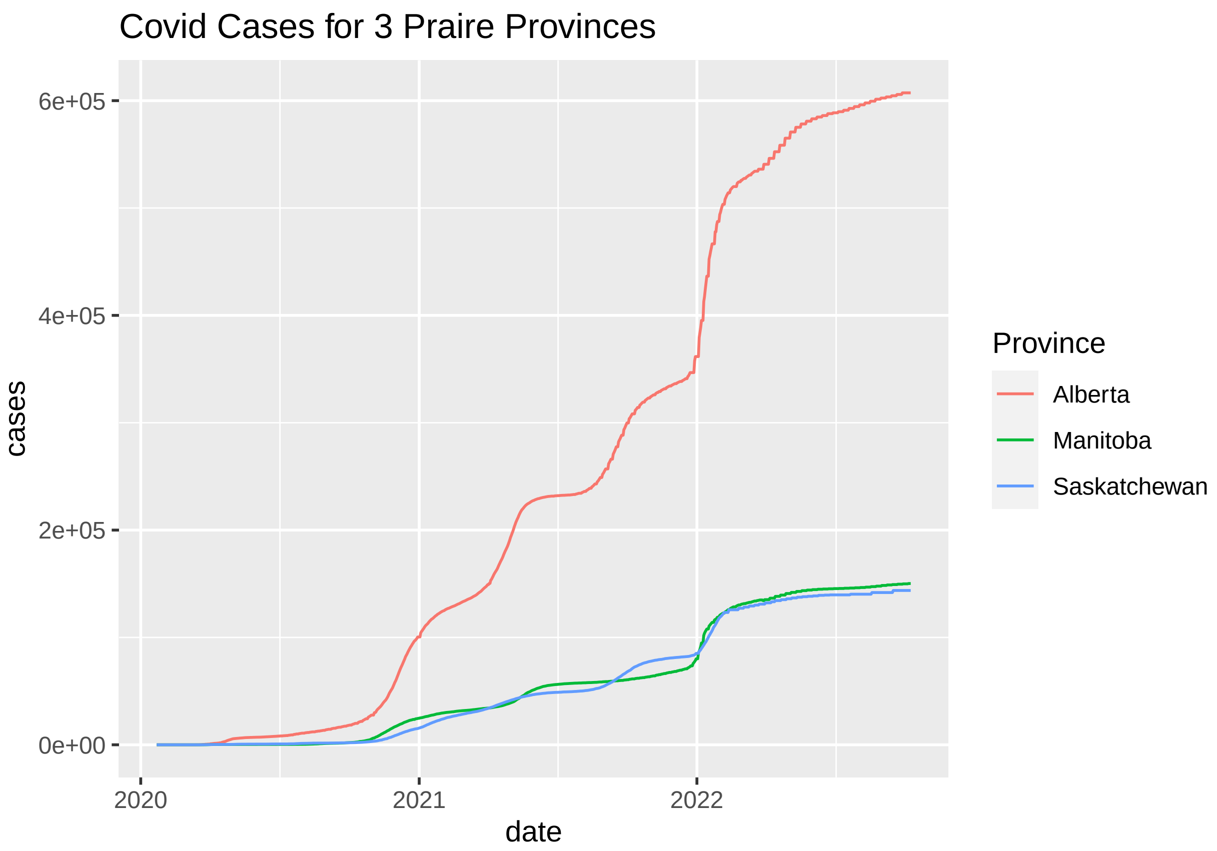
```
covid_case_clean <- covid_case_data %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
                 names_to = "date",
                 values_to = "cases") %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long)) %>%
  rename(Province = 'Province/State',
         Country = 'Country/Region')


covid_death_clean$Country <- factor(covid_death_clean$Country)
covid_death_clean$Province <- factor(covid_death_clean$Province)

covid_case_clean$Country <- factor(covid_case_clean$Country)
covid_case_clean$Province <- factor(covid_case_clean$Province)

str(covid_death_clean)
```

```
## tibble [286,688 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Province: Factor w/ 91 levels "Alberta","Anguilla",..: NA NA NA NA NA NA NA NA NA NA ...
##  $ Country : Factor w/ 201 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ date    : Date[1:286688], format: "2020-01-22" "2020-01-23" ...
##  $ deaths  : num [1:286688] 0 0 0 0 0 0 0 0 0 0 ...
```

