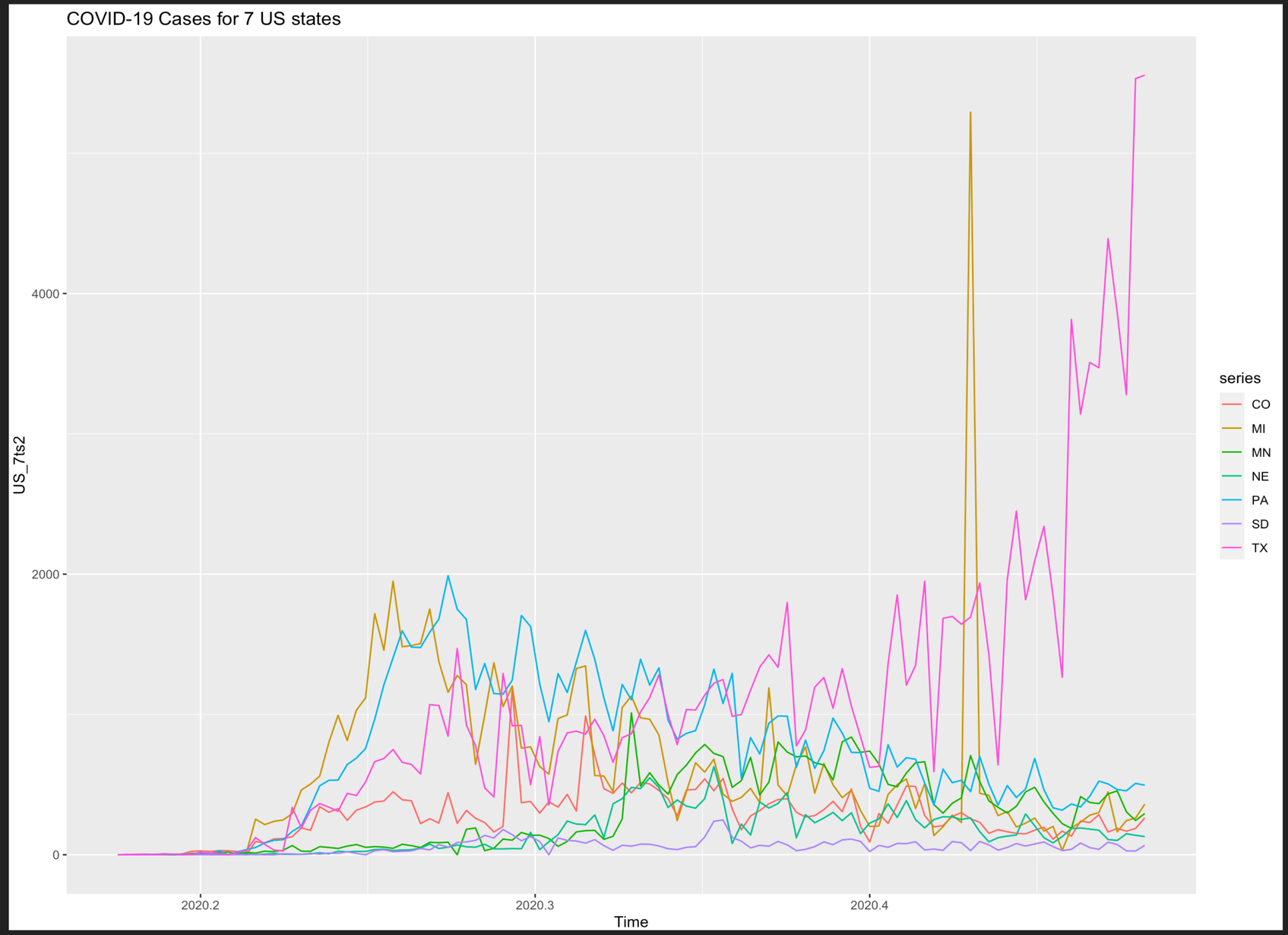
FORECASTING COVID-19 CASES AND DEATHS USING TIME SERIES ANALYSIS IN R

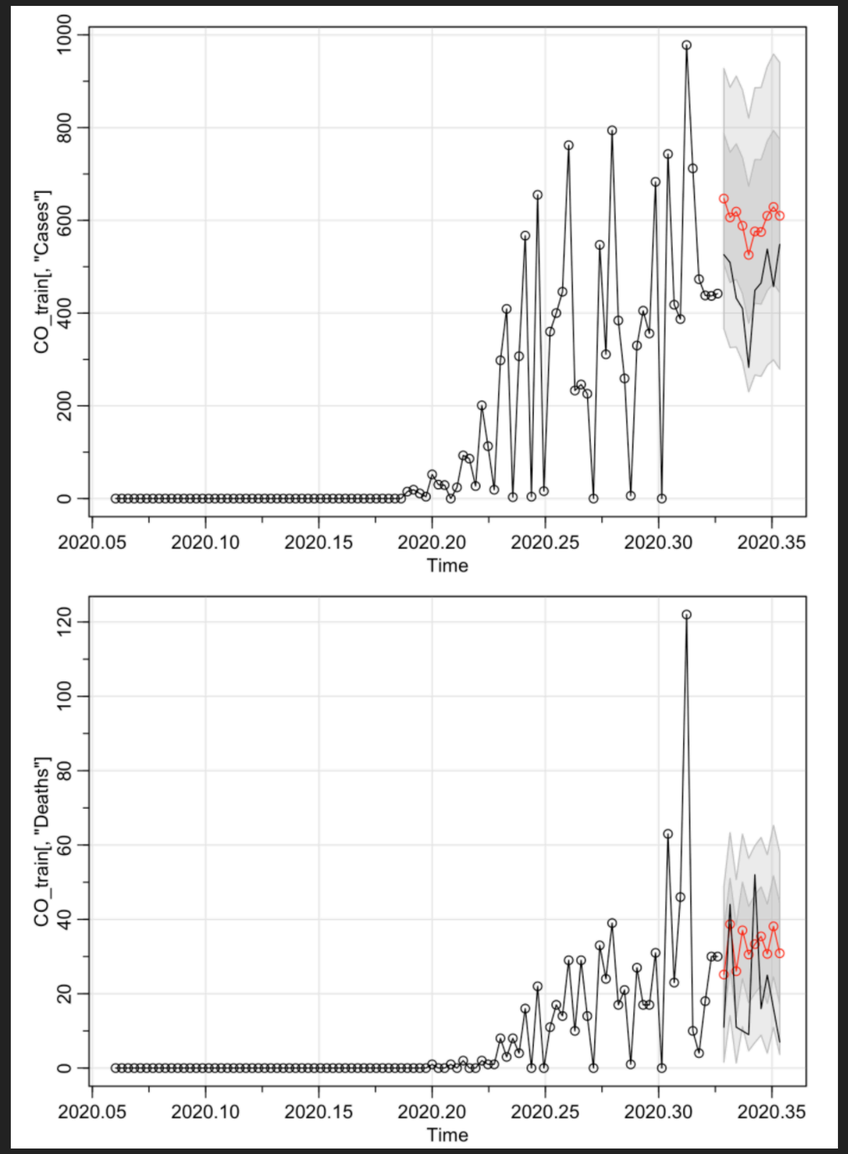
REGINA DUVAL,
SUMMER I 2020

MSDS692 PRACTICUM



THE PROBLEM

- The US started tracking COVID-19 cases and deaths on January 23, 2020, the date of the first recognized case in Washington state
- Case count and death count are available at the country, state, and county level
- Many things are unknown about the novel coronavirus, including details of treatment, spread, and how long the virus will be a global pandemic, making it hard to use the normal epidemiologic models
