Project 2

Group 3: Reha Patel, Niraj Sai Prasad, Sindhu Swaroop

12/5/2021

Introduction

In this report, we will be investigating prices obtained from developing world markets for various goods as obtained by the World Food Program. We acquired this dataset from Kaggle and it contains information such as country name, market name, commodity name, commodity price, etc. We will be analyzing the probabilities of specific events, how clusters form within the data, any semantic information hidden in text, and trends in prices over time.

Prior to beginning any analysis, we set the theme for the plots and graphs to be used throughout the analysis.

Data Wrangling

1. Discovering

Before we begin our analysis, it is imperative that we have an understanding of the dataset itself. In order to do this, we will start off by listing out the names of the columns as well as viewing a single row of each columns using the head function.

```
##  [1] "adm0_id"            "adm0_name"          "adm1_id"
##  [4] "adm1_name"          "mkt_id"             "mkt_name"
##  [7] "cm_id"              "cm_name"            "cur_id"
## [10] "cur_name"           "pt_id"              "pt_name"
## [13] "um_id"              "um_name"            "mp_month"
## [16] "mpyear"             "mpprice"            "mp_commoditysource"
## [19] "usd"
```

```
##   adm0_id   adm0_name adm1_id  adm1_name mkt_id mkt_name cm_id cm_name cur_id
## 1       1 Afghanistan     272 Badakhshan    266 Fayzabad    55   Bread     87
##   cur_name pt_id pt_name um_id um_name mp_month mpyear mpprice
## 1      AFN    15  Retail     5      KG        1   2014      50
##   mp_commoditysource       usd
## 1                WFP 0.4830918
```

Here we see that the columns names are not descriptive and can potentially cause confusion later in the investigation. Later in the data wrangling process we will ensure that we rename the columns to be more descriptive of the data stored in them. We also see that there are columns for country name as well as the commodity being sold. Because this dataset only includes developing countries, it is important to understand how many distinct countries and commodities are found in it. We will do this by checking for the count of each distinct country (adm0_name) and commodity (cm_name).