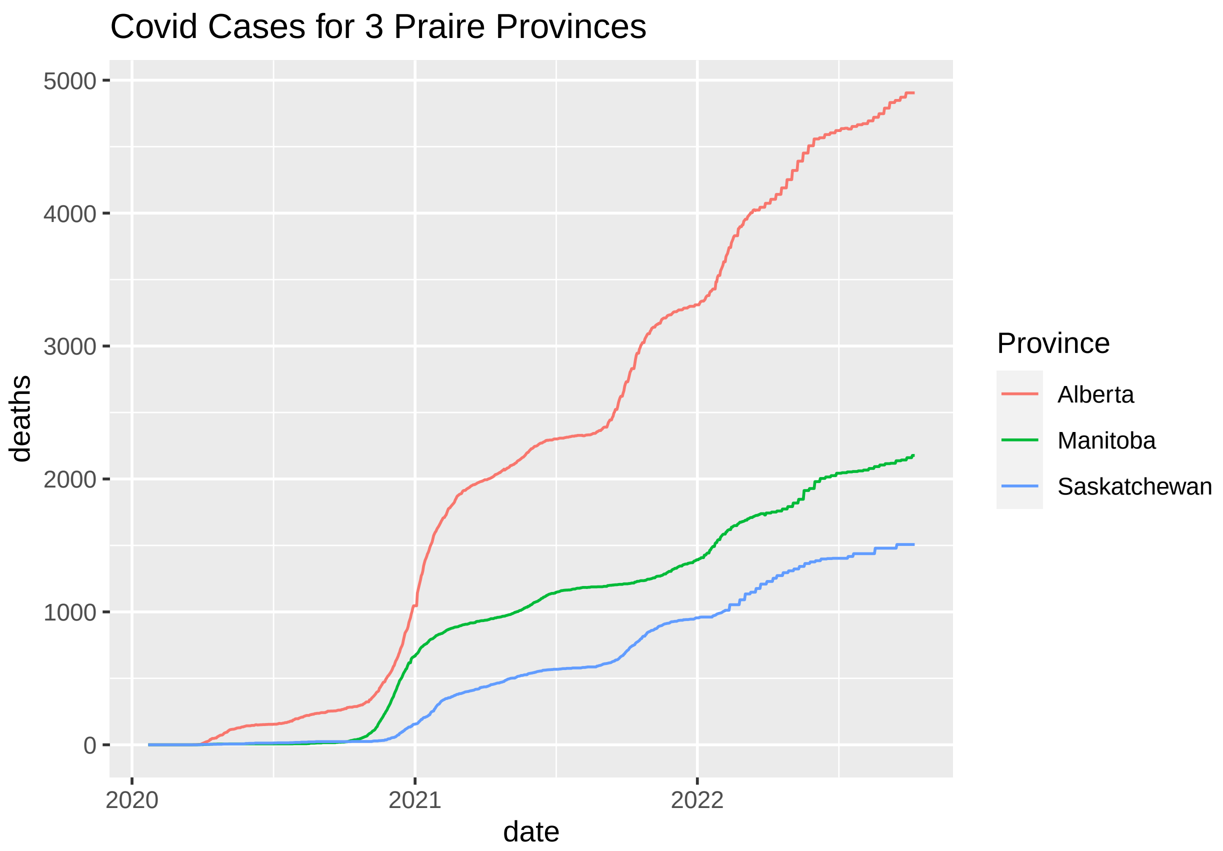
###Plotting and Analysis

Next we'll plot our case and death data to see if we can gain any initial information.

```
ggplot(subset(covid_case_clean, Province %in% c("Manitoba", "Saskatchewan", "Alberta")), aes(x = date, 
  geom_line(aes(y = cases, color = Province))
```



```
ggplot(subset(covid_death_clean, Province %in% c("Manitoba", "Saskatchewan", "Alberta")), aes(x = date,
  geom_line(aes(y = deaths, color = Province))
```