- The goal of this project was to use time series analysis in R to predict future cases and deaths for the US, a few key states, and several counties in my home state of Minnesota
- Constraints = lack of knowledge about the virus and short time frame of data

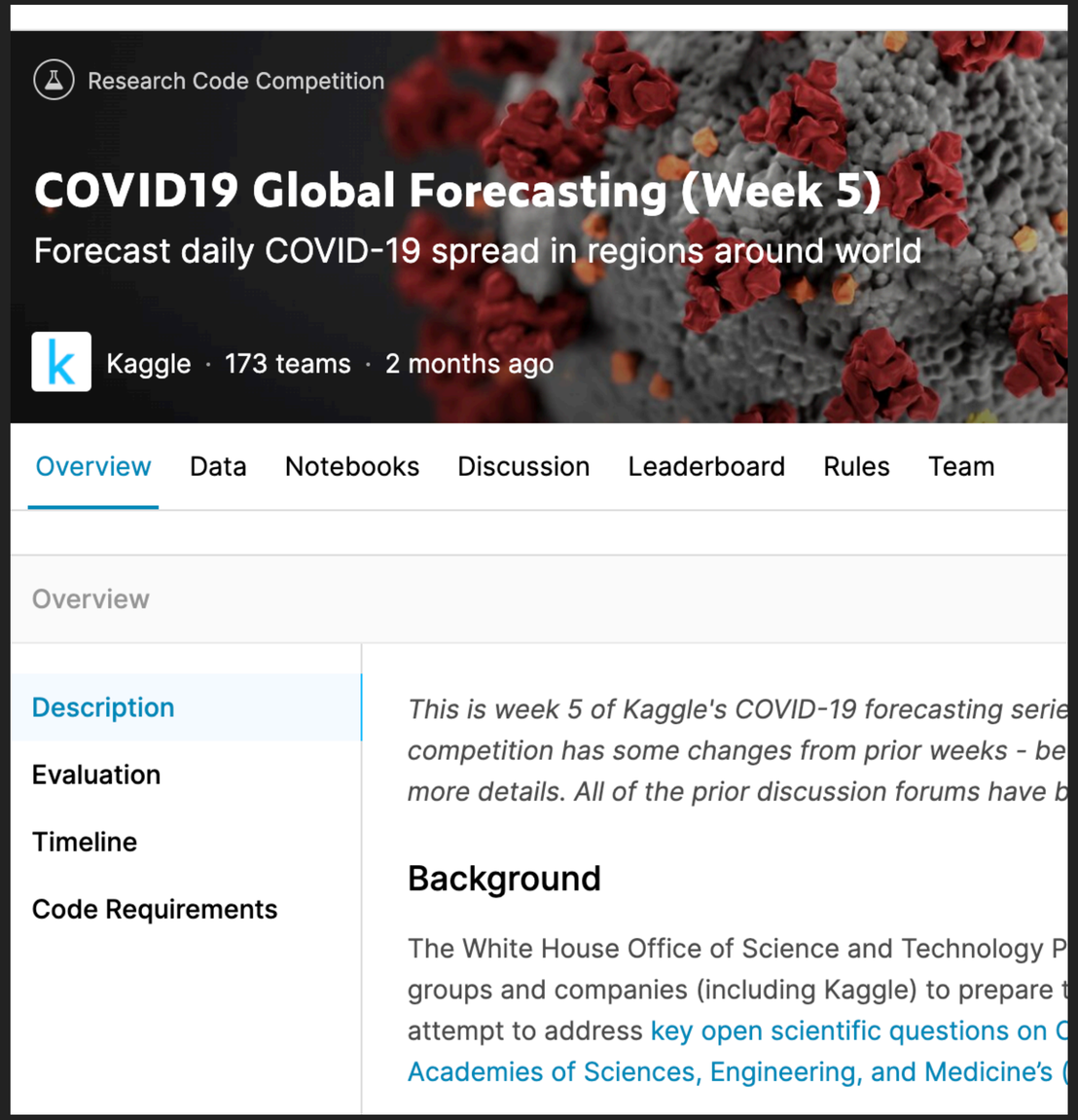


METHODOLOGY

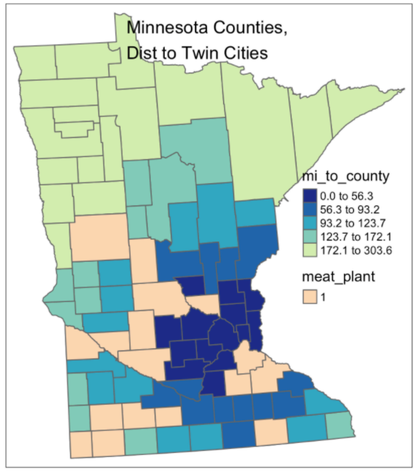
- Univariate time series methods like simple exponential smoothing, trend methods, and ARIMA were explored
- Multivariate methods such as Vector Autoregression was used to study groups of states and counties together, looking for relationships
- Linear regression was used to identify the impact of constants like population and hot spot indicators on COVID spread

