Report for Social Crop Challenge

Social Crop challenge was about time series data, actually stock prices of agriculture commodities. Many crops and APMC (Agricultural produce market committee)/mandi were collected with information about monthly price fluctuation.

Project description

For solving this challenge the project was split into some folders for better organization,

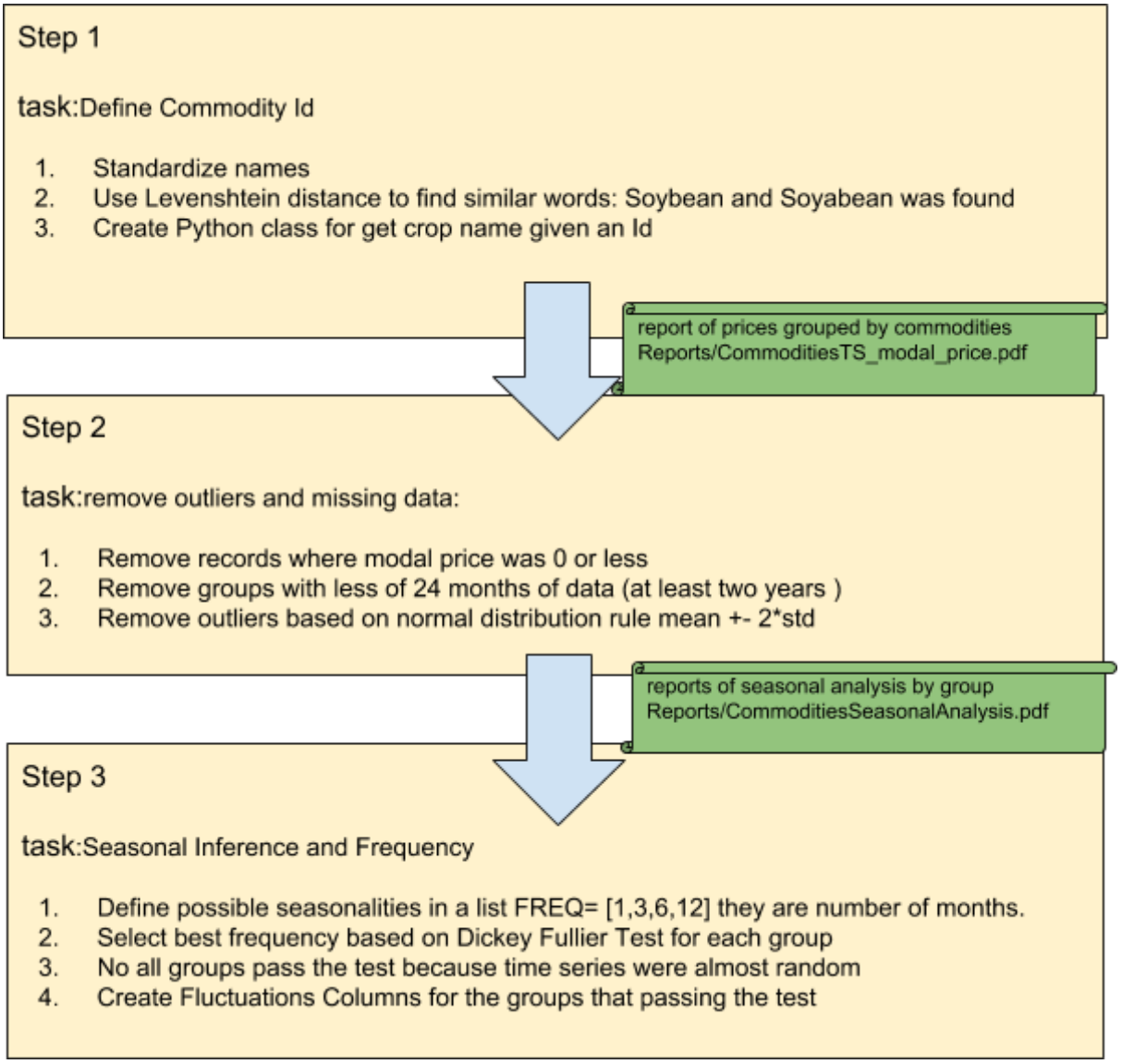
Factory	In this folder you will find the entity Commodity.py, I create this class for manage the Id of Commodities, so tables will not use differents names for same crops instead they should use the exactly Id, this is useful for Join operation and help to validate correctness of code.
PreprocessingExploration	Folder for preprocessing data before of analysis, outlier detection and duplicate data. CreateDBCommodity.py, this file creates database of commodities CreateMatcherCommodity.py this files is for setting commodity Id in monthly data CreateMatcherCommodityCMO.py this files is for setting commodity Id in CMO support price data OutliersDetection.py this file is for outlier detection, and remove missing data, further information will be show in the file PlotByCommodity.py this file is a visual report by commodities for visualice APMC price behavior
data	this is the data for scripts
Analysis	SeasonalityDetection.py this is a graphic report for understanding seasonality of commodities, and useful for select type additive or multiplicative for time series SeasonalDecomPriceSupport.py graphic report for visualize and compare seasonal data and deseasonalize data with the support price in this file we use commodity Id for Join operation.

