Alternative Data – GA 2049 Homework 2

Methodology Description and Gathering

- Create a folder in your Google Drive called AltDataGA.2049.**YOUR_NET_ID (change this)**
- Share with ge2068@nyu.edu (me) and Lingchao Zuo lz1060@nyu.edu
- **Create different subfolders for different homework**
- **Only one (1) submissions per group please**
- **Inform the instructor which student will be submitting the assigment**
- Create the following folder structure inside
    - In root
        - Create a README.txt
            - Your full name
            - Any comments on data features other users should be aware of when they use your data.
            - Instructions about how to run the different notebooks and look at the results
            - Instructions about any libraries
        -
    - data (put your data files here)
    - src (put your code here) – create your Jupyter notebooks here
        - libraries – include any libraries you needed to run your notebooks here and reference correctly (relative references only)
        - name your notebook file: alt_data_hw1.ipynb
- Important: Run all your code and generate the required output before submission. Ideally we should not have to run any of your code since the output to your code should be clearly visible below every cell.
- Add comments to show which part of the homework each cell belongs to and which cell has the final output.

Data

https://drive.google.com/file/d/1UEPZjw_R0Yu0b-c6uDHIO11sjC0FaELz/view?usp=sharing

Copy this data to your google drive folder.

Assignment

1. The goal is to impute the missing transactions from the data
2. First, aggregate the data on the following dimensions. There should be ~20k rows when you are done.
    a. Ourmonth

    b. Store
3. Use at the total sales (sum_dollar) amount as your key indicator.
4. Remove outliers – how to do this is up to you, but think about removing stores which don't have enough good data to make the valueable for predicting other store's missing data. Also remove months which don't seem to be complete, very often bordering the missing months of data.
    a. Remove entire outlier stores
    b. Remove outlier store months
5. Impute (complete) the missing months. All months from Jan 2017 to Dec 2018 should be now available for analysis. No stores should have missing months of data.
    a. The imputation model is up to you, but many off-the-shelf models are available in python.
6. Revisit step 2 for aggregation and now add ISIN to the aggregation (basically the original file).
7. Try to repeat steps 3-5 above, however this time predicting sub_dollar on an ourmonth, store and ISIN level. If you are not able to successfully do this, please write a few sentences on what went wrong and how you think such a task could be ultimately solved.