Agra District Price Report

b) I plotted the scatter plots for observing the pattern of prices for each market in the Agra District. From the plots, we can find out two things:

i) For Achnera, Agra, Fatehabad, Fatehpur Sikri markets, there is a decreasing trend in prices from approx 1400 to 700, from starting of January to end of February. After that there is an increasing trend from approx 700 to 3000 from starting of March till end of November. There is approximately linear trend. The prices are highest in November. And at last the prices drop down suddenly in December.

ii) For Jagnair, Jarar, Khairagarh, Samsabad markets, the trends are different for different markets. For Jagnair, Jarar and Khairagarh, prices decreases from approx 1250 to 600 from starting of Jan to end of February. For Jagnair, prices increases in March to approx 1400 and then remains constant in April to June. From July to November, market is almost closed and opened in December at 3000 and drops down to approx 600. For Jarar, there is an increasing and decreasing trend till June and the market is closed after that. For Khairagarh there is an increasing trend from March to April and then decreasing from May to June. The market is closed from July to November and then opened in December with a very high price and drops down suddenly in the same month. For Samsabad the prices decreased, increased and then decreased with constant price for each month. This is from January to June and the market was almost closed till December.

c) i) I would check the data types for each column. If they are not right, then I have to make the right data type.

I would check if there is any missing value in the data set. If yes, then I have to fill them. There are different criteria for that like if the values are numerical then fill it with mean or median, if the values are categorical then fill it with mode, if there are few missing values then we can drop those rows, we can fill the values with 0 also or if there are many missing values then we can fill them with random values.

I would check if there is any column that is not necessary for the model or it has only one value for all the rows. If yes, then drop that column. Here Grade column has only one value.

I would check if there are outliers in the dataset. If it is there then take log of that column or percentile value to remove the outliers.

c)ii) I can use Variety column as a feature. Since it is in string data type and machine learning doesn’t work with strings. So I will convert it into numbers. It is Nominal data and so I will use one hot encoding to do that.

c)iii) It can be framed as a machine learning problem by predicting the prices of Potato in year 2021 by using the data of year 2020. Here the target variable will be the price. It can be Max, Min or Modal price.

c)iv) Since the prices are numerical data so I can use regression model. It can be linear regression, multiple regression, Decision tree Regressor, Random Forest Regressor etc.

c)v) The loss function I would use will be Mean Squared Error, Mean Absolute Error, Root Mean Square Error, R2 Score.

c)vi) Before running the model I can split the dataset into training and test data set. I will run the model on train data set and can check the accuracy of model in test data set.