W271 Assignment 8

```
library(tidyverse)
library(magrittr)
library(patchwork)

library(lubridate)

library(tsibble)
library(feasts)
library(forecast)

library(sandwich)
library(lmtest)

library(nycflights13)
library(blsR)
```

```
theme_set(theme_minimal())
```

(14 points total) Question-1: Is Unemployment an Autoregressive or a Moving Average Process?

You did work in a previous homework to produce a data pipeline that pulled the unemployment rate from official BLS sources. Reuse that pipeline to answer this final question in the homework:

> "Are unemployment claims in the US an autoregressive, or a moving average process?

(1 point) Part-1: Why is the distinction important?

Why is it important to know whether a process is a AR or an MA (or a combination of the two) process? What changes in the ways that you would talk about the process, what changes in the ways that you would fit a model to the process, and what changes with how you would produce a forecast for this process?

(1 point) Part-2: Pull in (and clean up) your data pipeline.

In the previous homework, you built a data pipeline to draw data from the BLS. We are asking you to re-use, and if you think it is possible, to improve the code that you wrote for this pipeline in the previous homework.

- Are there places where you took "shortcuts" that could be more fully developed?
- Are the processes that could be made more modular, or better documented so that they are easier for you to understand what they are doing? You've been away from the code that you wrote for a few weeks, and so it might feel like "discovering" the code of a *mad-person* (Who even wrote this???)

(5 points) Part-3: Conduct an EDA of the data and comment on what you see.

We have presented four **core** plots that are a part of the EDA for time-serires data. Produce each of these plots, and comment on what you see.

(1 point) Part-4: Make a Call

Based on what you have plotted and written down in the previous section, would you say that the unemployment rate is an AR, MA or a mix of the two?

(6 points total) Part-5: Estimate a model

Report the best-fitting parameters from the best-fitting model, and then describe what your model is telling you. In this description, you should:

- (1 point) State, and justify your model selection criteria.
- (1 point) Interpret the model selection criteria in context of the other models that you also fitted.
- (2 points) Interpret the coefficients of the model that you have estimated.
- (2 points) Produce and interpret the model diagnostic plots to evaluate how well your best-fitting model is performing.
- (1 (optional) point) If, after fitting the models, and interpreting their diagnostics plots, you determine that the model is doing poorly – for example, you notice that the residuals are not following a white-noise process – then, make a note of the initial model that you fitted then propose a change to the data or the model in order to make the model fit better. If you take this action, you should focus your interpretation of the model's coefficients on the model that you think does the best job, which might be the model after some form of variable transformation.

(14 Points Total) Question-2: COVID-19

The United States Centers for Disease Control maintains the authoritative dataset of confirmed and probable COVID-19 cases.

- This data is described on this page [link].
- The data is made available via an API link on this page as well.

(1 point) Part-1: Access Data

Use the public API to download the CDC COVID-19 data and store in a useful dataframe. A useful dataframe:

- Should have useful variable names;
- Should be in a format that can be used for time series modeling;

- Should have appropriate time indexes (and possibly keys) set; but,
- At this point, should not have derivative features mutated onto the data frame; nor,
- Should it be aggregated or summarized.

(5 points) Part-2: Pick a State and Produce a Model

1. Choose a state that is not California (we are putting this criteria in so that we see many different states chosen);
2. Produce a 7-day, backward smoother of the total case rate; then,

3. Produce a model of COVID cases in that state. This should include:

- Conducting a full EDA and description of the data that you observe
- Estimating a model (either AR or MA) that you believe is appropriate after conducting your EDA
- Evaluating the model performance through diagnostic plots

(5 points) Part-3: Produce a Nationwide Model

1. Aggregate the state-day data into nationwide-day level data;

2. Produce a 7-day, backward smoother of the total case rate; then,
3. Produce a model of COVID cases across the US. Like the state model, this should include:

- Conducting a full EDA and description of the data that you observe
- Estimating a model (either AR or MA) that you believe is appropriate after conducting your EDA
- Evaluating the model performance through diagnostic plots

(3 points) Part-4: Write a few paragraphs about this modeling task

The nationwide model that you just produced contains much **more** data than went into your state-level model. Does this make it a better model? Why or why not?

Without a requirement that you actually produce the model that you propose: If you were trying to produce a nationwide model, knowing: (a) what you know about the state model that you fit; (b) what you know about the nationwide model that you fit; and (c) what you, as a citizen of this world who has lived through these past years: *propose a modeling strategy you think will produce the best nationwide forecasting model.*

This could be, for example, the nationwide model that you have fitted above. Or, you might propose some other forms of data aggregation before modeling, or model aggregation but not data aggregation. In writing about your strategy, justify choices that you are making.

Our goal with this question is to ask that you not only conduct the narrow technical work, but also that you do the higher-level reasoning about the technical work. We would like you to write in full paragraphs, rather than bullet points that address specific parts of the prompt above.