METCS521 Final Project Submission

Predicting Inflation rate with linear Regression

Author: Jinxiu Cao, Zilong Li, Jingyao Lu

1. How do we analyzing the data so far? Any suggestion?

2. Could we wirte over 5 pages?

Introduction

The economic cycle occurs repeatedly in the operation of the economy and generally consists of four stages: recovery, prosperity, recession, and depression. Affected by this, the price of cereal will fluctuate accordingly.

In this project, based on a dataset illustrating cereal price changes within last 30 years which consists of 9 columns (year, month, wheat price, rice price, corn price, inflation rate, and the inflated price for wheat, rice, and corn) and 360 rows (https://www.kaggle.com/datasets/timmofeyy/-cerial-prices-changes-within-last-30-years?select=rice_wheat_corn_prices.csv),

we want to understand, interpret, and visualize cereal price changes and inflation rate changes within the last 30 years.

The dataset was preprocessed using **Numpy and Pandas**. In addition, we used **Seaborn and Matplotlib** to visualize the data. Apart from that, **sklearn.metrics and model_selection** were also implemented to do data prediction.

In addition, machine learning algorithms (linear regression model) will be implemented in the project to predict future inflation rate.

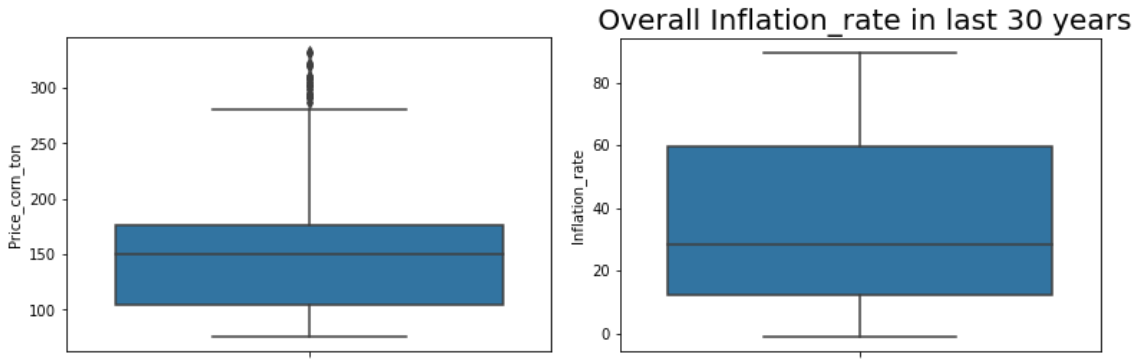
Two Jupyter notebook scripts (**script.ipynb** and **team project 521 (1).ipynb**) were used in the project to :

1. Download the dataset
2. Preliminary interpretation of the dataset
3. Exploration & Correlation
4. Linear regression of variables and prediction of test dataset
5. Clustering We used two machine learning methods in this project, supervised learning and unsupervised learning.In the future, we hope to use reinforcement learning to further study the relationship between grain prices and inflation rate .

Download the dataset

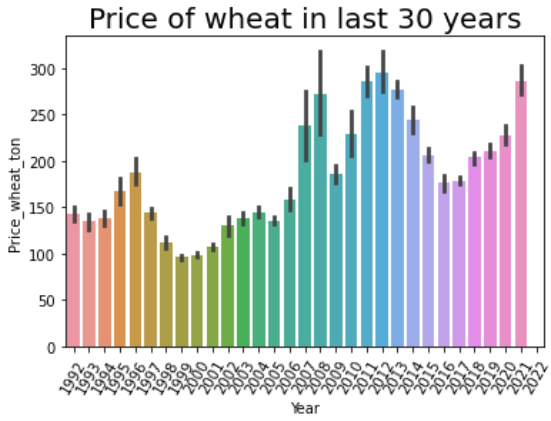
At first, we wanted to use **script.ipynb** to download the dataset programmatically. But one challenge lies in the raw dataset is not accessible because the dataset illustrating cereal price changes within the last 30 years was extracted from a website. Consequently, we downloaded the dataset csv file from the website and imported panda to read the csv file using pd.read_csv().