Initially this seems pretty noteworthy, but we have to considered and standardized for population to draw any conclusions so we'll set that up now in advance. While we could clean the data and use the entire table, considering we're only using 3 static data points for population we'll simply assign those the direct value and check that they're correct.

```
# select only columns required
# filter out unknown age groups


manitoba <- province_pop %>%
 filter(province_pop$GEO == 'Manitoba')

man_pop <- manitoba$VALUE

saskatchewan <- province_pop %>%
 filter(province_pop$GEO == 'Saskatchewan')

sask_pop <- saskatchewan$VALUE

alberta <- province_pop %>%
 filter(province_pop$GEO == 'Alberta')

alb_pop <- alberta$VALUE

man_pop
```

```
## [1] 1379888
```

```
sask_pop
```

```
## [1] 1178467
```

```
alb_pop
```

```
## [1] 4416682
```

The last thing I want to compare is the standardized rate to population of each of the 3 provinces to one and other. To do so will simply take the max count of cases and compare it to each provinces population.

```
can_covid_case = covid_case_clean %>%
    filter(covid_case_clean$Province %in% c("Manitoba", "Saskatchewan", "Alberta"))

can_covid_death = covid_death_clean %>%
    filter(covid_death_clean$Province %in% c("Manitoba", "Saskatchewan", "Alberta"))

man_per_person <- max(can_covid_case$cases[can_covid_case$Province =="Manitoba"]) / man_pop

sask_per_person <- max(can_covid_case$cases[can_covid_case$Province =="Saskatchewan"]) / sask_pop

alb_per_person <- max(can_covid_case$cases[can_covid_case$Province =="Alberta"]) / alb_pop

man_per_person
```

```
## [1] 0.1088588
```

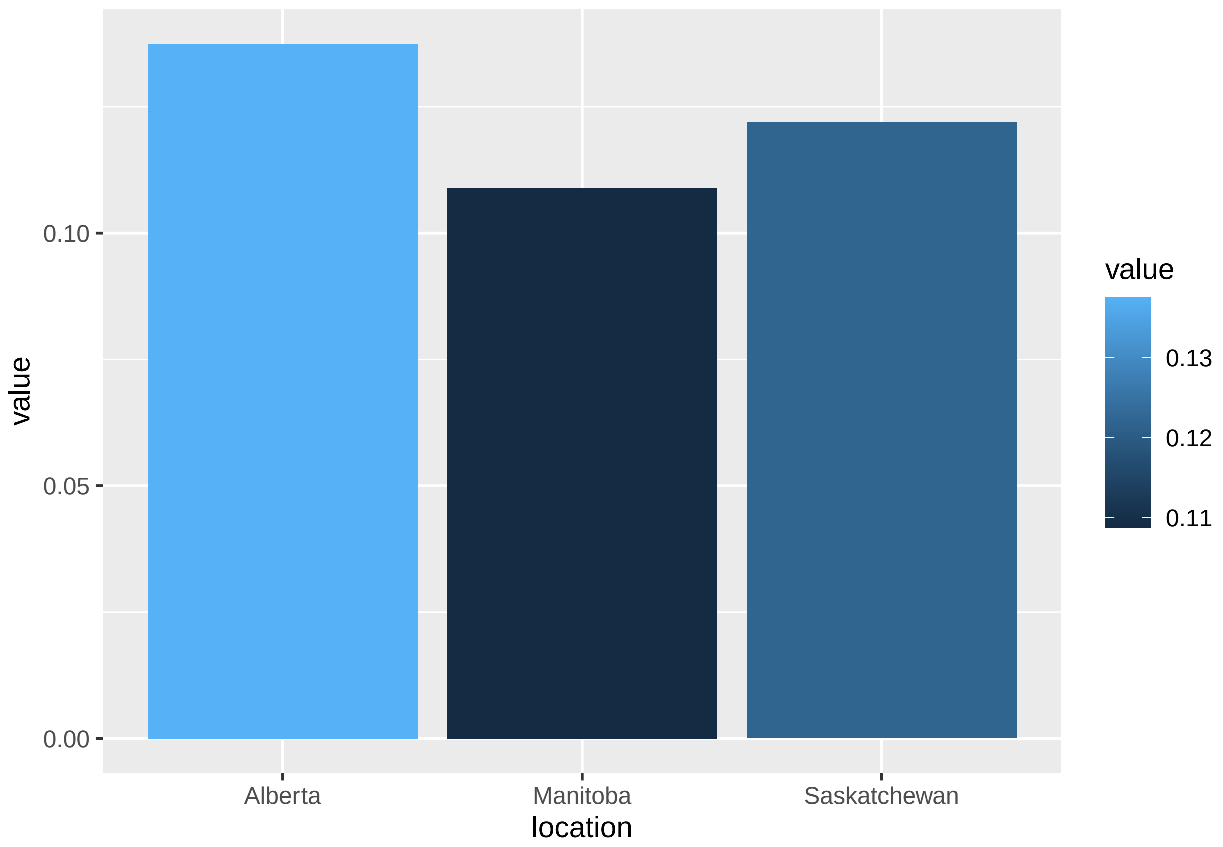
```
sask_per_person
```

```
## [1] 0.121978
```

