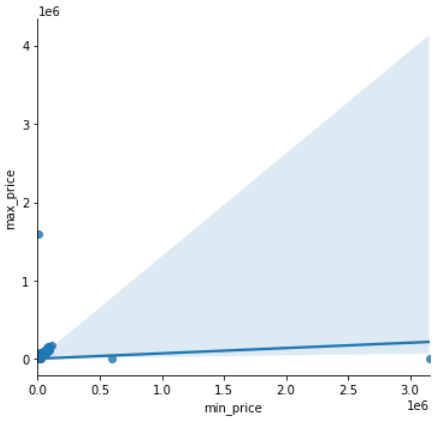
Dataset Chosen

Using agricultural-commodity-dataset by Sam Paul on Kaggle, and adding some nan values in it data_cmo_noisy.csv

Anomalies Found



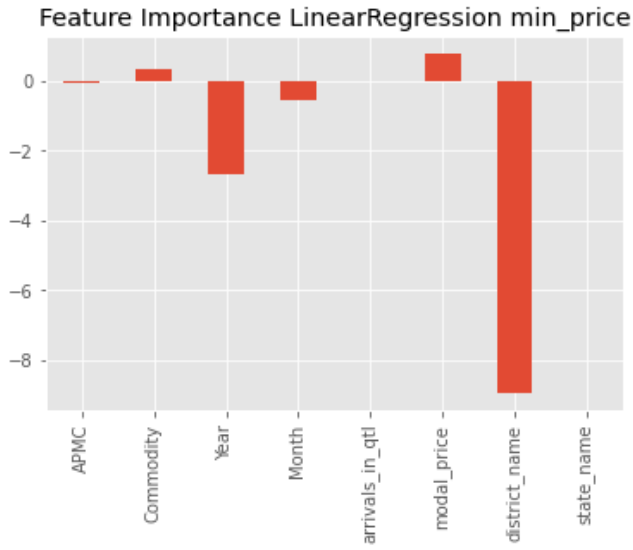
This graph displays the anomalies, those values are outliers which make the estimator take incorrect decisions, as we can see the regression line should have been different than it is.

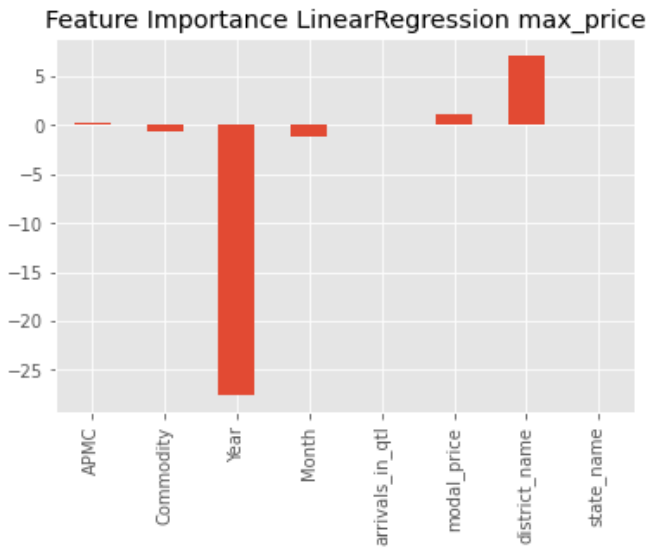
To clean it, we directly try to remove the outliers from the data set and we get a clean dataset.

We also have few nan values in the dataset, we fill some of them and drop the others which are hard to replace/fill.

After cleaning we save the clean data as data_cmo_clean.csv

Results of Code





This is the result of the code after feature importance, we see that the minimum price is most dependent on the district, and second-most on Year, for the maximum price it s a bit different, it is affected most by year, and second-most by the district,

the second thing we observe that arrival QTL is not much important as per the model (as we only have one state data, the state_name having no importance is expected.

Other graphs:

- Number of Arrivals by date
- Arrival QTL by date
- Total arrivals by month
- Arrivals in the district, APMC

Data interpretation

feature	defination
APMC	Agricultural produce market committee of the area
Commodity	The vegitable or crop
Year	Year in which the data was recorded
Month	The month in which the data was recorded
arrivals_in_qtl	The quantity of arrivals (units unknown)
modal_price	modal price of the commodity
district_name	The name of the district in which the data was recorded
state_name	The nae of the state in which the data was recorded, the data contains the data of Maharashtra, India only
date	the data on which the data was recorded
min_price	the minimum price of the commodity
max_price	the maximum price of the commodity

We see that our dataset has the data of commodity, the features are listed in the above table/image. From the analysis, we come to know that the prices are increasing year by year. We have the most data about Pune City in Maharashtra, Most arrivals are in the month of December and the least is around July-October, The arrival QTL increases as time goes. The commodity which has the most arrivals in QTL is onion, followed by potato and tomato.

Conclusion

Using this analysis we can train a Multiple-Regression model which will predict the minimum and the maximum price of a commodity, or it can also be used to collect the important information which contributes the most in predicting the price.

Solution by - Yash Pravin Pawar