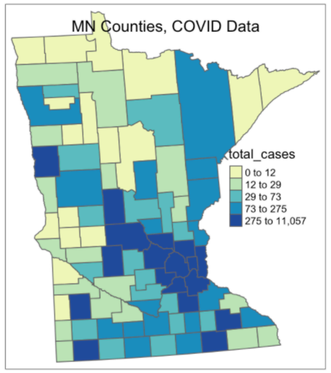
THE DATA

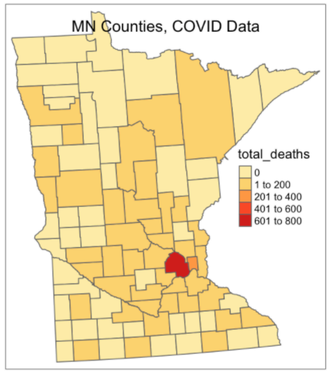
- The primary dataset was available via the Kaggle COVID Global Forecasting (Week 5) competition (“kaggle”)
- The secondary dataset contained county level population, distance data, and hot spot indicators compiled from various online sources (“county_data”)
- To update case and death count, data sets were pulled from the USAfacts website which updates cases and deaths daily down to the county level

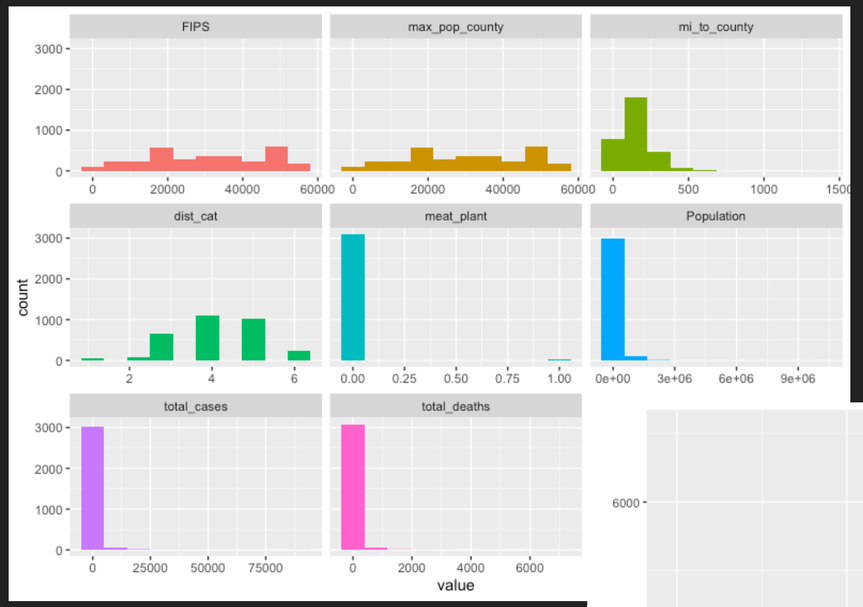


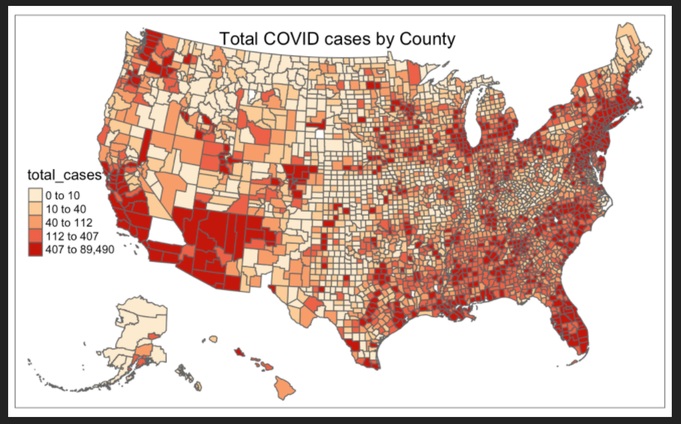
