Attached to this email is a zip file entitled ‘gro_homework.zip’ - it contains six .csv files.

Parameters:

- Use Python
- Part of the task is to tackle the question with only the information provided - if this requires you to make assumptions, please make those assumptions clear in the Google Colab notebook
- Please read the full set of instructions before proceeding

Goal:

We have been given the monthly production quantity for a certain agricultural product (let’s call it *Grople syrup*, note - no relation to actual Maple Syrup) in 10 different provinces of a country between January 2015 to December 2020. This *Grople syrup* comes from a fruit. It takes a few months for the fruits to grow on the trees which bear them. It also takes a few days to extract the *syrup* from the fruits after they have been harvested.

We would like to predict the production quantity for *Grople syrup* from Jan 2021 to Dec 2021.

Data:

- `Production Quantity.csv` has 4 columns
  - `start_date`, `end_date`: start day and end day of each month between January 2015 to Dec 2020.
  - `prod`: production quantity of *Grople syrup* in tonnes at monthly frequency
  - `region_id`: A unique identifier for the 10 provinces
- `Daily Precipitation.csv`: has 4 columns
  - `start_date`, `end_date`: start day and end day at a daily frequency between January 1, 2014 to Mar 13, 2022.
  - `precip`: Precipitation quantity (in mm) at daily frequency
  - `region_id`: A unique identifier for the 10 provinces
- `Daily Soil Moisture.csv`: has 4 columns
  - `start_date`, `end_date`: start day and end day at daily frequency between January 1, 2014 to Mar 6, 2022.
  - `smos`: Soil Moisture at 5cm depth (measured by the ratio Vol/Vol) at daily frequency
  - `region_id`: A unique identifier for the 10 provinces
- `Daily Temperature.csv`: has 4 columns
  - `start_date`, `end_date`: start day and end day at daily frequency between January 1, 2014 to Mar 13, 2022.
  - `temp`: Average daily temperature on the surface of the land (in celsius) at daily frequency
  - `region_id`: A unique identifier for the 10 provinces
- `Eight Day NDVI.csv`: has 4 columns

    - `start_date`, `end_date`: start day and end day at 8-day frequency between Dec 27, 2013 to Mar 13, 2022.
    - `ndvi`: Normalized Difference Vegetation Index (NDVI is a ratio which ranges between [-1, 1] and captures the vegetation abundance of an area) at 8 day frequency between the given periods**
    - `region_id`: A unique identifier for the 10 provinces
- `predicted_production_qty.csv`: has 4 columns
    - `start_date`, `end_date`: start day and end day of each month between Jan 2021 to Dec 2021.
    - `prod`: This column needs to be filled by the candidate with their predictions of *Grople syrup*.
    - `region_id`: A unique identifier for the 10 provinces

How to submit your results

1. `predicted_production_qty.csv` is the file that should contain your end results. The csv should only contain four columns - `start_date`, `end_date`, `region_id`, and `prod` . Rename the .csv file in the following format: `<your_email_address>.csv`. That is, if your email address superstar_modeler@gro-intelligence.com, then the filename should be 'superstar_modeler@gro-intelligence.com.csv'
2. Create a Google Colab notebook (https://colab.research.google.com/#create=true) which showcases all the work you did to arrive at the result. Create a text file named `<your_email_address>.txt` . This text file should only contain a link to that Google Colab notebook. IMPORTANT NOTE: ensure that you change the default sharing settings for the notebook to be 'Anyone with the link'.
3. Upload the csv file and the txt file (ONLY 2 files) via this link. The deadline for your submission is Apr 5 11:59 PM ET. Do not edit the Google Colab file after the deadline.

Evaluation:

The model will be evaluated using

1. MAPE: Mean Absolute Percentage Error between the prediction values and ground truth
2. R2: Coefficient of determination between the prediction values and ground truth
3. Run time: of the submitted google colab notebook. Note - we will only evaluate the run time of the feature engineering and model run steps (exploratory analysis that you may choose to showcase will not be part of this evaluation)

Note: Model metrics (MAPE, R2) will get higher weights than run time when evaluating. Spend no more than 5 hours on this.

**"Normalized difference vegetation index (NDVI) is a measure of the health and abundance of vegetation cover in an area. It is derived from satellite-observed light reflected from the Earth's surface in the near-infrared (NIR) and visible red (RED) portions of the spectrum. NDVI is calculated by the equation (NIR - RED) / (NIR + RED), and values range from -1.0 to +1.0. Chlorophyll in live green plants absorbs the red portion of the electromagnetic spectrum for photosynthesis, whereas the cell structure of leaves reflects the near-infrared portion of the spectrum as it is not useful to the plant. An area with a higher density of healthy green vegetation (more chlorophyll) will have a higher NDVI value than an area with less vegetation or vegetation that is dead or in poor condition (less chlorophyll)."
https://www.usgs.gov/special-topics/remote-sensing-phenology/science/ndvi-foundation-remote-sensing-phenology