The report of the Final Project

Made by: Abish Magzhan (Monday 2PM CSS 328-03P)
Kozhan Nurassyl (Monday 2PM CSS 328-03P)
Assentay Makhmud(Monday 2PM CSS 328-03P)

This project aims to give a brief overview of food supply trends in four cities in Kazakhstan and to study the price change of essential commodities such as potatoes and milk between different regions of Kazakhstan.

The main part of this report are as follows:

1. Data overview
2. Linear Regression
3. KNN and SVM

The study will include observations for 15 years, from 2005 to 2020.

Data overview

Data obtained from https://data.world/datasets/kazakhstan. We are uploading CSV file, which contains food types and prices.

The first thing we need to do is make sure that there are no empty entries and that all values are measured in the same unit. We read the data and grouped it by different regions of Kazakhstan: Nur-Sultan (Astana), Almaty, Kostanay and Aktau. This will allow us to see if there are trends in food items from 2005 to 2020.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3365 entries, 1 to 3365
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   date        3365 non-null   object
 1   admin1      3365 non-null   object
 2   admin2      3365 non-null   object
 3   market      3365 non-null   object
 4   latitude    3365 non-null   object
 5   longitude   3365 non-null   object
 6   category    3365 non-null   object
 7   commodity   3365 non-null   object
 8   unit        3365 non-null   object
 9   priceflag   3365 non-null   object
 10  pricetype   3365 non-null   object
 11  currency    3365 non-null   object
 12  price       3365 non-null   object
 13  usdprice    3365 non-null   object
dtypes: object(14)
memory usage: 368.2+ KB
```

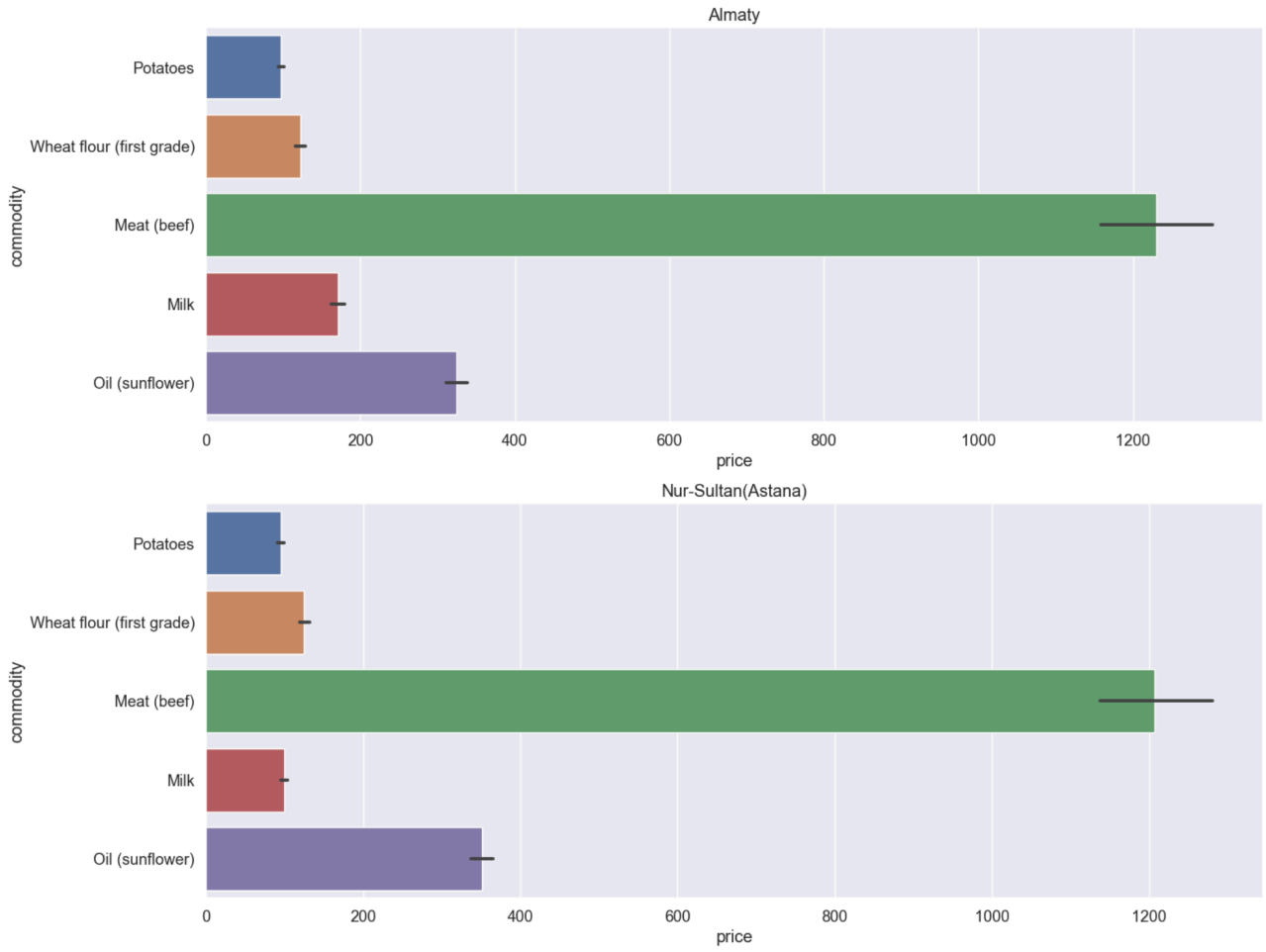
Pandas **dataframe.info()** function is used to get a concise summary of the dataframe. It comes really handy when doing exploratory analysis of the data. To get a quick overview of the dataset we use the **dataframe.info()** function: there are 3365 rows and 14 columns with no null values and each entry is in its correct type, in each column. The data is prepped for analysis.