EXPLORATORY DATA ANALYSIS

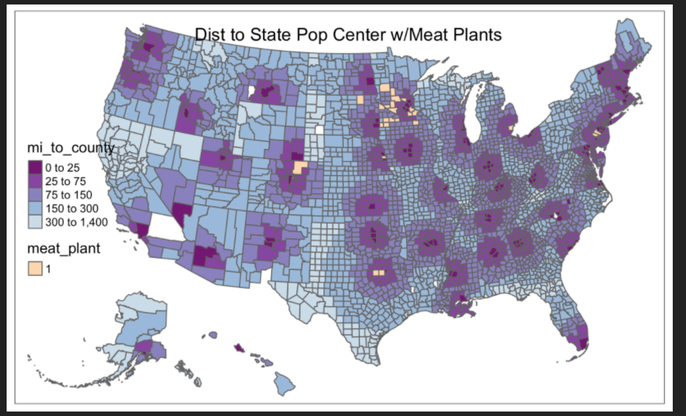


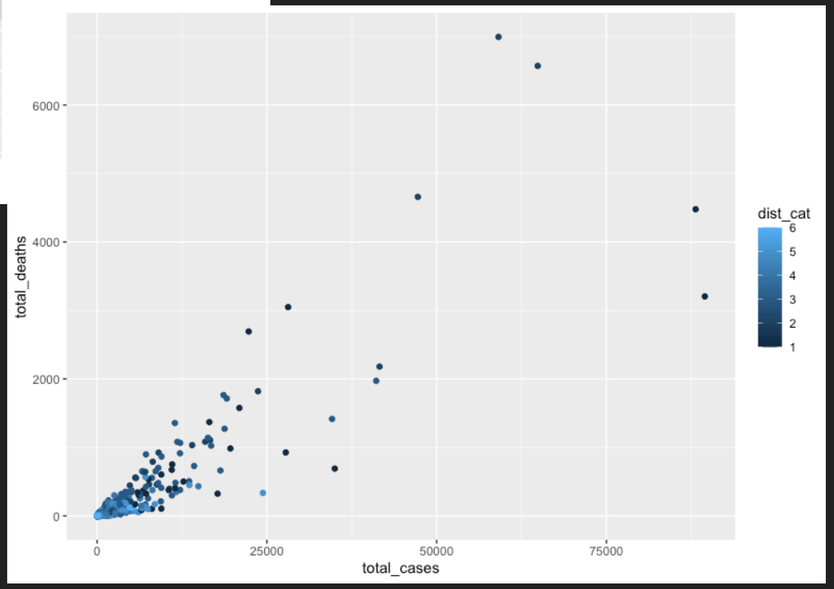






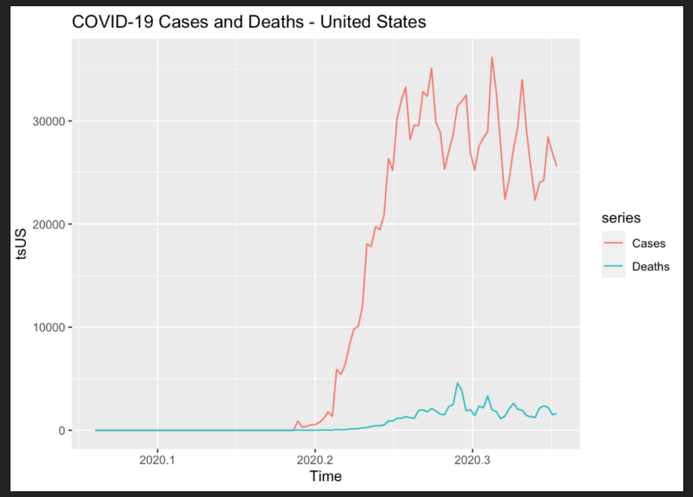


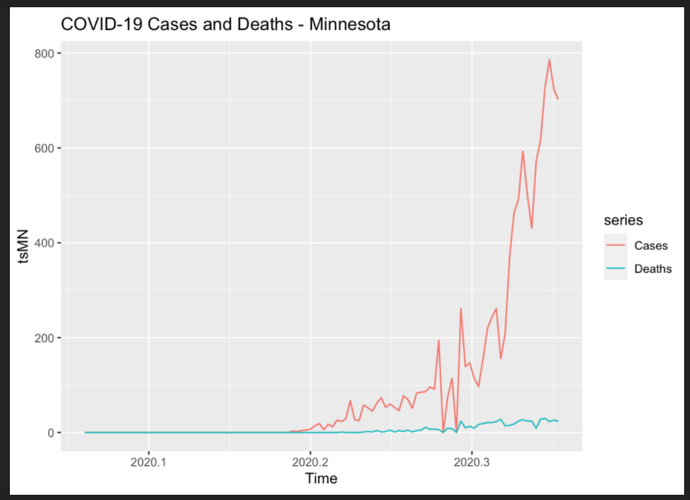


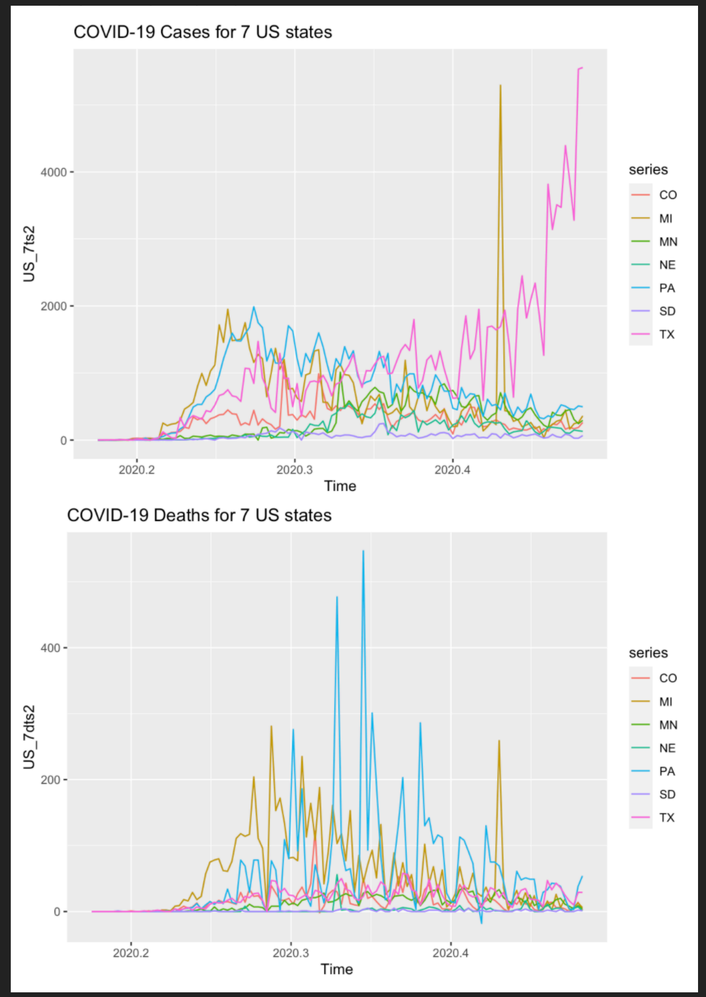


EXPLORATORY DATA ANALYSIS

- Deaths are caused by extreme cases, so we know that these two variables are related
- Both appear to have strong cyclical movement (peaks and troughs), but this could be related to reporting
- Deaths appear to be on the decline while cases are still rising or steady nationwide





