**Time Series Project: Spring 2020**

**The project:**

1. Pick a data set that has two or more variables recorded over time (similar to the Schumway LA air quality data from Unit 12. (dataset: *lap* from package: *astsa*) (can’t use this one ☺ )

2. Select a response from the data set.

3. Be creative and come up with a scenario as to why a client would want to analyze this data and why this response is important! Or better yet, use a real problem that you are interested in!

4. Fit at least one model from ***each*** of the following four categories (provide all plots and tables needed to ID these models: acfs, spectral density, factor tables, etc.):

a. ARMA / ARIMA / ARUMA / Signal Plus Noise (univariate analysis)

b. VAR with at least one explanatory model.

c. Neural Network

d. Ensemble model using at least two of the above. (this model does not have to “beat” your

other models.

5. Compare all models with the ASE… this does not mean you have to choose the model with the lowest ASE.

6. Pick a forecast horizon based on your “problem” from part 3 above and provide the forecasts and prediction limits.

7. Create a ppt and a 7-minute video describing your analysis and steps 1 – 6 above.

8. 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!

**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.

If however, you would like to work in a group of 3, let me know and I can make a special project with some data I have.

**Intended Audience:**

Your intended audience is the head health official in the state you are researching. This is usually the Director of the Department of Public Health but may vary per state. You can assume that this person directly advises the Governor of that State on COVID-19 related topics such as the reopening schedule for businesses and schools as well as the controversial topics such as mandatory mask ordinances, social distancing and even complete lockdowns. You can assume this person has taken an introductory course in statistics and another in time series and are familiar with ACFs, spectral density plots, ARMA/ARIMA models, AIC, ASE, and the tests that are associated with these models (ie Dickey-Fuller Test). However, they may be not be as familiar with VAR modeling and deep learning models as they were not as accessible when they went to school. Your goal is not to convince them of any political or medical opinion, rather to educate them on the facts (and some methods) so they can make their own decision for the state.

**Rubric:**

Knit RMD: 20%

Well organized and documented RMD knit to pdf or html.

Results / Analysis: 40%

Correct Interpretation

Creating Useful Models

Performing a Complete Analysis: Model ID, Model Building, Forecasting, Cross Validation

Presentation: 40%

* Communication and presentation of your findings are critical to being a successful data scientist. You will be graded on:
  + Voice inflection
  + Slide Organization / Content
  + Visualization
  + Composure: This will include **not reading** off of the slides and not smoothness of delivery.
  + Pace: Not going a second over 7 minutes. Your client is very strict on this point.

**Deliverables:**

**EDA:**

**Saturday March 21 at 11:59pm**

Deliverable:

1. 3-minute YouTube video: (You can use the same slides but each team member must make the full presentation with all the slides.)
2. Identify yourself and your team (if applicable).
3. Describe Data Set / Time Series (Who, What, When, Where, Why and How)
4. Stationary / Non-Stationary
5. ACFs and Spectral Densities just to explore
6. At least 2 candidate ARMA / ARIMA models
   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 with a Practical Horizon.

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

**Final Project Documentation and Presentation:**

**Saturday, April 11 at 11:59pm**

1. Everything listed above in “**The Project**” section.
2. Submission
   1. your pptx (or slides in whatever form (pdf, Prezi, etc.) (Put in “Final Project Presentation.”)
   2. an R markdown 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.
   3. Please make sure you
   4. r link is on your pptx and the Google Doc.

**Examples:**

Link to project from Fall 2019:

<https://docs.google.com/document/d/1DbgQO551orifyjHJBXe-lORp-CFa8PQ0JuHNCwqRZe4/edit?usp=sharing>

A particularly good one is Kristen Rollins’ presentation although there are several that are great examples.