After that, grouped the unique values from the 'market' column using **get_group()** method.

Let's visualize of our dataset using **seaborn.scatterplot()** function. We will only use the x (price), y (commodity) parameters of the function, while grouping data

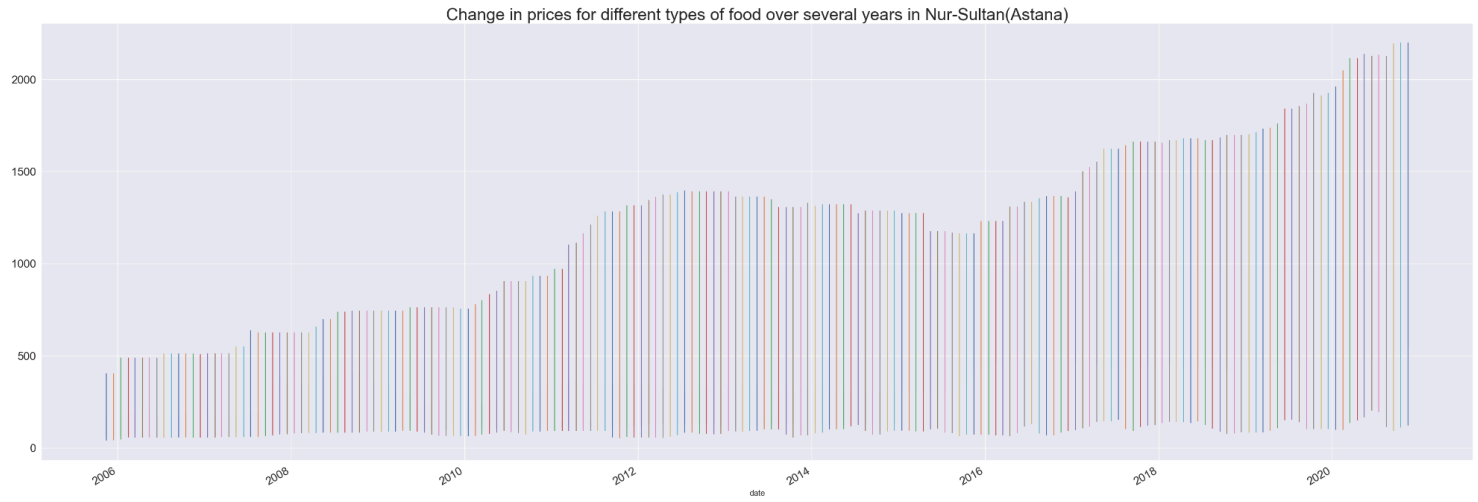
points on the basis of category, here as region. From the graph, the prices for the same food have been steadily rising in the given regions over the course of 15 years. The next step is to visualize each city in more detail.

Next, **seaborn.barplot()** method is used to draw a barplot. A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some information about the value of the prices of each given type of product.



As can be seen from the two bar plots, there is a price similarity between the two regions of Kazakhstan. Nur-Sultan(Astana) and Almaty have markedly higher meat prices, while Almaty has shown higher milk prices. This trend continues from 2005 to 2020. In general, the average indicator between these cities for the cost of meat(beef) is 1100 tenge.