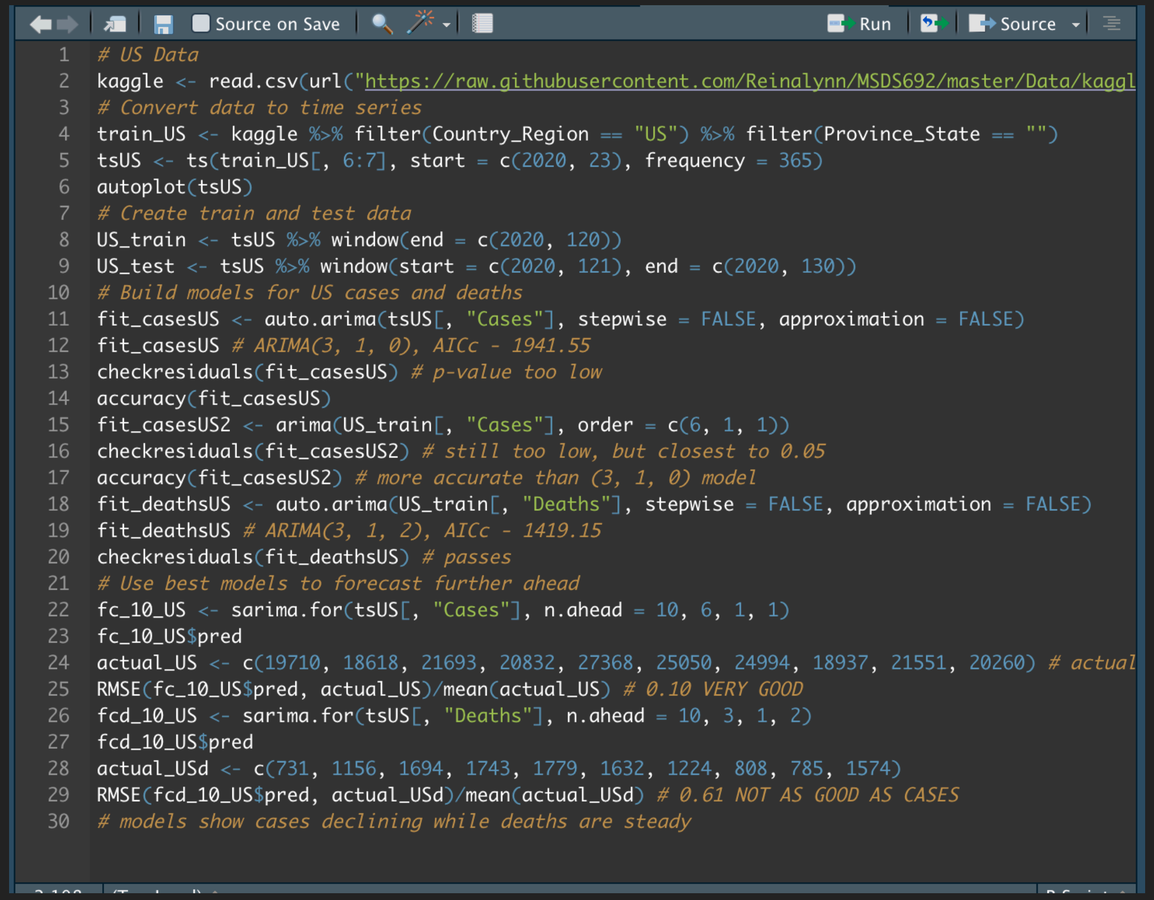


UNIVARIATE MODELS

- Multiple linear regression model showed that constants such as county population, distance from a population center, and the presence of hot spots like meat plants were significant, but they were not sufficient to predict cases or deaths well (adjusted R-squared values of 0.44 and 0.45)
- Univariate time series models like SES (simple exponential smoothing) and holt trend models (differenced, damped, diff and damped) were too simplistic
- ARIMA models were most effective at predicting both cases and deaths and providing interpretable forecasts

