COVID US Deaths Report

Bryan Strub

10/10/2021

Introduction

About the data:

> The dataset I will be reporting on is the time_series_covid19_deaths_US.csv data from Johns Hopkins. This dataset contains time series data of COVID related deaths by US county. It is updated on a daily basis for all counties. The data contains columns for a unique row identifier, iso2 code, iso3 code, FIPS, County name, State name, latitude, longitude, population, and a variable for every date since January 22nd, 2020. In order for analysis to occur, the data will be 'melted' so that the data is 'long' by date/location, instead of 'wide' by date.

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

Required packages:

This Rmd file depends on the following libraries:

- tibble
- magrittr
- data.table
- ggplot
- scales

Initialize RMD document and read in data

```
library(ggplot2)
library(magrittr)
library(data.table)
library(scales)
main <- read.csv("https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_1
```

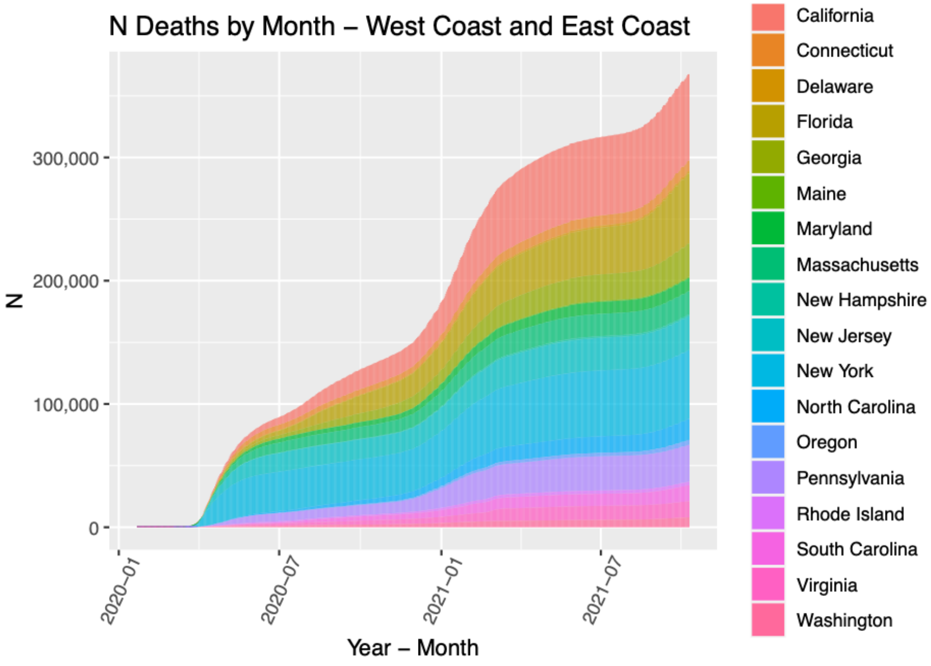
Tidy and Transform The Data

```
main <- melt(main,id.vars = c("UID","iso2","iso3","code3","FIPS","Admin2","Province_State","Country_Reg
main[, variable := as.POSIXct(gsub("\\.","/",gsub("X","",variable)),format = "%m/%d/%y")]
```

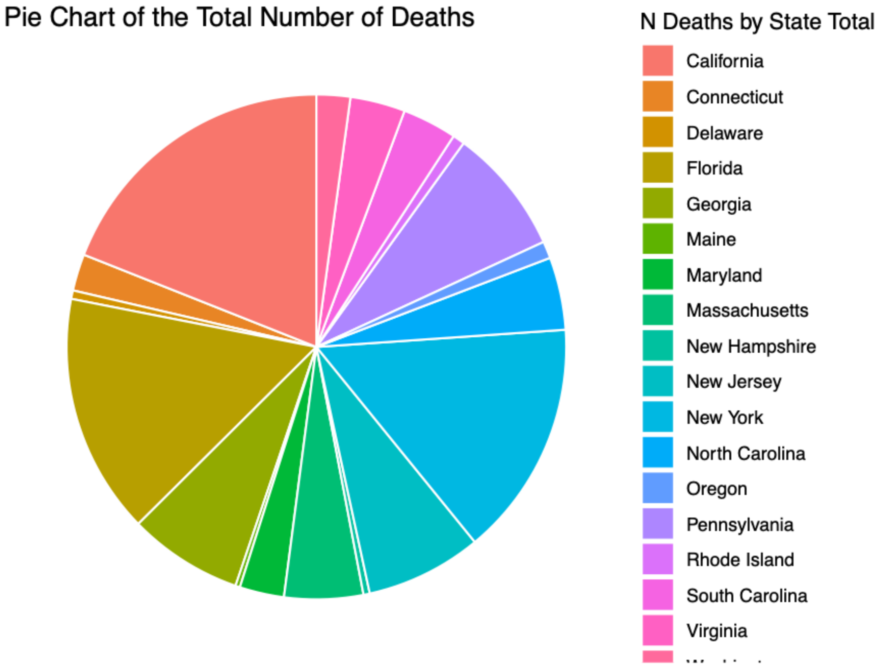
Add Visualizations and Analysis

```
main[, c("UID","iso2","iso3","code3","FIPS","Country_Region","Lat","Long_","Combined_Key") := NULL]
east.coast <- main[Province_State %in% c("Connecticut","Delaware","Florida","Georgia",'Maine',"Maryland
east.coast[, COAST := "EAST"]
west.coast <- main[Province_State %in% c("Washington","Oregon","California")]
```

```r
west.coast[, COAST := "WEST"]

both.coasts <- rbind(east.coast,west.coast)
both.coasts[, value := sum(value), by = c("Province_State","variable")]
both.coasts[, Population := sum(Population), by = c("Province_State","variable")]
both.coasts[, Admin2 := NULL]
both.coasts <- unique(both.coasts)

# both.coasts[,variable := substr(variable,1,7)]
# both.coasts[, value := sum(value), by = c("Province_State","variable")]
# both.coasts <- unique(both.coasts)


ggplot(both.coasts, aes(fill=Province_State, y=value, x=variable)) +
    geom_bar( stat="identity") +
    scale_y_continuous(labels = comma) +
    labs(title = "N Deaths by Month - West Coast and East Coast", x = "Year - Month", y = "N") +
    theme(axis.text.x = element_text(angle=65, vjust=0.6))
```

