**Time Series Project: Summer 2021**

**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!

In fact… .here is an idea from an NPR radio show I listened to in January: [Fertility Rate Freakonomics](https://19743.mc.tritondigital.com/OMNY_FREAKONOMICSRADIO_FREAKONOMICSRADIO1_P/media-session/c4ef7614-613b-40e8-851a-f57f2306ce9e/d/clips/aaea4e69-af51-495e-afc9-a9760146922b/14a43378-edb2-49be-8511-ab0d000a7030/4b4069df-4211-4098-80ab-acb00025bd30/audio/direct/t1610596861/447_How_Much_Do_We_Really_Care_About_Children.mp3?t=1610596861&in_playlist=d1b9612f-bb1b-4b85-9c0c-ab0d004ab37a&utm_source=Podcast&siteplayer=true)

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 variable.

c. Neural Network (mlp)

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

other models.

5. Pick a short and long term forecast horizon based on your “problem” from part 3 and compare all models with the ASE and the rolling window ASE for both the short and long term forecasts … this does not mean you have to choose the model with the lowest ASE.

6. Provide the forecasts and prediction limits for both the short and long term forecasts.

7. Create a ppt and a 7-minute video (with Zoom or YouTube) describing your analysis (more info below).

8. Post that video to you-Tube and the (private) link (or the Zoom 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! (Note: if you use the Zoom link, make sure you make it public so that I and your peers don’t need a password.)

**Groups:**

This is an individual project. You are encouraged to help each other out, but all analysis should be produced and presented by each student. Each project should be composed of unique data / scenario (unless using the fertility suggestion above.)

**Intended Audience:**

You can assume your audience have 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.

**Deliverables:**

**EDA: (10 % of overall grade in the class.)**

**Sunday July 11th at 11:59pm**

Deliverable:

1. 3-minute YouTube (or Zoom) video:
2. Identify yourself.
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. Rolling Window ASE
  4. Visualization of Forecasts for both the short- and long-term Horizons.

1. Strategy / Plans for the rest of the analysis.
2. Submit your slides to 2DS
3. Submit Rmd file or Jupyter Notebook
4. Submit Knit Rmd File or Jupyter Notebook as a pdf, html or docx.
5. Make sure your video URL is on the Google Doc.

**Google Doc Link:** [**https://docs.google.com/document/d/1JZTmpdhSH466Ih491WXWrWoQoljrpfyVxAUkXluLiOk/edit?usp=sharing**](https://docs.google.com/document/d/1JZTmpdhSH466Ih491WXWrWoQoljrpfyVxAUkXluLiOk/edit?usp=sharing)

**Final Project Documentation and Presentation: (20% of overall grade in the class.)**

**Saturday, July 31st 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. Knit rmd file in pdf, docx or html.
   4. Please make sure your link is on your pptx and the Google Doc.
3. Presentation: 7 minutes.

**Rubric for both EDA and Final Presentation;**

Knit RMD: 20%

Well organized and documented RMD knit to pdf or html or docx.

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
  + Animation
  + Composure: This will include **not reading** off of the slides and smoothness of delivery.
  + Pace: Not going a second over time. (3 min for EDA and 7 min for Final Presentation.) Your client is very strict on this point.

**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.