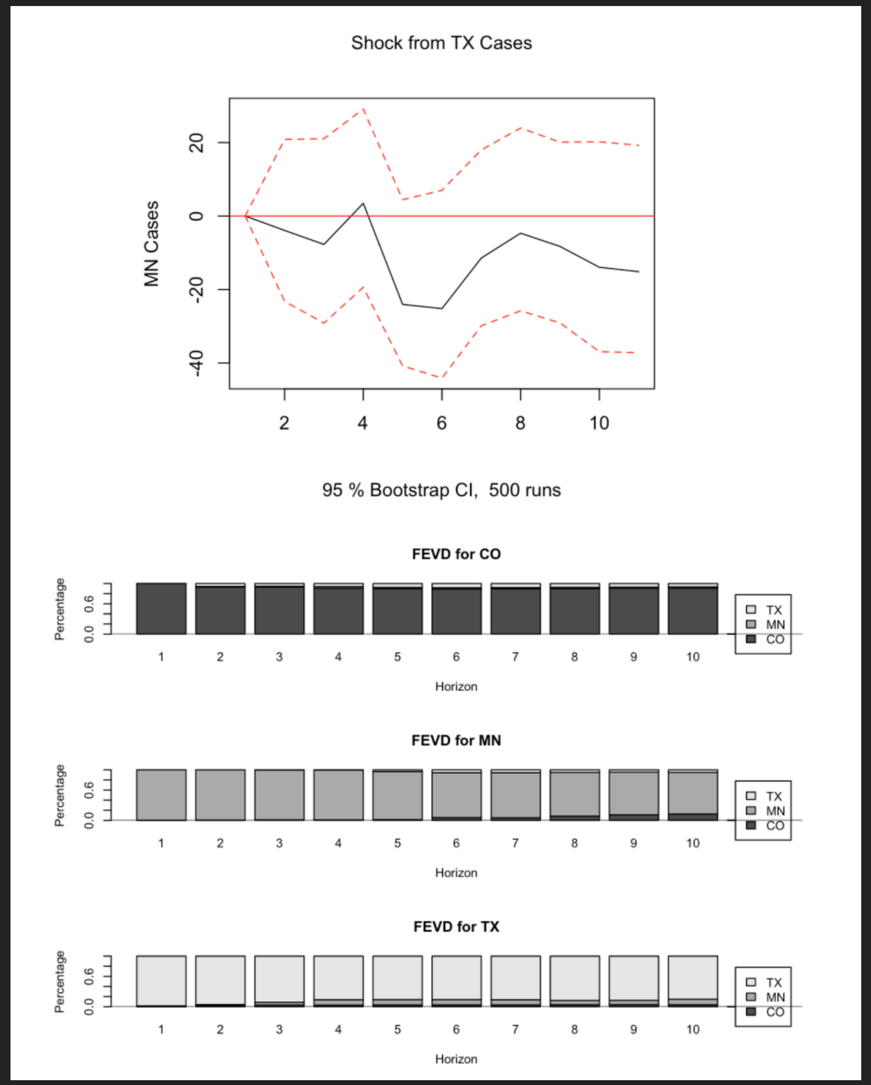
BUILDING ARIMA MODELS

- Convert data to time series using ts() function
- Plot ts using autoplot()
- Use auto.arima() to find base model
- Check the residuals to see if the model is adequate (desired outcome is p-value greater than 0.05)
- Tweak model as necessary
- Use optimal model to forecast and test against actual data




```r
# US Data
kaggle <- read.csv(url("https://raw.githubusercontent.com/Reinalynn/MSDS692/master/Data/kaggl
# Convert data to time series
train_US <- kaggle %>% filter(Country_Region == "US") %>% filter(Province_State == "")
tsUS <- ts(train_US[, 6:7], start = c(2020, 23), frequency = 365)
autoplot(tsUS)
# Create train and test data
US_train <- tsUS %>% window(end = c(2020, 120))
US_test <- tsUS %>% window(start = c(2020, 121), end = c(2020, 130))
# Build models for US cases and deaths
fit_casesUS <- auto.arima(tsUS[, "Cases"], stepwise = FALSE, approximation = FALSE)
fit_casesUS # ARIMA(3, 1, 0), AICc - 1941.55
checkresiduals(fit_casesUS) # p-value too low
accuracy(fit_casesUS)
fit_casesUS2 <- arima(US_train[, "Cases"], order = c(6, 1, 1))
checkresiduals(fit_casesUS2) # still too low, but closest to 0.05
accuracy(fit_casesUS2) # more accurate than (3, 1, 0) model
fit_deathsUS <- auto.arima(US_train[, "Deaths"], stepwise = FALSE, approximation = FALSE)
fit_deathsUS # ARIMA(3, 1, 2), AICc - 1419.15
checkresiduals(fit_deathsUS) # passes
# Use best models to forecast further ahead
fc_10_US <- sarima.for(tsUS[, "Cases"], n.ahead = 10, 6, 1, 1)
fc_10_US$pred
actual_US <- c(19710, 18618, 21693, 20832, 27368, 25050, 24994, 18937, 21551, 20260) # actual
RMSE(fc_10_US$pred, actual_US)/mean(actual_US) # 0.10 VERY GOOD
fcd_10_US <- sarima.for(tsUS[, "Deaths"], n.ahead = 10, 3, 1, 2)
fcd_10_US$pred
actual_USd <- c(731, 1156, 1694, 1743, 1779, 1632, 1224, 808, 785, 1574)
RMSE(fcd_10_US$pred, actual_USd)/mean(actual_USd) # 0.61 NOT AS GOOD AS CASES
# models show cases declining while deaths are steady
```

MULTIVARIATE MODELS



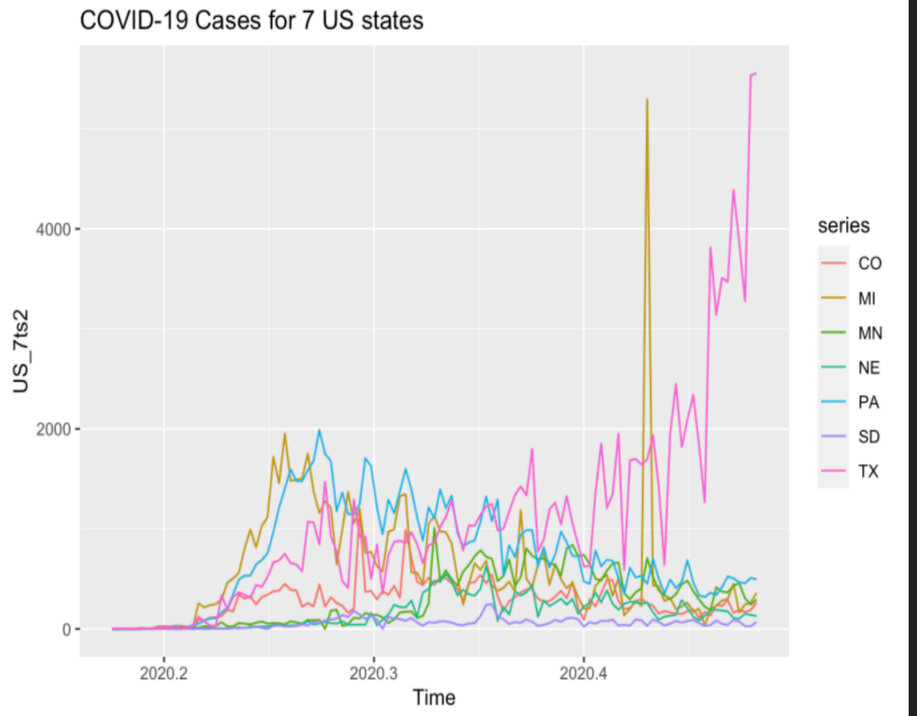
- Used in real world data when variables affect each other (instead of a unidirectional relationship)
- Vector Autoregression models treat all variables symmetrically
- All variables are endogenous instead of exogenous (note: ARIMA models do allow for external regressors but they are not very sensitive to these variables)
- VAR models can show the impact of one time series on another time series

