Project Description (Alternate Route for Exam 4)

1. Extract 100 time-series, which should contain the following:
    a. 50 companies from the S&P 500 Companies – using List of S&P 500 companies - Wikipedia while ensuring that each of the obtained series has at least 500 observations). *You can choose to aggregate each time-series by week or by month.* To get this data into R, you will need the following:
        i. Using "rvest" to scrape the Wikipedia table (preferably from the permanent, non-changeable Wikipedia link above). The table will allow you to obtain the tickers/symbols for those companies, and capture their founding dates.
        ii. Using "tidyquant"/"quantmod" to get the data for this.
        iii. Inside your tq_get() call, you can set the *periodicity* to "daily" (default), "weekly" or "monthly" to aggregate the data differently.
    b. 15 indexes capturing the performance of main indexes and sectors: (can be obtained using "tidyquant" / "quantmod")
        i. ^DJI
        ii. ^NYA
        iii. ^GSPC
        iv. ^IXIC
        v. ^SP500-10 (Energy Sector Index)
        vi. ^SP500-15 (Materials Sector Index)
        vii. ^SP500-20 (Industrials Sector Index)
        viii. ^SP500-25 (Consumer Discretionary Sector Index)
        ix. ^SP500-30 (Consumer Staples Sector Index)
        x. ^SP500-35 (Healthcare Sector Index)
        xi. ^SP500-40 (Financial Sector Index)
        xii. ^SP500-45 (Technology Sector Index)
        xiii. ^SP500-55 (Utilities Sector Index)
        xiv. ^SP500-60 (Real Estate Sector Index)
        xv. ^N225
    c. 35 cryptocurrencies of your choice (possible tickers can be obtained from CoinMarketCap. I recommend using the cryptocompare API to extract these data (since it will not contain missing data as opposed to the Yahoo Finance API). You should use a daily time-scale.

2. Once you have all the datasets loaded, you will need to convert the data for each series into a time-series. **Data preparation:**
    a. I would recommend using the timetk::tk_ts() function for that purpose. Note that some of you had issues installing the "timetk" package, but this should be installable on the FSB computers/ remote desktop.
    b. Ensure that none of the time-series has any missing data. If you have missing data, you can pick an appropriate technique from the "imputeTS" package.
    c. For the naïve forecast, hw, and auto.arima methods, you will only need the past values of the time-series for forecasting. However, you are expected to examine at least 1-2 machine learning methods for the sake of comparison. Hence, you will need to generate features for those methods.

i. The timetk::tk_augment_timeseries_signature() function expands out the timestamp information column-wise into a machine learning feature set, adding columns of time series information to the original data frame.

ii. Common technical indicators used by stock traders can be easily computed for each time-series using the “TTR” package. Obviously, there is no need to create features based on all 50 indicators (so pick a few 5-10 based on your research).

3. For each time-series, use 80% of the data for training and use the remaining 20% for validation.

4. Methods to be used in your analysis:
   a. Naïve forecast
   b. Holt-Winters (you can optimize, but then you will have to record your alpha, beta, and gamma and use it over the entire time-series).
   c. auto.ARIMA (similar to above, once you have picked the model based on the training data, you will have to apply the obtained model over the entire time-series).
   d. autoML (see Demo Week: Time Series Machine Learning with h2o and timetk (business-science.io))
   e. Bonus points: Prophet -- prophet.pdf (r-project.org)

5. Report the validation results for the five approaches for the 100 datasets.

Deliverable: A HTML from your R Markdown capturing the entire process, with plots and/or tables capturing the results. Your results should include the MAE, RMSE, and MAPE for each dataset and method. In the HTML, you should also provide an overall result (e.g., using the RMSE metric, the naïve forecast, hw, ARIMA, autoML, and prophet were the top performing methods for 5, 20, 30, 20, and 25 datasets, respectively) – this can potentially be better captured in a plot.

Comment: If you decide to take this route, you should not take exam 04. Otherwise, the grade from exam 4 will be used for your final grade.

Due Date: 5/10/2021 at 5:00 pm eastern time