Preliminary interpretation of the dataset

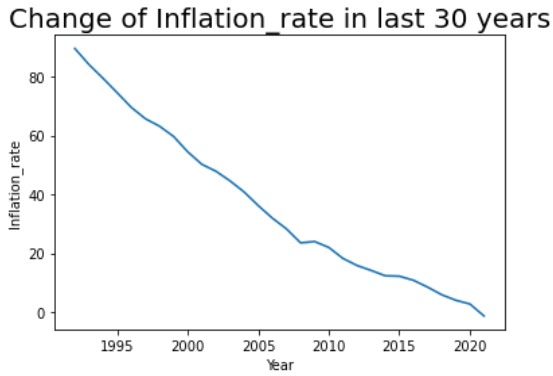


In order to understand the value distribution. Boxplot was used to figure out the minimum, the maximum, the sample median, and the first and third quartiles of the price of corn per month (the left one) and the overall inflation rate in the last 30 years(the right one). From the boxplot. We can infer that for the price of corn per month, the highest recorded outlier is at ~175; the lowest recorded outlier is above 100; and the median is ~1

50. For the overall inflation rate, the highest recorded outlier is lower than 60; the lowest recorded outlier is above 10; and the median is ~30.



Apart from that, a bar plot was also used to illustrate the price trend of wheat in the last 30 years. Based on the bar plot, we can conclude that in 1992-1996, 1999-2004, 2005-2008, 2009-2012, 2017-2022, the price of wheat is increasing while in 1996-1999, 2004-2005, 2008-2009, 2012-2017, the price of wheat is decreasing.

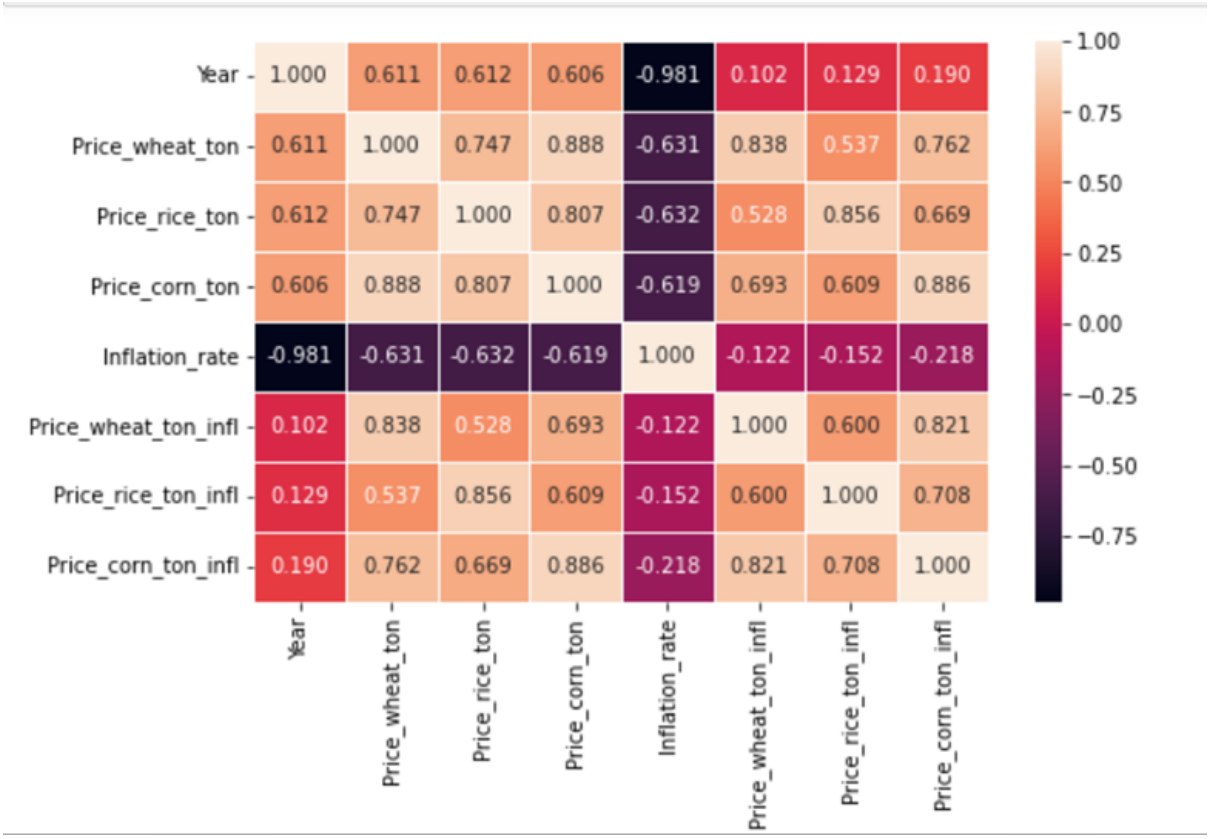


In addition, we also noticed an attention-worthy phenomenon which is that the inflation rate keeps decreasing in 1995-2020.

Exploration & Correlation (Extension 1)

To build a more accurate linear regression model, we first analyzed the correlation between variables. The linear regression model that we build will predict the inflation rate based on the rest of the variables. We used the Pearson method to perform our correlation heat map.