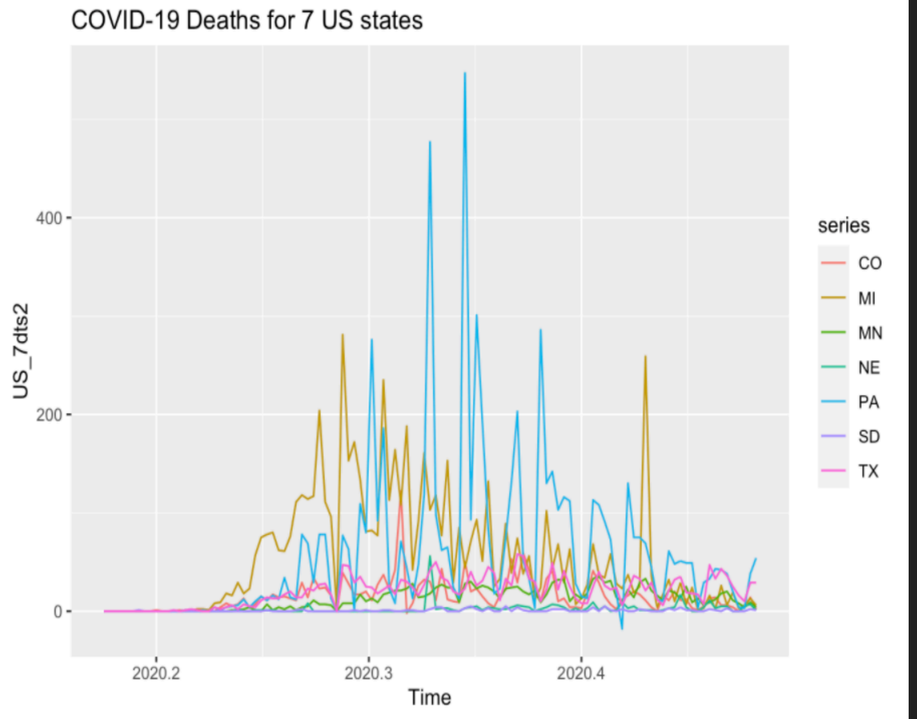
COMPARING ARIMA AND VARS

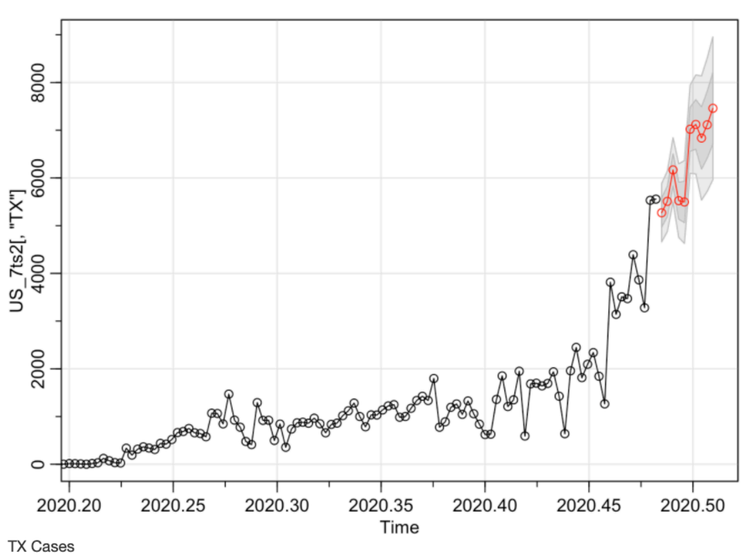
- Complexity - ARIMA is much easier to build and interpret
- Results - ARIMA appears to be more accurate, at least with COVID cases and deaths data at the state level, forecasting 10 days out
- Utility - ARIMA models require fewer variables and thus easier to maintain but are limited in the information they provide (forecasts only, VAR has secondary uses)

Minnesota	Column1	Column2
Variable	Model	RMSE (normalized)
Cases	ARIMA	0.19
Cases	VAR	0.55
Deaths	ARIMA	0.31
Deaths	VAR	0.64

Colorado	Column1	Column2
Variable	Model	RMSE (normalized)
Cases	ARIMA	0.32
Cases	VAR	0.41
Deaths	ARIMA	0.93
Deaths	VAR	2.45

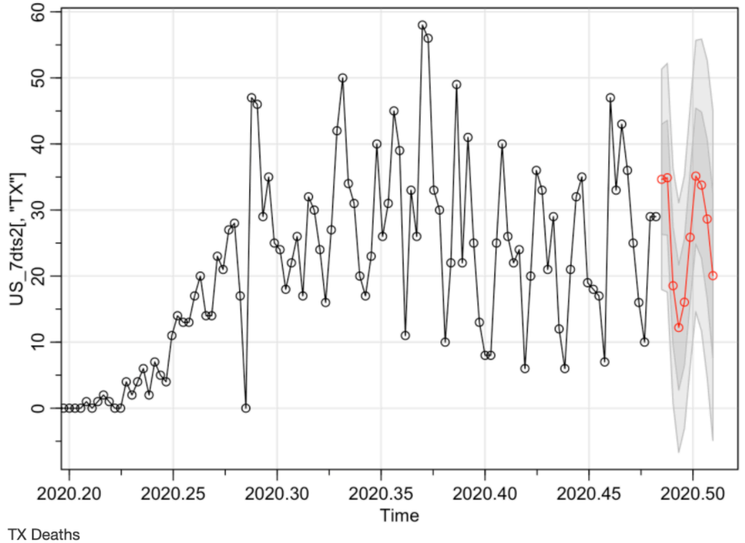






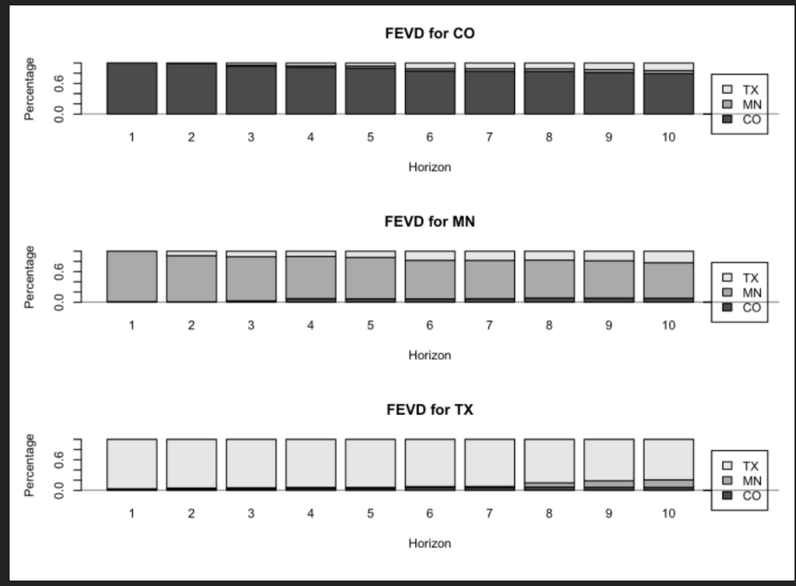
TX Cases

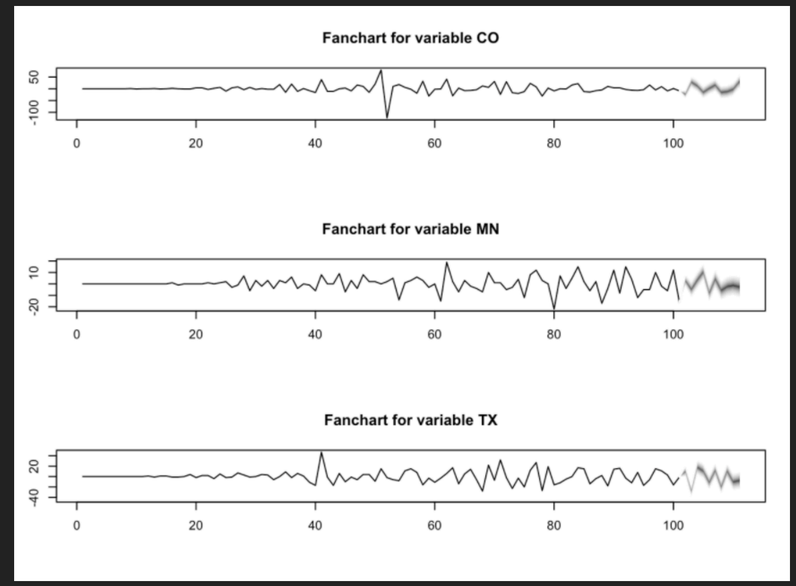
```
## Time Series:
## Start = c(2020, 178)
## End = c(2020, 187)
## Frequency = 365
##  [1] 5270.443 5509.166 6168.777 5524.268 5496.354 7020.696 7121.484 6836.454
##  [9] 7114.152 7460.074
```