Pandas **dataframe.resample()** function is used for time series data. A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Resampling generates a unique sampling distribution on the basis of the actual data. We can apply various frequency to resample our time series data. In this case we used the frequency - M(month end frequency)



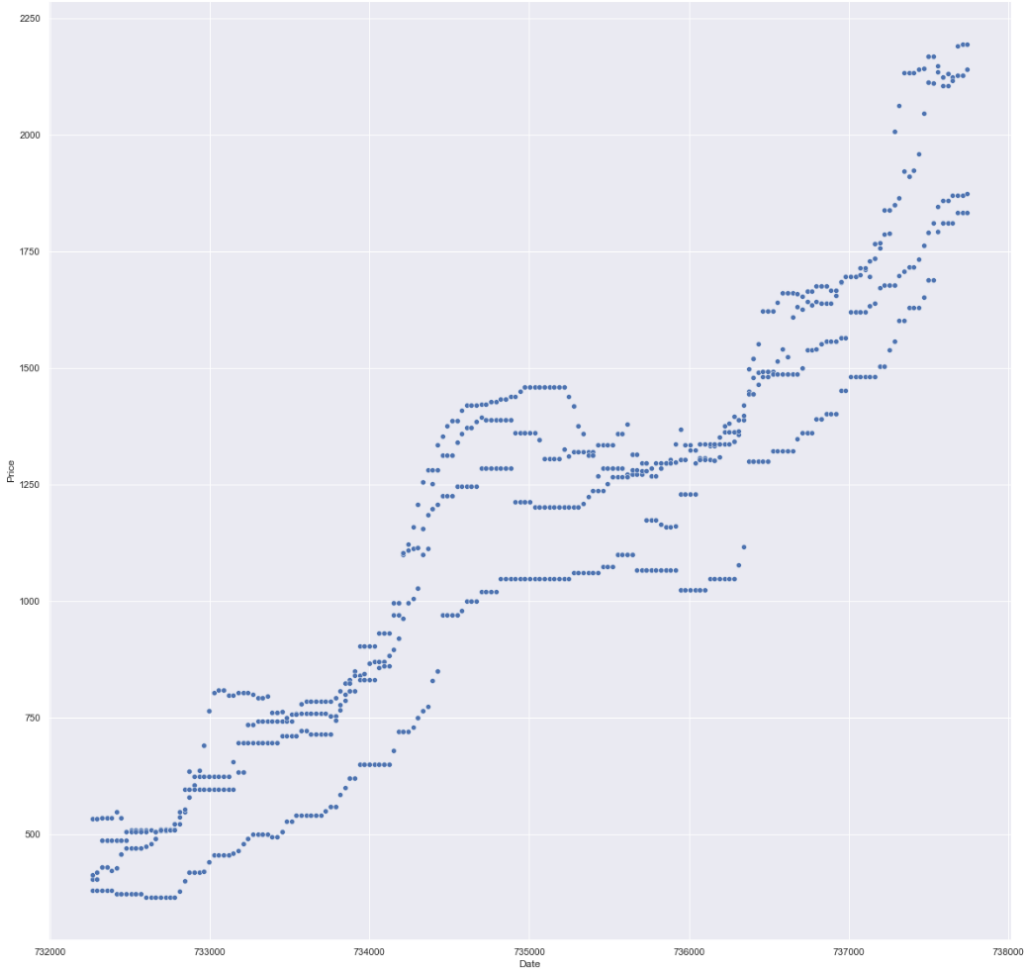
According to these, cities of Kazakhstan such as Nur-Sultan and Almaty have shown higher prices while other regions such as Kostanay and Aktau have also seen higher food prices. These 2 cities observed from the graphs of price changes can show how much it increases, which may indicate that regions in Kostanay and Aktau tend to have lower food prices.