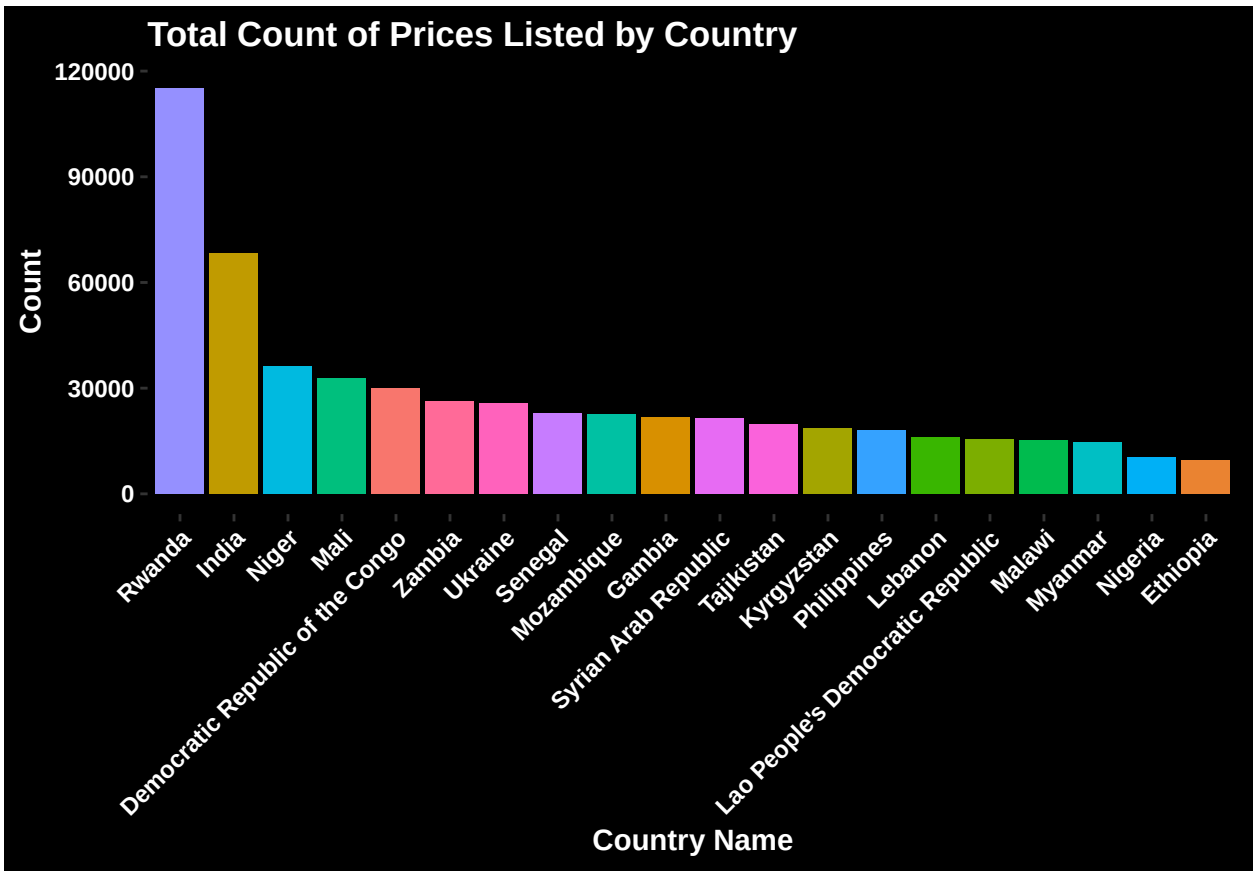
Table 1: Total Unique Countries Listed

total_records
74

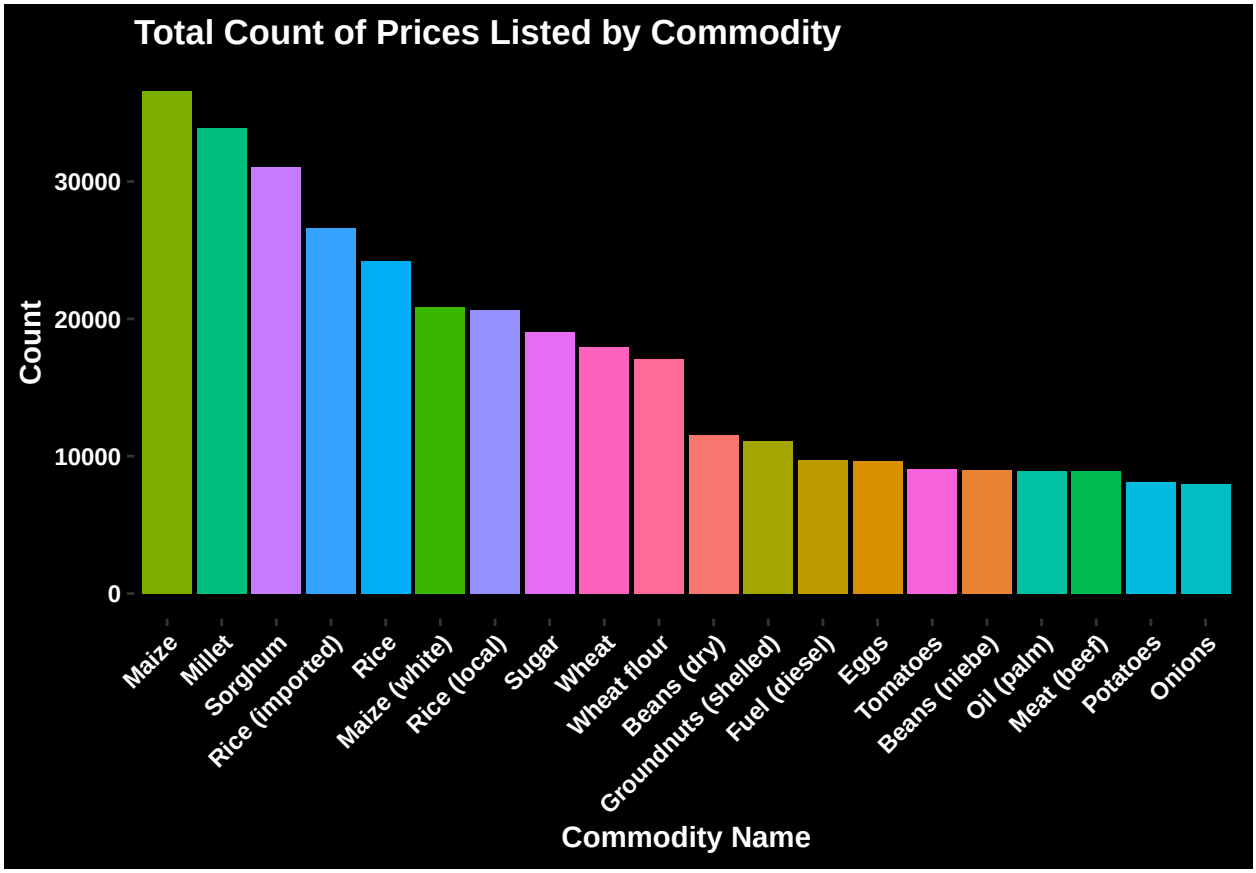
Table 2: Total Unique Commodities Listed

total_records
321

The data source on Kaggle told us that it focuses on markets in developing countries. Using the head() function above we saw that Afghanistan was one of the country's listed, but now we will see how many records are listed for each of top 20 of the 74 distinct countries. This is important because later in our analysis, we may want to focus on one country and study the relationships within this country.