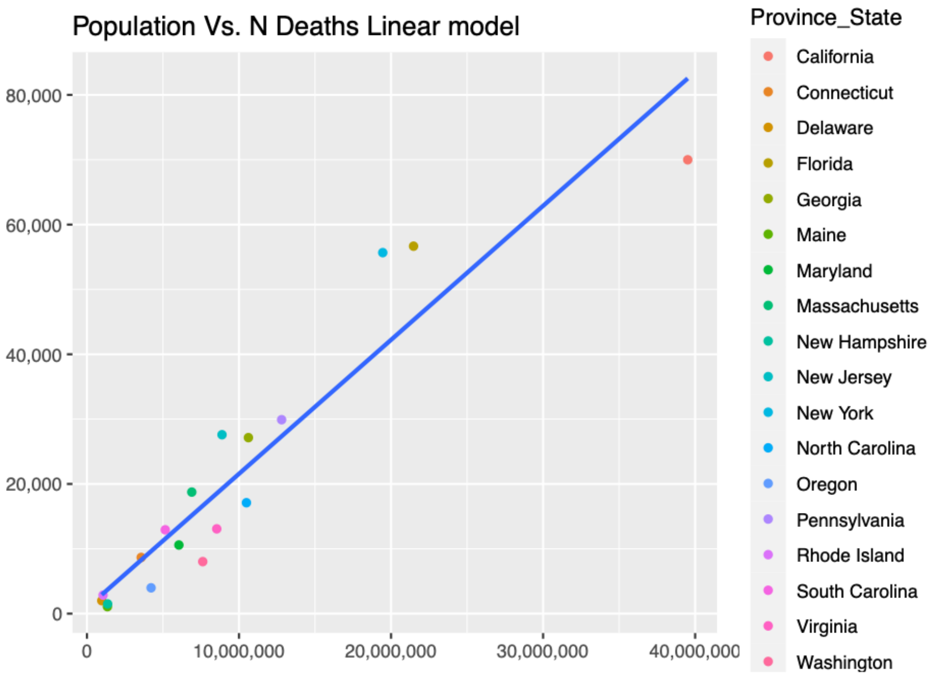


```r
both.coasts.agg <- copy(both.coasts)
both.coasts.agg <- both.coasts.agg[variable == max(variable)]

ggplot(both.coasts.agg, aes(x="", y=value, fill= Province_State)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  labs(fill="N Deaths by State Total",
       x=NULL,
       y=NULL,
       title="Pie Chart of the Total Number of Deaths") +
  theme_void()
```



```r
ggplot(both.coasts.agg, aes(Population,value)) +
  geom_point(aes(col= Province_State)) +
  geom_smooth(method='lm',se = FALSE) +
  scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma) +
  labs(
       x=NULL,
       y=NULL,
       title="Population Vs. N Deaths Linear model")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Conclusion and Bias Identification

This has been a short introduction to the Johns Hopkins COVID Deaths data set and some key statistics and takeaways from the data. In general, population is a factor when determining the number of COVID deaths you would expect to see in a given state. There are some exceptions to this rule, however, as you would expect population influences number of COVID deaths.

In regards to potential bias, as someone who lives on the west coast, I have not experienced what many others around the United States have expereinced in regards to the Coronavirus Pandemic. There may be factors that I am missing that should be taken into account when modeling the number of COVID deaths by State.

Session Info

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
## [1] scales_1.1.1      data.table_1.14.0 magrittr_2.0.1    ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9       pillar_1.6.2    compiler_4.1.1  tools_4.1.1
##  [5] digest_0.6.27   evaluate_0.14   lifecycle_1.0.0 tibble_3.1.4
##  [9] gtable_0.3.0    nlme_3.1-152    lattice_0.20-44 mgcv_1.8-36
## [13] pkgconfig_2.0.3 rlang_0.4.11    Matrix_1.3-4    yaml_2.2.1
## [17] xfun_0.25       fastmap_1.1.0   withr_2.4.2     stringr_1.4.0
## [21] knitr_1.34      vctrs_0.3.8     grid_4.1.1      glue_1.4.2
## [25] R6_2.5.1        fansi_0.5.0     rmarkdown_2.10  farver_2.1.0
## [29] ellipsis_0.3.2  htmltools_0.5.2 splines_4.1.1   colorspace_2.0-2
## [33] labeling_0.4.2  utf8_1.2.2      stringi_1.7.4   munsell_0.5.0
## [37] crayon_1.4.1
```