Linear Regression

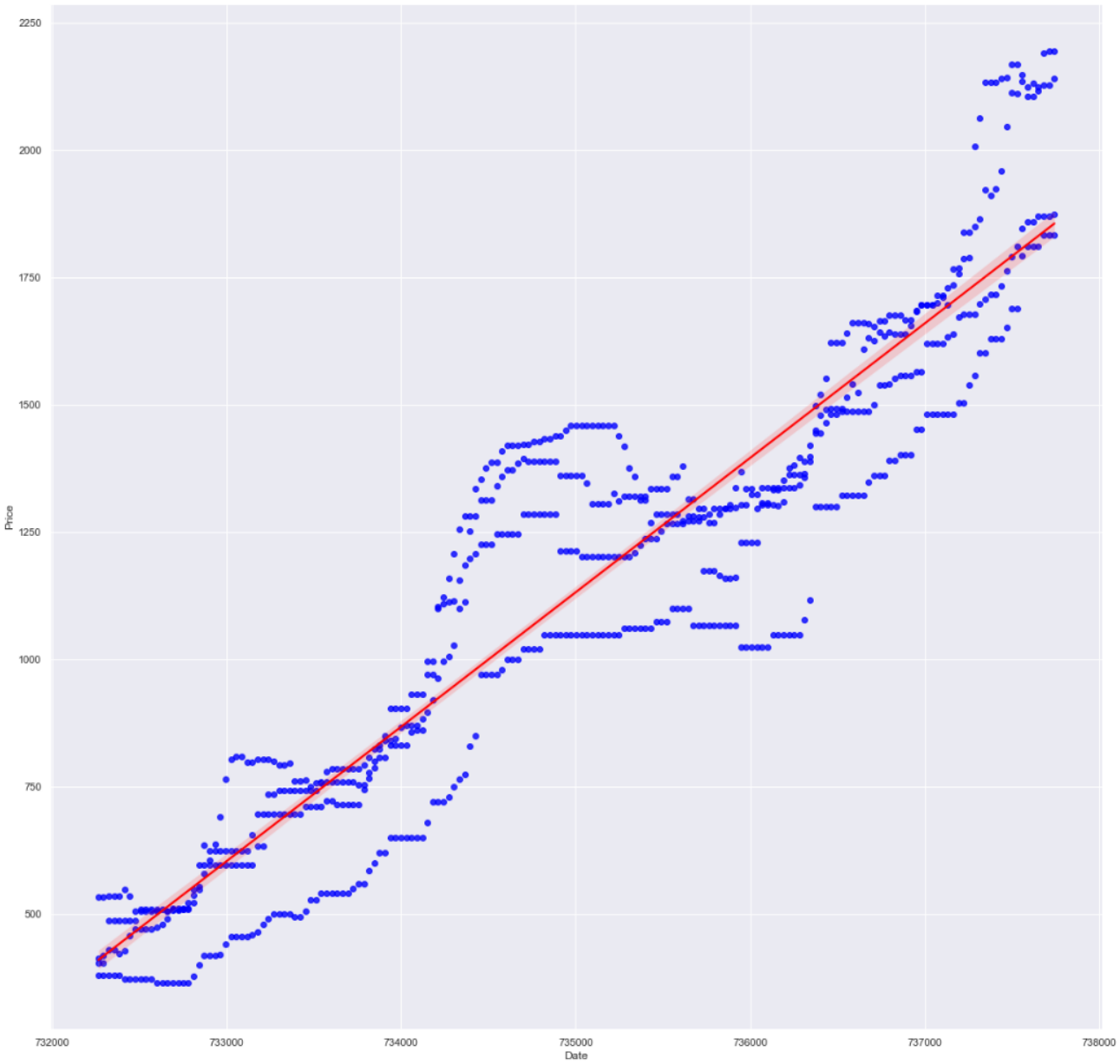
We have data on the prices of the products that are most needed for each family (potatoes, flour, meat beef, milk and sunflower oil) from 2005 to 2020 in the cities (Nur-Sultan (now Astana), Almaty, Kostanay, Aktau). Many people say that products have become very expensive, especially meat. To prove that let's consider all the regions that we have listed, for example, let's find how much meat has risen in price visually with a linear regression. And then we can make a prediction for future years of how the price will vary.

