SocialCops Submission

Notebook:	Internship Resume and Data		
Created:	04-11-2018 05:27 PM	Updated:	05-11-2018 04:01 PM
URL:	http://localhost:8888/notebooks/Data%20Science%20Projects/SocialCops/Mandi%20…		

main.ipynb:

This Jupyter Notebook contains Data Manipulation and Comparison of Minimum, Maximum and Modal price with MS price for all the commodities for which MS price was known.

Data Manipulation:
There were many records in which a single commodity was registered under two different names pertaining to inconsistent use of capitals. So, I merged such records into one and removed extra records.

Comparison with MS Price:
For comparison of prices, I did an Inner Join on ["APMC"and "Commodity"] for both the data sets and then for each cluster, plotted all the four prices.

Detecting Outliers:
For outliers detection, I used two methods:

1. IQR (points 1.5*IQR above and below Q3 and Q1 respectively)
2. Z Scores (points more than Z Score of 3.5)

However, for both the methods, no outliers were detected for any of the three listed prices.

Flagging highest fluctuations:
For flagging highest fluctuations, I at first group the data for every month for all the three years. Then, for a particular month and year, I calculate fluctuations for all the commodities(maximum value for that month and year - minimum value for that month and year) and then select the commodity with maximum fluctuation for that particular month and year.

I have saved all the manipulated data files with same name in excel.

temp.ipynb:

I have used mandi02.ipynb as a temporary notebook where I did most of the calculations and operations on data.

Detecting Seasonality and Trends:
For detecting seasonality and trends in prices of the commodities for each cluster, at first, I employed seasonal_decompose() function available in statsmodels library.
As data provided to me was recording on monthly basis, I set the frequency of calculation to be 12.
Upon observation of plots for each cluster, I found that the this method was not able to extract trends and seasonality from the data.
I think this is due to a very few number of observations available for each cluster (a maximum value of 26 for Pigeon Pea (Tur) at Jalgaon Jamod-Aasalgaon).

Also, Rolling Mean or Rolling standard deviation failed for window size of 12(monthly data, therefore 12) as in most of the cases, there were not as many as 12 observations for each cluster. I think, similar results can be observed by doing Differencing of time series data.

Then, I tried using autocorrelation function on the time series but in most of the cases, I got negative autocorrelation. In some which I did get positive AC were at lags of from 3 to 5.