	SeasonalInferenceFrequency.py this script perform the task of guess frequency , deseasonalize data, and create fluctuations columns
Notebooks	Small script step by step for explain all the process

**Further information can be find in each file

Workflow

The implementation of the solution could be easily explain with this flow, despite many details exists inside it, this could give you a good general understanding.



Main Files descriptions

PreprocessingExploration Folder

CreateDBCommodity.py
this scripts is for standardize the Commodity Code, We get the most of the names from the file of monthly data to finally create the database(file) DB_Commodity.csv with an Id for each Crop Commodity In this process we transform the names to uppercase and remove special characters, to word space we use "_" as word separator

CreateMatcherCommodity.py
this script if for add the column commodity Id to make easy to join this data with other data using the same entity but sometimes with different name and formats we take the raw data ./data/Monthly_data_cmo.csv and return a new file with the addition of the id

CreateMatcherCommodityCMO.py
this script if for add the column commodity Id to make easy to join this data with other data using the same entity but sometimes with different name and formats we take the raw data ./data/CMO_MSP_Mandi.csv and return a new file with the addition of the id, sometimes words are similar but write in a different way we use Levenshtein distance to find the crop commodity with similar name, like soybean and soyabean, and use the same Id, when we don't find any crop in the database similar we should consider to add this one as a new commodity

OutliersDetection.py

Outliers detection, missing data, and not enough records
This script is for clean the data, some steps perform here are:

* Remove records with price equal to zero
* Remove groups with time series with less than two years of data
* Remove outlier based on normal distribution mean +- 2*std
* Remove duplicates

You could un comment viewPlots function call to see histograms and graphs given an Dataframe
this was used for a small basic exploration

the final result was the file Monthly_data_cmo_step2.csv

PlotByCommodity.py

This a simple report script for visualize all the crops commodities join in one graphic all the institutions that sell this product, then two study how is the general behavior of each crop

the results were tree pdf reports CommoditiesTS_<var_name>.pdf
Crops are ordered by the number of records, descending so you will see the most populated in the first plot

Analysis Folder

PlotHighestFluctuation.py

This script is to show the groups with a high fluctuation
this step use Monthly_data_cmo_step3 where we create a column to save
the rate of fluctuation to monthly and frequency(inside the group) prices
finally this file return a new file flagging every group which show a fluctuated behavior
I use the quantile 90% to take by means of the by commodity set

PlotSeasonalDecomPriceSupport.py

This script is for showing the relationship between some commodities and the support price, and to visualize the deseasonalize data

PlotSeasonalityDetection.py
Multiplicative or Additive? This script is for exploring the data and decide the type of seasonality if it is multiplicative or additive, after to see all the plots we could decide that are additive models because changes in trend didn't imply changes in the seasonality

SeasonalInferenceFrequency.py
This Script depends of the data from step 2, This script is for analyze the monthly data for each group of commodity-APMC the first task is to type the date column for use date functions and create time series objects for each group the script will try to find the best seasonal behavior taking the Dickey Fullier Test and iterating over multiple frequencies [1,3,6,12] I defined the p value THRESHOLD_pvalue=0.3, to be acceptable over 30% because data is not so uniform after run every test the script select the frequency that is better for deseasonalize the time series Finally the script creates new columns for save the monthly and current frequency fluctuations Note: some groups didn't pass the test so they won't be taking into account