First, from this table we sort the commodity we want, from the columns of the commodity we derive only beef meat. Because we have a date type saved in the table as an object we first convert the date type to **datetime** using the **to_datetime** method. Next, to work with dates we can use the method **datetime.toordinal()** which returns

the proleptic ordinal number of the date according to the Gregorian calendar, where January 1 of year 1 has the ordinal number 1. This conversion will help us when we find a linear regression. We also need to convert the price to float since before all the data was stored as an object. Now that all the data is ready, we use the **seaborn** library to draw a scatter plot, where the x-axis is our date and the y-axis is the price of the meat. This is how we can see how the price has varied:



Next, using the ready-made **regplot** method, we can figure out the best option (desaturated red is all options that can claim as a regression line but only saturated is the best among all) for regression lines:



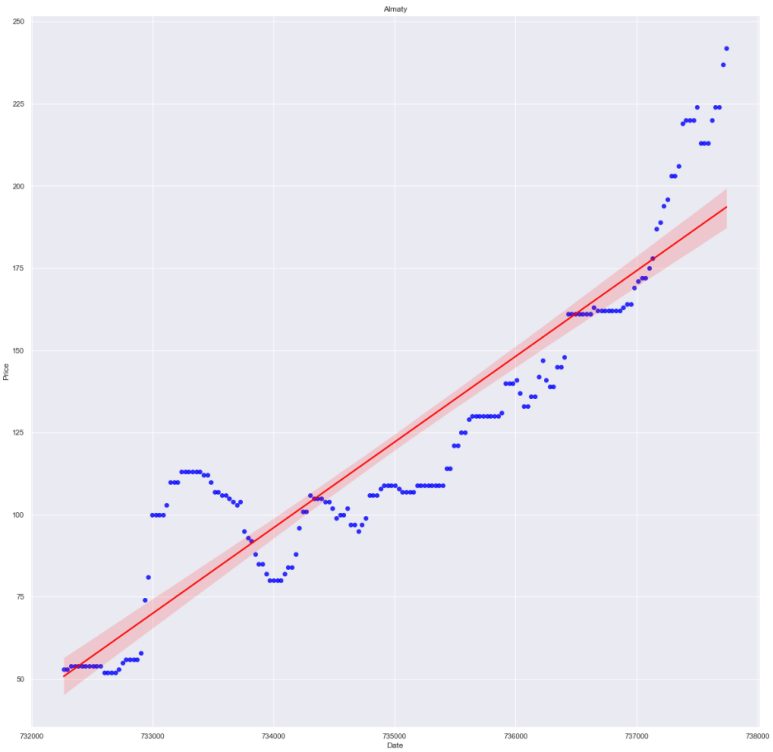
Here we can see visually how the price of meat rose linearly, starting in 2005 from about ~300 tenge to 2020 about ~1900 tenge, if we count the six times increased. The line also shows a clear increase in price.

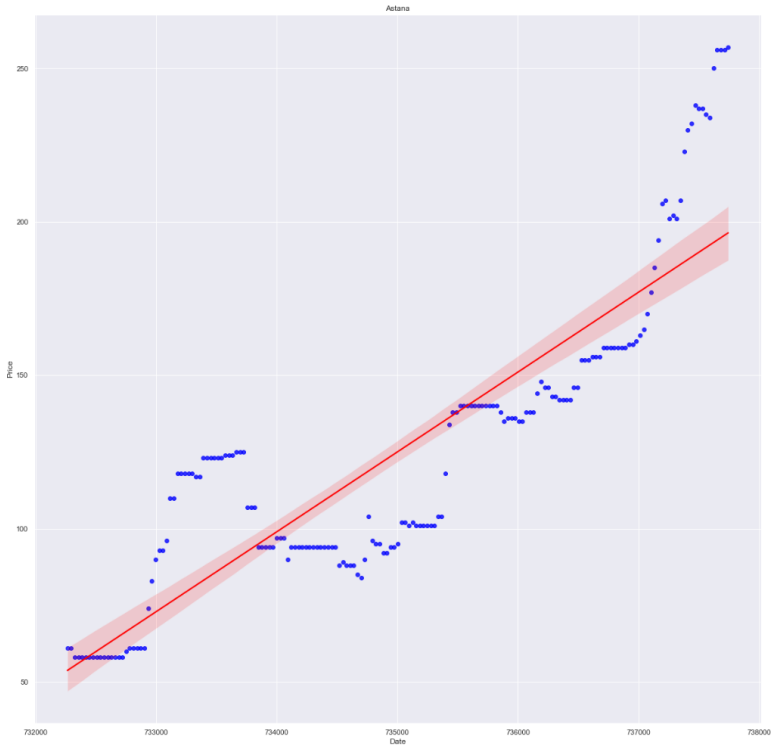