TX Deaths

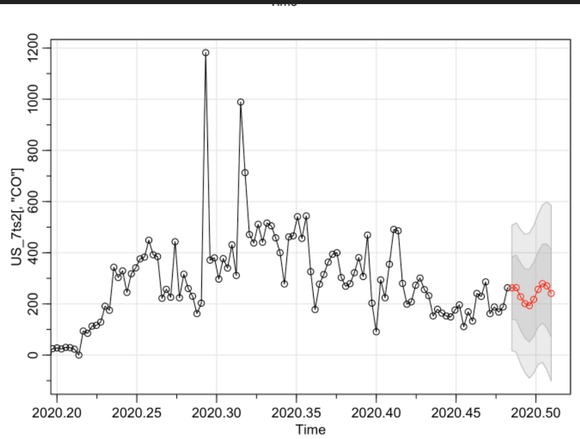
```
## Time Series:
## Start = c(2020, 178)
## End = c(2020, 187)
## Frequency = 365
##  [1] 34.64500 34.88198 18.54188 12.20248 16.04625 25.87062 35.13658 33.77441
##  [9] 28.63990 20.06938
```





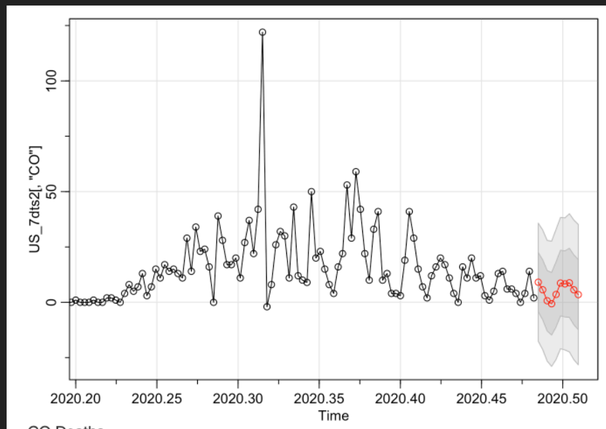
FINAL THOUGHTS

- Is time series analysis in R effective for forecasting COVID cases and deaths?
  - Decide for yourself!
  - Using ARIMA, I predicted the cases and deaths for Colorado from June 25, 2020 to July 4, 2020
  - Compare to actual values on your own or visit my GitHub to find forecasts for other states
  - Github link: https://github.com/Reinalynn/MSDS692



CO Cases

```
## Time Series:
## Start = c(2020, 178)
## End = c(2020, 187)
## Frequency = 365
##  [1] 262.5499 263.5916 228.1819 201.2255 193.2711 217.5006 256.7161 279.1208
##  [9] 270.5932 241.0802
```



CO Deaths

```
## Time Series:
## Start = c(2020, 178)
## End = c(2020, 187)
## Frequency = 365
##  [1]  9.0875560  5.6590338  0.7176019 -0.6552971  3.5564503  8.6917856
##  [7]  8.3167922  8.7985093  5.6477377  3.4760598
```

REFERENCES

https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
https://afit-r.github.io/ts_exploration
https://arxiv.org/pdf/2004.07859.pdf
https://blogs.oracle.com/datascience/introduction-to-forecasting-with-arima-in-r
https://bookdown.org/singh_pratap_tejendra/intro_time_series_r/multivariate-ts-analysis.html
https://www.cdc.gov/mmwr/volumes/69/wr/mm6918e3.htm?s_cid=mm6918e3_x
https://www.census.gov/acs/www/data/data-tables-and-tools/
https://cran.r-project.org/web/packages/tmap/tmap.pdf
https://cran.r-project.org/web/packages/vars/vars.pdf
https://data.census.gov/cedsci/table?g=0400000US27.050000&layer=VT_2018_040_00_PY_D1&y=2018&tid=ACSST5Y2018.S1603&t=Language%20Spoken%20at%20Home&vintage=2018
https://data.nber.org/data/county-distance-database.html
https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/
https://earth.esa.int/documents/10174/1573054/Factors_that_have_an_influence_on_time_series.pdf
https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
https://www.ewg.org/news-and-analysis/2020/05/ewg-map-counties-meatpacking-plants-report-twice-national-average-rate
http://www.healthdata.org/us-county-profiles
https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc451.htm
https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=558
https://www.kaggle.com/c/covid19-global-forecasting-week-5/data
https://www.kaggle.com/imdevskp/corona-virus-report
https://kevinkotze.github.io/ts-7-tut/
https://machinelearningmastery.com/time-series-forecasting/
https://www.medrxiv.org/content/10.1101/2020.04.17.20069237v1
https://www.medrxiv.org/content/10.1101/2020.04.17.20069237v1.full.pdf
https://otexts.com/fpp2/
https://orca-mwe.cf.ac.uk/62788/1/'Horses%20for%20Courses'%20in%20demand%20forecasting.pdf
http://past.rinfinance.com/agenda/2013/talk/RueyTsay.pdf
http://people.duke.edu/~rnau/411flow.gif
https://www.r-econometrics.com/timeseries/varintro/
https://rstudio-pubs-static.s3.amazonaws.com/274358_9fbc895fea2b443aaa60ad1a75c75687.html
https://simplemaps.com/data/us-cities
https://www.statista.com/statistics/1103185/cumulative-coronavirus-covid19-cases-number-us-by-day/
https://www.statmethods.net/advstats/timeseries.html
https://www.statmethods.net/stats/regression.html
https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783552078/1/ch01lvl1sec08/multivariate-time-series-analysis
https://thefern.org/2020/04/mapping-covid-19-in-meat-and-food-processing-plants/
https://towardsdatascience.com/top-5-r-resources-on-covid-19-coronavirus-1d4c8df6d85f
https://www.tutorialspoint.com/r/r_time_series_analysis.htm
https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
https://www.usnews.com/news/healthiest-communities/articles/2020-05-01/cdc-nearly-5-000-meat-plant-workers-infected-by-coronavirus
https://www.wired.com/story/why-meatpacking-plants-have-become-covid-19-hot-spots/
http://zevross.com/blog/2018/10/02/creating-beautiful-demographic-maps-in-r-with-the-tidycensus-and-tmap-packages/