**Answers to question c**

c)

* First the data that I have extracted from the site has all the numerical data points as floating point numbers . So I will convert **S\_Number , Min Price , Max Price** to integer form as present in the site.
* There is a column named **Price Date which has dates in string format** . I will change that to **datetime format** so that I can extract weekdays and months from that .
* I will check for missing values in each column . I will impute that depending on the column which has missing values.
* Encode the categorical values in the **columns District Name , Market name , Commodity , Variety and Grade** into numerical values using Label Encoder.

ii.

* From the data it is understood that , **S\_Number , District Name and Grade** would'nt play any role for predicting prices.S\_Number is an independent variable that increases throughout the data. Also District Name and Grade are constant thoughtout the data.**So I will exclude these features.**
* From the plots of price pattern for major market places **,** I realised that market nameis important for predicting prices However Variety doesn't play a role in predicting prices .
* From the plots , I also realised that months play an important role in predicting the prices and week of the months also affect prices.

So overall features that I will be using are **Market Name , month and week\_no and Modal Price .**

iii. **Formulation of ML Problem**

This is a **regression type** problem where I have to predict the prices of potato for a particular market over some time. As I will be predicting real values using some features so I am treating it as a Regression Problem.

The **target variable** will be **Modal Price (Rs./Quintal)** . Also I will find the relation between Modal Price (Rs./Quintal) ,Min Price (Rs./Quintal) and Max Price (Rs./Quintal) using some regression technique and derive those from the predicted Modal Price.

iv. I will be using **XGBoost algorithm** for predicting prices .Although there are many algorithms for regression problems ,I prefer using XGBoost because of it performance.

After performing some Hyperparameter Tuning on parameters like **n\_estimators , max\_depth , subsample** and other parameters , it can yield really good results.

For deriving Max and Min Price , I will find the slope and intercept using Linear Regression on the original data and then use these coefficients to calculate the max and min prices

from Modal Price .

v. The loss function that I want to use is Root Mean Squared Error (RMSE).The formula for RMSE is . where yi is actual value and y i^ is predicted value.

A perfect RMSE value is 0 .