Now let's make some predictions. Since the data is relevant for 2020, let's say that 2022 is supposedly from the future for this table. With the help of the coefficient already found (model_LR.coef_= 0.26403365) and intercept_(model_LR.intercept_= -192933.51545354436) we can predict the price in two years (from 2020 to 2022), or rather, predict the supposedly today's price and compare it with the current average price for beef in Kazakhstan. Since we have in the table ordinal date number by Gregorian calendar, we take the last date is 737744 (this is the real date 2020-11-15) and add to it two years in days so 737744 + (365*2) = 738474, we use the formula y=kx + b. That is 738474 * 0.26403365 + (-192933.51545354436) = 2048.471 tenge. That is, according to the prediction with a linear regression at this time the price of meat should be 2048 tenge, but given mean_absolute_percentage_error (~11%) at this

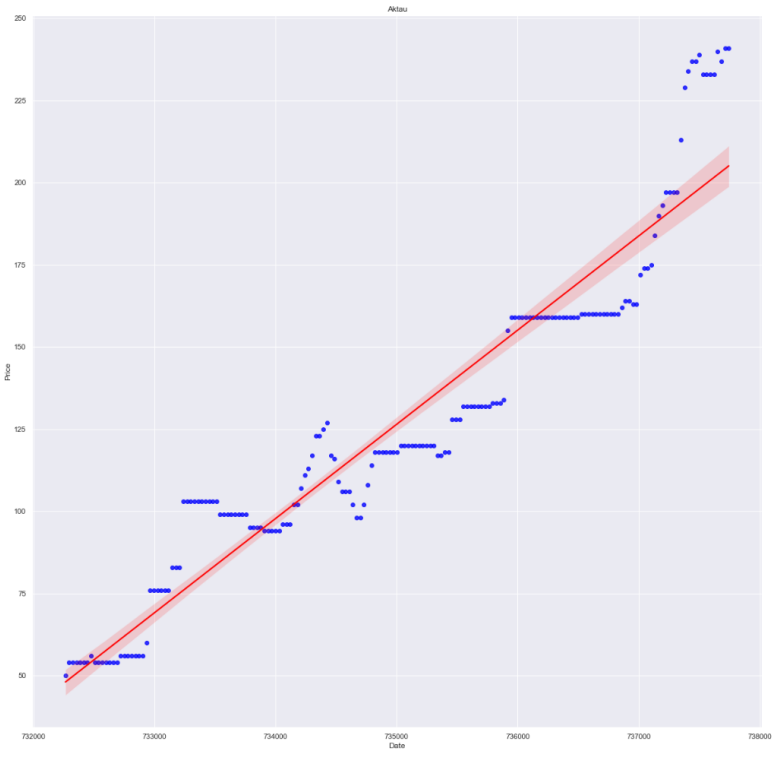
time the price of beef should be in the range [1823,2273], now at this site (https://vecher.kz/tsena-na-myaso-v-kazakhstane-prodolzhaet-rasti), the average retail price of Kazakhstani favorite beef in February 2022 was 2082 tenge per kilogram. 1823 < 2082 < 2273 -that is, our prediction coincides!!!

Now if we take flour, but for each region separately we find a linear regression, then for this product, it will look like this:

If we draw a linear regression for each, it will look like this:





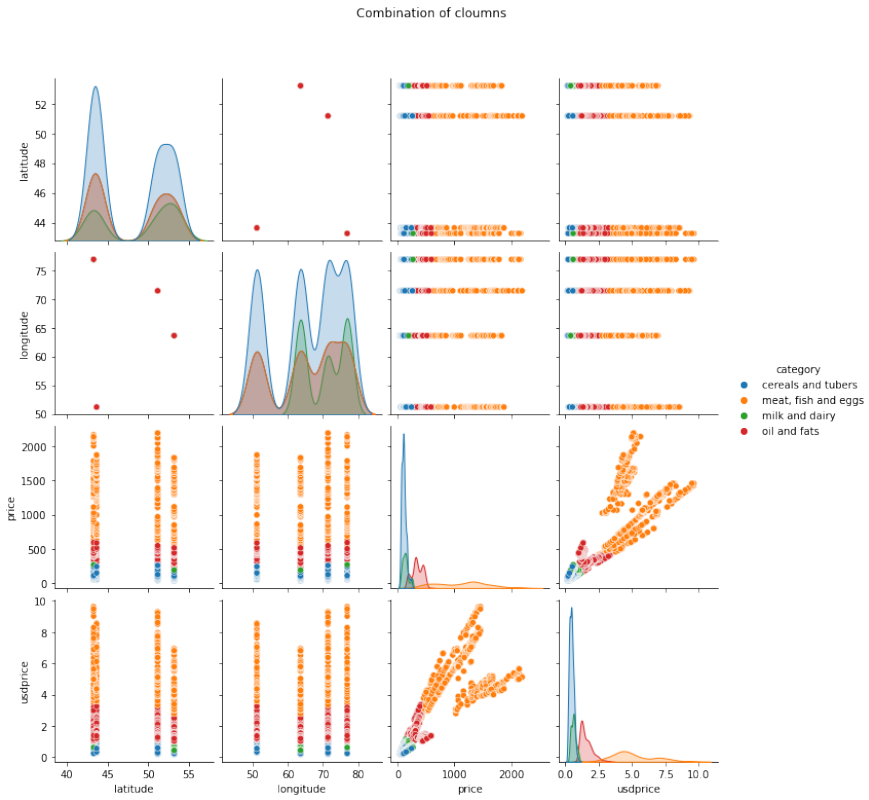


Just as we calculated above for meat, we can also make flour or for all the products that we have, taking into account all regions or a specific region.