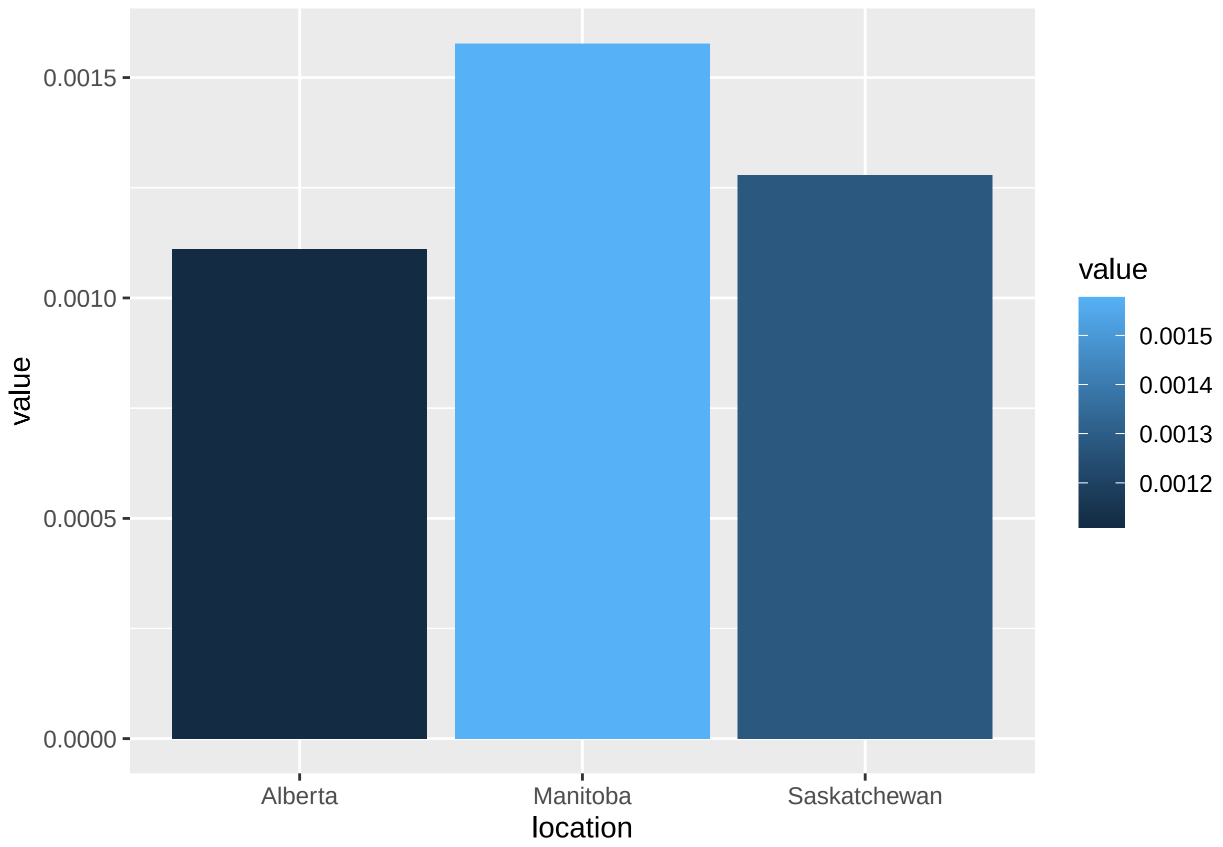
```
alb_per_person
```

```
## [1] 0.1374989
```

```
chart_time <- data.frame(rate = c("1", "2", "3"), location = c("Alberta", "Saskatchewan", "Manitoba"), 

ggplot(data = chart_time, aes(x = location, y = value)) + geom_bar(stat = "identity", aes(fill = value)
```



```
d_man_per_person <- max(can_covid_death$deaths[can_covid_death$Province =="Manitoba"]) / man_pop

d_sask_per_person <- max(can_covid_death$deaths[can_covid_death$Province =="Saskatchewan"]) / sask_pop

d_alb_per_person <- max(can_covid_death$deaths[can_covid_death$Province =="Alberta"]) / alb_pop

chart_time <- data.frame(rate = c("0.05", "0.1", "0.15"), location = c("Alberta", "Saskatchewan", "Mani

ggplot(data = chart_time, aes(x = location, y = value)) + geom_bar(stat = "identity", aes(fill = value))
```