As We can see from the heat map, the most correlated two variables are Year and Inflation Rate and has a negative correlation which means as the Year increases the Inflation Rate decreases. The price of wheat, corn, and rice per ton are highly correlated to each other too, and have positive correlation. Correlation summarizes the strength and direction of the linear (straight-line) association between two quantitative variables(GeeksforGeeks, 2019).



Linear Regression (Extension 2)

We used the linear regression model to predict the inflation rate, so, in our linear model, the Inflation Rate will be our output, and other variables will be the input. And we set our train set as 60% of our data, the rest will be the test set to test how our model is performed. The intercept of our model is 2925.24. The accuracy of our model is 0.9792, which we used the score() to compute.

The elevation of our model is performed by using the predicted dataset. As shown below, our R squared which gives the information about how many predicted points fall on the regression line is 97.96. Our MAE of our model

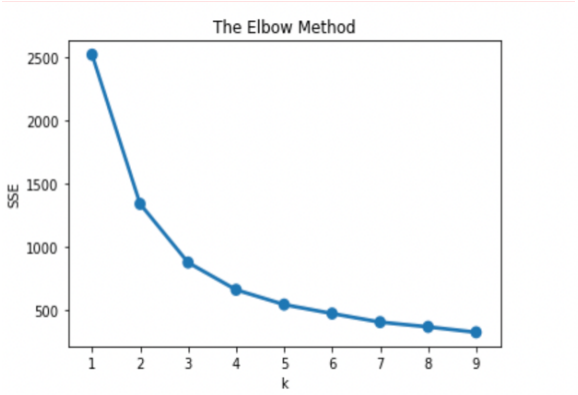
which shows the difference between the actual of true values and the predicted values is about 2.8018. Our MSE that tells the average of the square of the difference between the original and predicted values of the data is 14.8302. The RMSE which is the standard deviation of the errors which occur when a prediction is made on a dataset is 3.851.

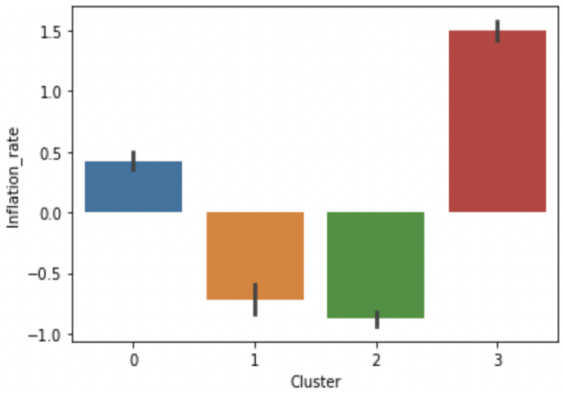
Overall, the Multiple Linear Regression model performs well, because 97.96% of the predictions fit the regression model. But the mean absolute error, mean square error, and the root mean square error are on the average level.

Clustering (Extension 3)

In order to better understanding our inflation rate, we use the K-mean algorithm do the clustering. The first step is find out the best k value by using elbow chart. We plotted the model of k from 1 to 9 and found the most suitable k is 4 through the chart.Next, we are divide the price_wheat_ton, price_rice_ton, price_corn_ton and inflation rate by cluster of mean. Since the inflation rate and price of wheat doesn't use same unit so we are normalized them and we get our bar plot for the inflation rate for our 4 groups. The cluster 0 is middle inflation rate group inflation rate from 0-0.5, the cluster 3 is high inflation rate from 0-1.5, cluster 2 and cluster 1 have similar inflation rate but the cluster 2 have larger range of It, the range is from 0 to -0.8. Since the trend of inflation rate is continually dec

reasing from 1922 to 2020, we know the negative of inflation is most likely happen in recent year.





In addition, the price of cereal will be affected by various factors, around 1990's agricultural industrialization had not been popularized, and the production of food had not fully met people's daily supply needs, so the price was greatly impacted on the environment at that time, such as weather, politics, international trade policy and other factors and Now the supply and demand of rice, wheat, corn are in balance so Inflation rate will not be as high as before.

Conclusion

As we shown above, we try to gain more understanding of our dataset by analyzing the useful variables, try to figure out the changes in the cereal inflation rate over the past three decades and the relationship between different cereal by different method: linear regression, clustering and correlation heat map. Through these methods, we know the relationship between the year and the inflation rate, using Linear Regression Model to predict the cer

eal prices in next couple years. We used two machine learning methods in th is project, supervised learning and unsupervised learning.In the future, we hope to use reinforcement learning to further study the relationship betwee n grain prices and inflation rate .