Final Project

Submissions
Each group must submit

- **Project topic submission 4/3**
- **Project submission (.rmd and .html) 4/25**

In lieu of a final exam we will have a team-based final project that will allow you to further practice the concepts and methods learned in this class on a real-world dataset. Additionally, this project will help build your portfolio which has become a common request among employers.

Project
The goal of this project is to apply the data wrangling, exploration and forecasting methodologies learned in class on a real-world time series dataset.

For this project you will be submitting an R markdown file and HTML report of the file. The code should be executable from anyone's computer so don't load data from your local hard drive.

Sections:

- Project setup
    - Install and load required libraries
    - Import dataset
- Introduction
    - Describe the dataset, where it's from, who created it, the index, keys and variables (what are the forecast variable and predictor variables (if applicable)).
    - Why did you choose this dataset?
    - How can forecast on this data be leveraged to make smarter decisions?
- Data wrangling
    - Convert to tsibble
    - Deal with missing data
    - Create new variables to aid in forecasting
    - Aggregate time series to desired format for forecasting
- Exploratory analysis and visualization for the dataset
    - Visualize the dataset and comment on characteristics of time series
    - Comment on the any anomalies in the data
    - Describe trend/seasonality/cycles with supporting charts
- Model fitting
    - Split dataset into training and test sets
        - Why did you choose to split where you did?
    - Fit TSLM, ETS and ARIMA model(s)
        - Use AICc to select within model families
        - Evaluate the residuals

    - If there are predictor variables – fit a TSLM with predictor variables, Regression with ARIMA errors
        - Use AICc to select within model families
        - Evaluate the residuals
    - Fit benchmark methods
    - Extra credit – a NNETAR model or Prophet model (FPP3 Ch 12)
- Accuracy
    - Compare model performance using accuracy measures of your choice on the test dataset
    - Why did you select those metrics for assessing accuracy?
    - Select a final model – why did you select this model?
- Forecast
    - Produce a forecast for the future
    - How many steps into the future are you forecasting? Why?
    - Considerations or watchouts when putting forecast into production?

Time series data

Requirements

- Time series data not from fpp3 or tsibbledata
- Time series should have sufficient amount of data
    - Seasonal data should have at least 4 seasons of data
    - Remember lecture 1 – the smaller the dataset the harder the forecast will be to produce

Encouraged

- Dataset should be of interest to group!

Suggestions

FRED

- https://fred.stlouisfed.org
- TONS of datasets available

http://www.timeseriesclassification.com

https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=ts&sort=nameUp&view=table

Kaggle

- https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather/data

Google's dataset search engine

- https://datasetsearch.research.google.com

Bureau of Economic Analysis

- GDP, Personal Income and Real Personal Income, Consumer Spending and Real Consumer Spending, Employment
- https://www.bea.gov/data/by-place-states-territories

Sports websites

- Attendance
- https://www.pro-football-reference.com/years/2018/attendance.htm

Misc repositories:

- https://data.world/datasets/time-series
    - Crime data, airport delays, public debt/gov finance data, air pollution, taxi

Rubric:

Section	Standard	Points Possible	EC
Project Setup	1.1 - are all required libraries installed/loaded? 1.2 - can dataset be imported to local R session 1.3 - is entire .rmd executable without errors?	30	
Introduction	2.1 - describe the dataset 2.2 - did team explain why dataset was chosen? 2.3 - how can forecast be used?	20	
Data Wrangling	3.1 - conversion to tsibble object 3.2 - properly dealing with missing data 3.3 - explanation of any data manipulation steps	20	
Exploratory Analysis	4.1 - plot of time series or sample of series, seasonality, components, autocorrelation 4.2 - observations from plots 4.3 - anomalies within the dataset that should be omitted?	30	
Model Fitting	5.1 - split of dataset into train and test, or CV stated 5.2 - fitting of TSLM, ETS, ARIMA (ARIMA errors with regression if predictors) 5.3 - AICc compared within model families 5.4 - Residuals evaluated for top model within family 5.5 - Extra credit option for additional methods	50	20
Accuracy	6.1 - Compare model performance using accuracy measures of your choice on the test dataset 6.2 - Why did you select those metrics for assessing accuracy? 6.3 - Selection of final model with rational	25	
Forecast	7.1 - Produce a forecast for the future 7.2 - How many steps into the future are you forecasting? Why? 7.3 - Considerations or watchouts when putting forecast into production?	25	
Total		200	220