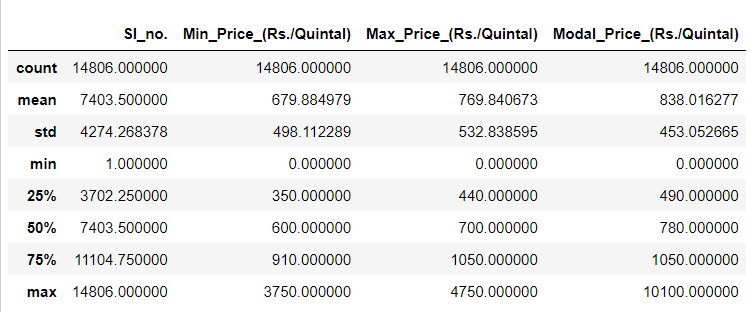
1. **What are the data pre-processing / cleaning techniques you would apply?**

****

For data cleaning, in the min\_price, max\_price and modal\_price columns, by looking at the max row, there are clearly outliers present.

So we can remove outliers by filtering the data frame on the conditions that the value of a variable should lie in the range (25% - 1.5\*IQR , 75% + 1.5\*IQR) for that particular variable.

Here IQR is the interquartile range which is given by 75% - 25% or Q3 - Q1.

After performing the above outlier detection, it was observed that the outliers were present only on the higher side.

There were 595 rows having at least one variable of the three as outlier.

Since it is a very small number as compared to the entire dataset, we can simply drop these rows.

On the lower side however there are 2048 rows having at least one of these three variables as zero and also it was observed that the variables are either zero or have an acceptable realistic value. So it is clear that the value is zero because the data was unavailable during collection.

Since 2048 is a large number, we cannot drop all these rows.

To treat this problem, we will first remove the outliers on the higher side and then for each outlier on the lower side, we will take the mean of the entire column without any outliers and fill all zeros with that mean value.

For the Market\_Name and Variety features I will use OneHotEncoding as they are nominal type of categorical variables.

1. **What are the features you would use to create the model?**

Since the features like District\_Name, Commodity and Grade have only 1 unique value, they cannot provide any useful information to the model.

Hence, I would use the remaining features which are as follows:

Market\_Name

Variety

Min\_Price

Max\_Price

Modal Price

Price\_date

1. **How would you frame this problem as a machine learning problem? What are the possible options ? Mention why you choose to go with a particular option.**

We can keep the target variable as Modal\_Price\_(Rs./Quintal) and using ML predict it for a particular day, at a particular market for a particular variety using the trends observed by the ML model in the previous years.

This can be framed as a multivariate time series forecasting problem where we are forecasting the modal\_price of the commodity.

Since the problem revolves around the fact that prediction of the crop price is important for efficient crop planning, we need to forecast the prices of the crop based on past observations.

The other option is that we extract the month and year from the price\_date column and treat month as an ordinal variable and the year as a continuous variable.

Using this approach we can use regression techniques to predict the price of the crop for a particular month and year.

However I would go with the former approach as it can provide more granularity as getting predictions on a scale of days is not feasible with the latter approach.

1. **Which algorithm would you use for price prediction?**

I would use the SARIMA(Seasonal Autoregressive Integrated Moving Average) algorithm which to predict the price because SARIMA allows for the incorporation of seasonal patterns which are important to predict prices of crops as they are highly dependent on the climate which in turn have strong seasonality at least in India.

1. **What would be the loss function you would use?**

I would use the MSE loss function as it is a regression problem and MSE loss is computationally cheap and highly intuitive making it easy to measure the model’s performance in case of time series forecasting.

1. **Any other comments you want to add?**

One more approach that we can try is to group the data by the market\_place variable and then for each market place again group it by the variety variable and then perform time series forecasting on each of the 48 groups using only the continuous variables. It will reduce the number of input features for the forecasting algorithm and the influence of one class on another class of the same variable during forecasting can be reduced. However it comes with a drawback that there will be 48 different models to train. This approach can be used for suc lower numbers but for situations having variables with many classes, the above forecasting method can be preferred.