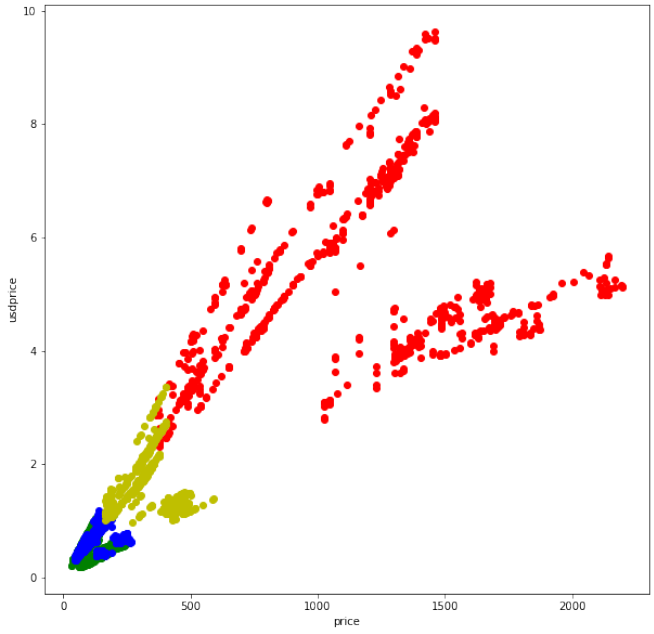
KNN and SVM

In this stage lets classify products. **.** For classification we need target column it would be “category”. However for solving this problem we need illustrate data. To

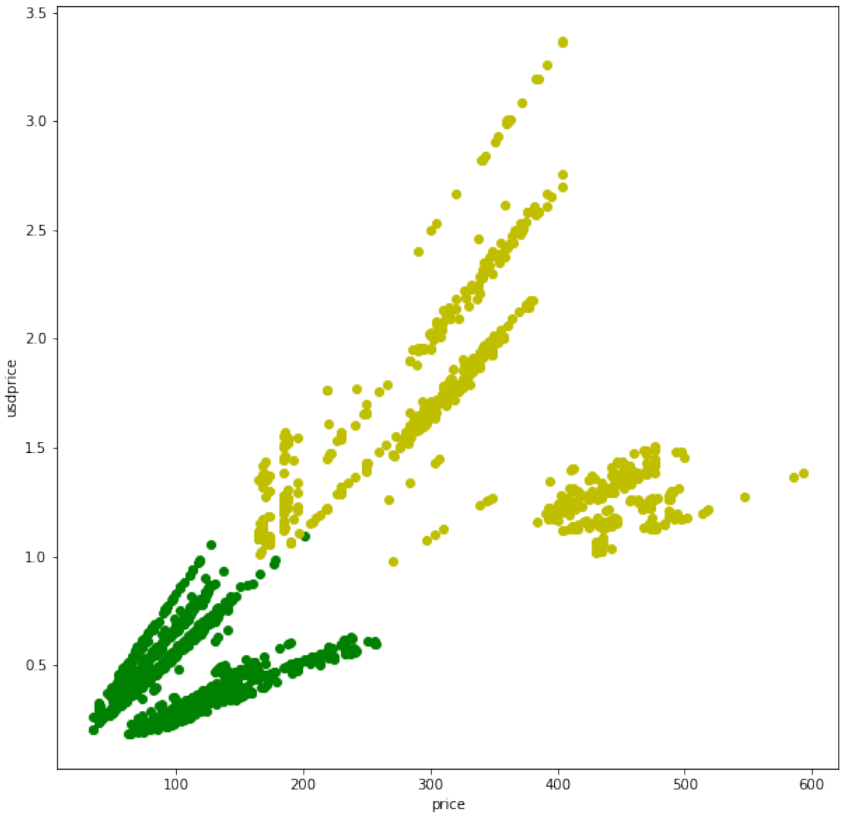
better show how data is distributed between each other in python, seaborn library has **pairplot.** It shown below.



