**Summer 2020 Time Series Project**

It seems that the Covid pandemic will be an issue the World will deal with for at least the near future. The pandemic has shown to hold serious implications for the economy, education and mental and physical healthcare. It stands to reason that the ability to forecast Covid cases could provide society with the information it needs to, among other things, apply valuable resources (financial, medical, etc) in a timely, efficient and effective manner. This project focuses on providing these invaluable forecasts through the use of the time series models and methods we have studied this semester.

**Goal 1: Data collection**

As we all know, much of the data scientist’s (time series analyst’s) initial work on a project is gathering and cleaning the correct, or at least useful, data. Currently it seems that the media and the government prefer to report the total number of cases and deaths per time period and geographic region. Lately however, we have been hearing more about the “positive percentage”. Research this statistic and decide which measure you feel is more telling about the current severity of the pandemic and the progress that the nation (or certain locals) are making in curbing it. Report specifically on:

* 1. What the “positive percentage” statistic is.
  2. Where you found raw data and information on the total number of cases, total number of deaths and the “positive percentage”.
  3. Provide an exploratory data analysis using data for your given state and the nation as a whole. (Every group will be assigned their own State although each group will also analyze the nation as a whole.). The EDA should simply include a slide or two with time series plots, answers to the questions above and any other tables or charts that help convey the information you have gathered and that you think will be important in providing the forecasts.

**Goal 2: Univariate Analysis**

This analysis will be focused on providing both short- and long-term forecasts for your given state and the nation as a whole *using only the single time series* (no additional explanatory variables). The short-term forecasts will be forecasts for the next 7 days while the long-term forecasts will be for the next 3 months. Note: you may use different models for the short- and long-term forecasts. This analysis should address:

* 1. Stationarity versus Non-Stationarity and any concerns you may have about the integrity of the data (if any.)
  2. Any non-stationary modeling of the data (differencing, etc.) if any.
  3. Model IDing of any stationary models.
  4. Model Building
  5. Comparing / Assessing / Evaluating Candidate Models (ASE, AIC, etc.)
     1. At least 1 ARMA / ARIMA type model
     2. At least 1 MLP / RNN (neural network) type model
     3. At least 1 ensemble model (this can be a simple average of the forecasts of the models above.)
  6. Effectively producing, presenting and explaining your short- and long-term forecasts (including confidence intervals).

**Goal 3: Multivariate Analysis**

This analysis is very similar to the univariate analysis although you are now able to add explanatory variables to the model. This analysis will be focused on providing both short- and long-term forecasts for your given state and the nation as a whole using the time series realization and at least 1 exogenous / explanatory variable of your choosing. The short-term forecasts will be forecasts for the next 7 days while the long-term forecasts will be for the next 3 months. Note: you may use different models for the short- and long-term forecasts. This analysis (one for the short term and one for the long-term forecasts) should address:

1. Stationarity versus Non-Stationarity and any concerns you may have about the integrity of the data (if any.)
2. Any non-stationary modeling of the data (differencing, etc.) if any.
3. Model IDing of any stationary models.
4. Model Building
5. Comparing / Assessing / Evaluating Candidate Models (ASE, AIC, etc.)
   * 1. At least 1 VAR and / or MLR with correlated errors type model
     2. At least 1 MLP / RNN (neural network) type model
     3. At least 1 ensemble model (this can be a simple average of the forecasts of the models above.)
6. Effectively producing, presenting and explaining your short- and long-term forecasts for both your given state and the nation as a whole (including confidence intervals).

**Groups:**

This is either a group project of 2 or less or you may complete it alone (group of 1). To be clear, groups must be 2 members or less, no groups of three or more. Communication between groups is encouraged although it is assumed that the analysis, write up (RMD or equivalent) and presentation materials are performed by the members of the group. Also, you may partner with someone from the other class as well if you like.

**Rubric:**

Initial Project Presentation: 5%

EDA see below

Initial Project Documentation: 5%

EDA see below

Final Project Presentation: 40%

* Communication and presentation of your findings are critical to being a successful data scientist. You will be graded individually on:
  + Presenting at least 3 minutes
  + Voice inflection
  + Slide Organization / Content
  + Visualization (including clear, visually appealing and labeled plots).
  + Composure: This will include **not reading** off of the slides and smoothness of delivery.
  + Pace: Not going a second over 8 minutes. Your client is very strict on this point.

Final Project Documentation 50%

* Knit RMD: 10% (Same grade for both team members.)
  + Well organized and documented RMD knit to pdf or html.
* Results / Analysis: 40% (Same Grade for both team members.) (Double check the entire project, not just your part.)
  + Correct Interpretation
  + Creating Useful Models
  + Performing a Complete Analysis: Model ID, Model Building, Forecasting, Cross Validation

**Deliverables:**

**EDA:**

**Saturday July 25th at 11:59pm**

Deliverable:

1. 4-minute (max) YouTube video: (As a group, each student must present for at least a minute.)
2. Identify yourself and your partner (if applicable). Goal 1
3. Describe the Data Set / Time Series (Who, What, When, Where, Why and How) Goal 1
4. Stationary / Non-Stationary Goal 2
5. ACFs and Spectral Densities just to explore Goal 2
6. At least 2 candidate ARMA / ARIMA models (univariate) Goal 2
   1. The models in factored form or at least separate the stationary and non-

stationary factors with standard deviation or variance of the white noise.

* 1. AIC
  2. ASE
  3. Visualization of Forecasts (Short and Long Term / State and National)

1. Strategy / Plans for the rest of the analysis.
2. Submit your slides to 2DS under *Initial Project Documentation* and make sure your video URL is on the Google Doc.

**Final Project Documentation and Presentation:**

**Saturday, August 8 at 11:59pm**

1. Addressing a summary of Goal 1 and fully addressing Goal 2 and 3.
2. PowerPoint Presentation (1 per group)
3. RMD or Jupyter Notebook (1 per group)
4. 8-minute (or less) You-Tube video presenting your PowerPoint (or Presi or LaTex) slides. (1 per group but each student must speak for at least 3 minutes.)
5. Post that video to you-Tube and the (private) link to the Google-Doc and submit your ppt and Rmd File (or Jupyter notebook) to 2DS. Please leave the link on the Google Doc for a week so others can learn from your presentation. Please check out at least 3 of your peer’s presentations and please watch your own presentation as well. It is often very useful (although always a bit awkward for me at least ;) to watch yourself present!
6. Submission
   1. The link to your you-tube presentation on the Google-Doc
   2. your pptx (or slides in whatever form (pdf, Prezi, etc.) (Put in “Final Project Presentation.”)
   3. an R markdown (RMD) or Jupyter notebook or equivalent (Put in “Final Project Documentation.”) this file should contain all of your EDA, modeling and forecasting code and be very organized and well commented.