Every commodity goes through a long supply chain before reaching to consumers, involving different stakeholders at every stage of supply chain. Every stakeholder adds his share of profit to the price of original commodity. Some stakeholders are involved in illicit market manipulation techniques(like artificial supply deficit) with intentions of earning more profit.

These illicit market manipulation techniques can be observed by analyzing several characteristics of commodity like its production, wholesale price, retail price etc. As a use case we have taken onion.

For Onion, we have following data for every day since 2006:

1. Arrival in tonnes
2. Wholesale Price at Mandi level
3. Retail Price at Center level

Wholesale price is inversely proportional to arrival in the market and we assume it is the only factor on which wholesale price is dependent. We also noticed in our literature review that most of the farmers don’t hoard in India. So, from this relation, we have the following hypothesis:

***H1. If there is increase in arrival pattern, there should be decrease in the wholesale price and if there is decrease in arrival pattern, there should be increase in the wholesale price considering a lag factor of 15 days.***

Retailers purchase onions from the wholesale markets, mandis, etc. So, rate at retail level should be directly proportional to wholesale price in that region. We assume here that demand remains constant and there is no supply shock created because of excessive export of onion So from here we get the following hypothesis.

***H2. If there is increase in the wholesale price of onion, then there will be corresponding increase in the retail price and vice-versa assuming demand remains constant and there is no supply shock created because of excessive export of onion.***

Similarly, we can state that if there is no change in the arrival for some period, then wholesale may remain same and if wholesale price remains same for some period than retail price may remain same. So from that we get following two hypothesis:

***H3: The deviation between arrival - wholesale and wholesale- retail should not vary much compared to values for same time in past years.***

Also, one should make a note of that, even in H1 and H2, when wholesale price or retail is increasing, then it should be in considerable amount. It should not be like, there is marginal increase in wholesale price and retail price is boosting up or there is little decrement in arrival and wholesale price goes up by unacceptable level.

Also, we assume that mandis in the same region will behave in similar manner, because production and effect of other factors will be same in one region. So, based on that we have following hypothesis:

***H4: Mandis in the same region should follow the same relationship between arrival and wholesale price, as that of, taken whole region combinely.***

Our attempt would be to create a library which takes all these series as an input and try to detect possibilities of above stated anomalies or highlight all the important unusual behavior of parameters seen in the input series.

Test Criterion for H1:

* Starting with granularity of year-wise, find out cross-correlation, considering various lag factor up to 15, between arrival and wholesale and check that overall it is positive or negative.
* Go on decrease the granularity. Next will be season-wise. For a particular season, how arrival and wholesale price are behaving. Point out where, this correlation becomes positive.
* Similar thing can be done for month-wise as well as fortnight data.
* Note that, correlation mentioned here is just one of the method. There can be various other methods to define the behaviour of 2 time-serieses.
* Above method breaks time. It can also be done considering space as well. It can be mandi-wise, region-wise, state-wise.

Test Criterion for H2:

* Same techniques as described in H1 can also be applied here.
* Since both the parameters (i.e Wholesale price and retail price) have same unit of measurement. We can take the relative difference of the retail price and wholesale and plot another series which will indicate Retail Price- Wholesale Price.
* Positive spikes in this series indicates large increase in retail price compared to wholesale price.

Test Criterion for H3:

* One can use prediction based model like ARIMA [2] to test this hypothesis.
* If method mentioned in [2] is used then there is no need to align time-series into phase, as method mentioned in it take care of it.
* One can train the model considering some period for 8 years and can be tested on remaining 2 years. Difference in actual and predicted value is seen and one above some threshold value is reported.
* Also, while generating model, different window size can be considered.
* If some other technique is used, then one may need to align time series into common phase. That can be done using cross-correlation method, considering lag factor of around 15 days. Lag with the highest correlation value is considered.
* One can also apply graph based anomaly detection technique as mentioned in [3] to find out malicious behaviour.

Test Criterion for H4:

* Here, we need to combine data of multiple places into one, to find out the combined behaviour and then compare it with each of the mandi.
* One of the technique, to combine multiple series is,
  + For Arrivals: take sum from all Mandis
  + For Wholesale-price: Take average
  + Apart from center, if one wants to combine statewise, then for retail price also, one can take average for retail price
* Other technique, to combine multiple time series is, subspace based transformation, as mentioned in [1].
* Then, to compare behavior of each particular mandi time-series with the aggregated one, techniques mentioned to test hypothesis 1 or 2 can be used.

**Note:**

Results returned by H1 and H2, consists of all the time periods where arrival-wholesale price and wholesale price-retail price pairs go out of line. There may be case where, for each time of that year it may be going out of line and might not be anomaly. So, to remove such false positives, we can take intersection of results of H1 and results of H3. Similar thing can be done with H2 also, by taking intersection of results of H2 and results of H3.

System Design:

Assumptions:

1. Given time series should be of the same time period.
2. Each time series should be present in different file.

Flow:

* Take input, How many time series are there? (Let’s say user gave n time series)
* Give option to select each time series from a file, along with that also give an option whether user wants to smooth time series or not?
* Generate n\*n editable matrix with some default value, where user can specify the relation between any two time series as follows:
  + 1: Positive Correlation
  + -1: negative Correlation
  + 0: Random Relation
  + Default: User is not aware of any relation and also any doesn’t exist according to user
  + 2: User is not aware but is interested in finding relation, if exist any
* Where correlation exists, take lag factor to consider as input:
  + Number of Units for lag
  + Consider only positive, only negative or both type of lags
* Run the functions
* Display the output to the user, what tests have been conducted and their results. Detailed values will be provided if asked specifically. Different Graphs/Charts possible with the output can be generated on user demand.
* Provide an option to user whether he would like to see the results with some different threshold value other than the taken by library.
* Provide all the functionalities present in the system and ask user, if he wants to run some specific tests on the input time series.

Processing Modules:

* General Modules
  + Exponential Smoothing
  + Average Smoothing
* For H1
  + Correlation in Sliding Window
    - Window Size: Default 15 days (on trial basis, we are yet to decide final value, which we will finalising after some runs)
  + Slope Based Detection
    - Calculate the slope of both time-series daywise and compare

Assumption: Data is smoothed. Since day-wise is taken, if data is not smoothed then it may generate many “spikes”

* + - . So if data is not smoothed, then before applying this technique, either apply smoothing or take weekly average. If granularity of data is daywise, then smoothing can help. If data is reported once-twice weekly, then taking average weekly and then calculating slope day-wise may help.
* For H2
  + 2 methods mentioned above for H1

Spike Detection Method, if both the time-series have the same units.

References:

1. Cheboli, D., 2010. Anomaly detection of time series (Doctoral dissertation, University of Minnesota).
2. Hillmer, S.C., Larcker, D.F. and Schroeder, D.A., 1983. Forecasting accounting data: A multiple time‐series analysis. Journal of Forecasting,2(4), pp.389-404.
3. Cheng, H., Tan, P.N., Potter, C. and Klooster, S.A., 2009, January. Detection and Characterization of Anomalies in Multivariate Time Series. InSDM (pp. 413-424).