Shape of the df (2733, 3)

Data imputation

There are quite many missing values. If the value are missing in the beginning or at the end of the observation period it may be better to drop it. But if the y are lost in the middle of the observation term it might be useful to replace them according to the tendency

Изображение

Number of the countries submitting the data is quite constant throughout the years:

But not all of the counties submit every parameter. More data is available about harvested area and yield then about a price.

We could see that the Yi

Изображение

Further data visualisation reveald that with years total world yield were increasing whereas total world area under the crop was increasing.

Question: Why is it so? Is there correlation.

First of all, this could be due availability of the data or due to technological development and improvement of agricultural practices. Yield is calculated from Area under the crop and Total harvested area, both parameters are equal present in the dataset. Thus, we could pass this data for further regression analysis. There was quite large difference between submitted data in 1991 and 1992, it might have sense to drop 1991 even though there may be historical reasons for that as well.Изображение

For the first glance with the years producer price was growing together with the yield over the years. But there are many missing values what will happen if interpolate the values?

Изображение

This is just beautiful

With interpolation (very primitive). So let’s have a look if I can run regression with thatИзображение