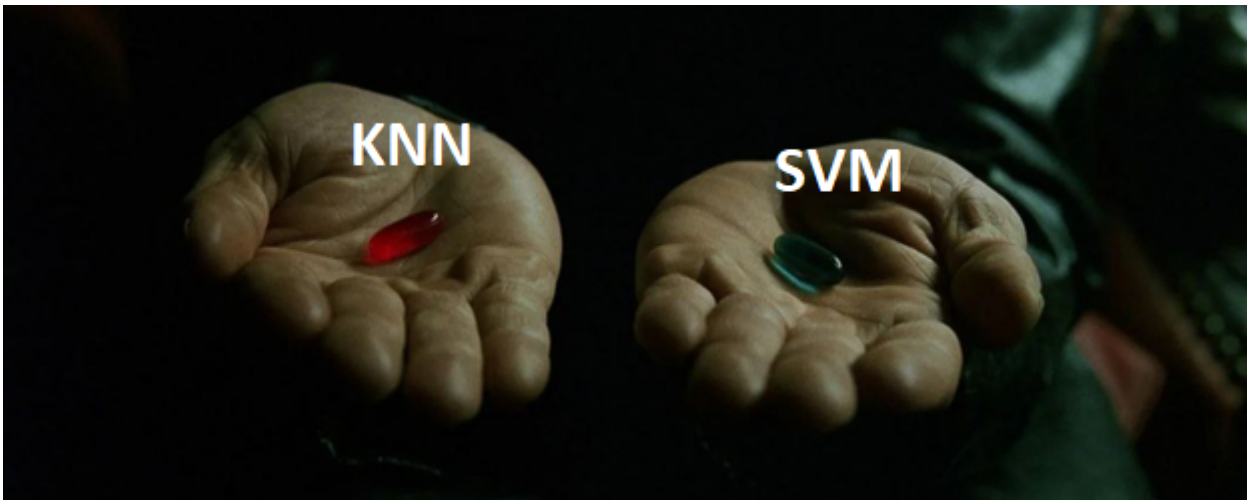
If we saw these graphs we can identify that columns “price” and “usdprice” are somehow distributed. In addition to that much closer for classification.



Meat and oil intersects on the most of data. When it comes milk and cereals it is impossible to take them at least in this situation. For classification, it would be better if we take cerals and oil.



So we prepare proper data. To classify them we can you two algorithms KNN and SVM. After that compare which of them fit better for this data.



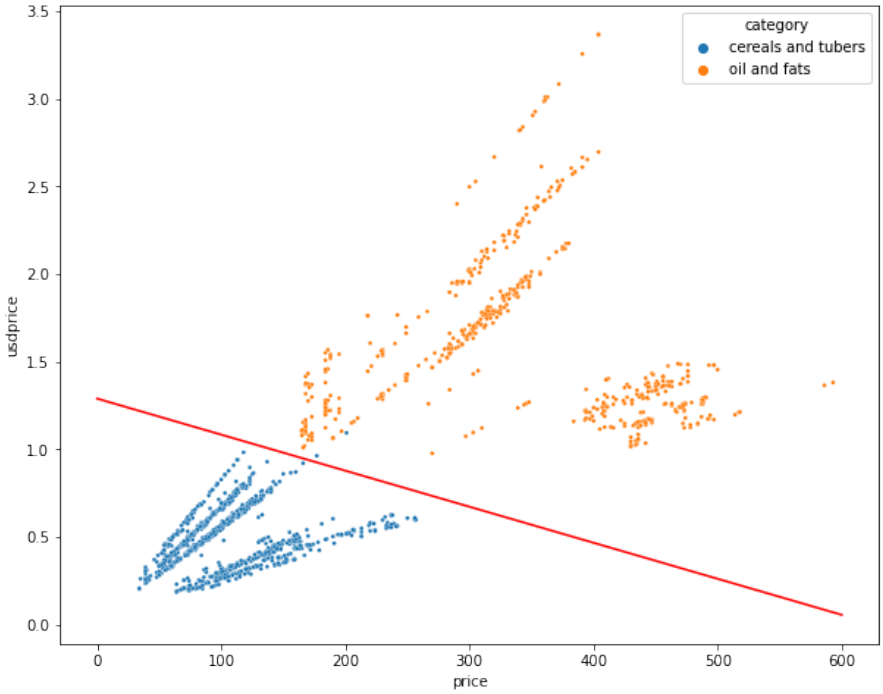
KNN - K-Nearest Neighbours

The first step starts with dividing data into train and test data with percentages of 80 and 20. Further, we need to normalize data to make the graph more suitable.

After implementing, all required functions we compare it with inbuilt functions. It shows 100 percent of similarity. In addition to that with an accuracy of 97 percent.

SVM - Support vector machine

For here we need to take already separated data in the KNN part. When we find accuracy inbuilt python libraries. It gave a 67 percent of accuracy.



For this sort of data we can conclude that KNN better suit.