Interestingly, while Manitoba has the lowest case count per person, it has the highest death rate and Alberta the opposite. This leads me to believe that on the surface Alberta's handling of the pandemic was indeed more effective at keeping people safe and alive.

###Modelling

Lastly, we'll do a couple simple linear models to see how strong of, if any, the relationship is between cases in a province of our sample. First off, cases by province.

```
mod_case = lm(cases~Province, data = can_covid_case)

summary(mod_case)
```

```
##
## Call:
## lm(formula = cases ~ Province, data = can_covid_case)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -248311  -57032  -15471   74678  358978
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             248311       4283   57.97   <2e-16 ***
## ProvinceManitoba       -189570       6058  -31.30   <2e-16 ***
## ProvinceSaskatchewan  -191279       6058  -31.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 134900 on 2973 degrees of freedom
## Multiple R-squared:  0.3071, Adjusted R-squared:  0.3066
## F-statistic: 658.9 on 2 and 2973 DF,  p-value: < 2.2e-16
```

Next we will do the same regarding deaths.

```
mod_death = lm(deaths~Province, data = can_covid_death)

summary(mod_death)
```

```
##
## Call:
## lm(formula = deaths ~ Province, data = can_covid_death)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2162.24  -587.18    -6.02   721.82  2742.76
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2162.24      34.99   61.79   <2e-16 ***
## ProvinceManitoba     -1207.37      49.49  -24.40   <2e-16 ***
## ProvinceSaskatchewan -1575.05      49.49  -31.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1102 on 2973 degrees of freedom
## Multiple R-squared:  0.2717, Adjusted R-squared:  0.2712
## F-statistic: 554.5 on 2 and 2973 DF,  p-value: < 2.2e-16
```

Looking at R-squared values, it is somewhat interesting to note that while there is a minor correlation likely between cases and deaths in the three provinces it is not very strong with r-squared values of 0.30 and 0.27 respectively. Those border on what is generally classified as weak/moderate correlation.

## Bias

First and foremost, I am a part of the data, which means I have prior knowledge, handling and conclusions drawn prior to even starting. While I feel I didn't proceed any differently that cannot be discounted.

Additionally, conclusions drawn do not factor in any other outside biases on data that exist. For example, Manitoba might have a vastly older or young population. Alberta may have significantly inferior health care (these are examples, not accusations), and as such these biases would likely impact the rate at which deaths occur from cases.