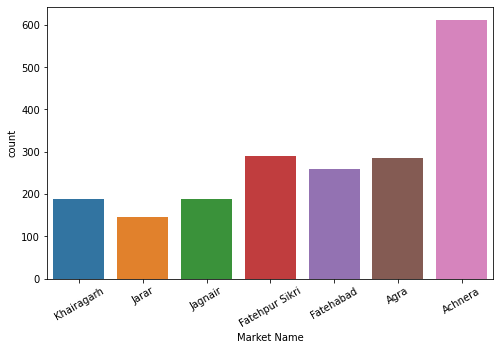
**NAME: Harshita Gupta**

**This doc deals with the answers to the questions given in the problem statement.**

**Using the data from 01-01-2020 to 01-12-2020**

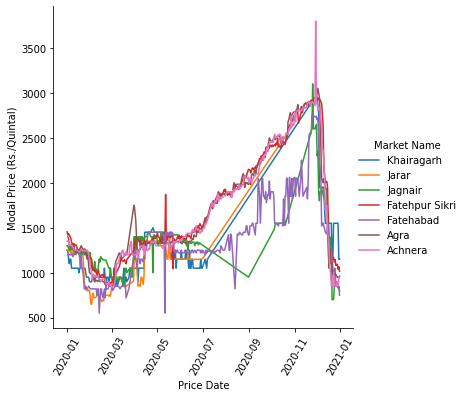
**PART B**

**Major markets for the district “Agra”**



Since Achnera has the maximum entries (count) thus it is one of the major market of Agra district.

**Price patterns for each of the markets.**



From the above graph it can be seen that the price of the potato per quintal increased at the end of the month. Since it was the time of pandemic so there was a supply shortage, rise in demand pushed potato prices up by 40% in major cities.

**PART C**

1. Data Preprocessing techniques-
2. The modal/min/max prices in the data are left skewed, so they should be transformed to normal distribution.
3. **Market names** and **Variety** of the potatoes can be one hot encoded.
4. Various columns such as Sl no. , District Name, Grade, Commodity can be removed as they do not provide any valuable information.
5. Price Date column can also be removed if the data is not treated as time series data.
6. Feature scaling of modal/max/min price can be done to bring them in the same range.

**2.** Features use to create the model-

1. As stated in 1(c) these features are either unique or same for all, so they don't provide any valuable information and thus are discarded.
2. And since the data is only for a year, we can transform the data as a time series since we can identify any pattern over the year. So Price Date column is also discarded.
3. Thus can use the model/max and min price of the potato, market and variety of the potato for modeling.

**3.** This problem can be framed as a machine learning problem to predict the price of the potato per quintal (dependent variable) from the other features provided (dependent variable.)

Modal Price would be the target variable.

**4.** Since from the graph of modal price, we can see an increasing trend in the prices for the year 2020, we can apply linear regression for price prediction.

(For LR models we don't need to transform the data to normal distribution.)

**5.** Loss function could be RMSE/MSE loss as we have to minimise the difference in the actual price and the predicted price.

**6.** Additional comments:

1. Some feature engineering can be done to improve the model performance.
2. Previous year data can be taken to convert the problem into a time series problem.
3. The data is imbalanced as major entries are from Anchera market.
4. On analysing the data based on variety, it shows a similar trend as the modal price (increasing at the later months of the year.)