Based on the table above, we see that Rwanda, India, Niger, Mali, and the Democratic Republic of the Congo are the five countries with the most commodity observations in the dataset. If we choose to perform an analysis on a single country or on a group of countries, these five would be viable options. Next we will look at similar counts, but instead we will look at the top 20 by the type of commodity.



Of the 321 unique commodities found in the dataset, these are the top 20. Similar to the top countries, these 20 commodities would be suitable for individual or group analysis.

2. Cleaning

To begin the cleaning process we will first address the column names. As mentioned previously, the column names are not descriptive and can potentially cause confusion later in the investigation. In order to avoid any confusion, we will rename the columns to more appropriate values.

```
##  [1] "Country_ID"       "Country_Name"     "Locality_ID"      "Locality_Name"
##  [5] "Market_ID"        "Market_Name"      "Commodity_ID"     "Commodity_Name"
##  [9] "Currency_ID"      "Currency_Name"    "Market_Type_ID"   "Market_Type"
## [13] "Measurement_ID"   "Unit_of_Goods"    "Month"            "Year"
## [17] "Price"            "Commodity_Source" "USD"
```

When observing the different rows in the dataframe, we noticed that some locality names were preceded by a dollar sign. In order to avoid any errors later in our analysis because "$" is also a symbol used in R, we will remove it from the data.

```
##   Country_ID Country_Name Locality_ID Locality_Name Market_ID Market_Name
## 1          1  Afghanistan       99878     $Daykundi       275        Nili
##   Commodity_ID Commodity_Name Currency_ID Currency_Name Market_Type_ID
## 1           55          Bread          87           AFN             15
##   Market_Type Measurement_ID Unit_of_Goods Month Year Price Commodity_Source
## 1      Retail              5            KG     1 2014 51.72              WFP
##         USD
## 1 0.4997102
```

```
##  [1] Country_ID       Country_Name     Locality_ID      Locality_Name
##  [5] Market_ID        Market_Name      Commodity_ID     Commodity_Name
##  [9] Currency_ID      Currency_Name    Market_Type_ID   Market_Type
## [13] Measurement_ID   Unit_of_Goods    Month            Year
## [17] Price            Commodity_Source USD
## <0 rows> (or 0-length row.names)
```

3. Enriching

As a part of the enriching process, we converted the prices found in the dataset to USD. This step took about 3 hours to complete, so it was done prior to loading the dataset in order to avoid having to constantly run the block. This was important to our analysis because it would allow us to compare the prices of commodities across countries where the currencies may differ. Here is the code which was used to convert the commodity prices to USD:

In addition to this, we checked for instances when the price of the commodity equaled 0 or NaN. This would impact the average prices of a commodity so it was important to remove it. Additionally, we removed country names, country IDs, locality names, locality IDs, commodity names, and commodity IDs which were NaN. Including these could potentially impact our analysis so it was important to remove them.

```
## [1] FALSE  TRUE
```

```
## [1] FALSE
```

4. Validating

Finally, we will perform some validating steps on our dataset. This will include ensuring that the NaN values discovered above were truly removed. We will also ensure that the prices of commodities were non-negative numbers because that would indicate errors in the data.

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## [1] FALSE
```

Business Questions

Probability

Question 1: What is the probability that the price trends across all countries in this dataset are descreasing over the years?

We first selected all the unique countries in our dataset, and then found the average of the earliest prices of commodities for a particular country. We compared this with the most recent prices of commodities for the same country by assigning 1 if its greater, and 0 if lesser - then based on count of 1s, we found the required probability.

```
## [1] 0.1692308
```

Observations & Conclusions We made the following observations from the calculation above:

- The probability that prices decrease over time across all countries is ~ 0.17.
- This is an expected value because the dataset has mostly developing countries. Inflation leads to increase in prices.
- 17% of the countries that do show decreasing trends could be improving in terms of trade, localized farming, and many other factors.

Question 2: In the United States it is not uncommon for prices of goods and services to be higher in bigger cities. Do developing countries face a similar scenario? What is the probability that commodities are more expensive in the capital city than in the other cities across all countries in the dataset?

First we wrote code to get K values. K values sum up to find the final probability for all countries. The function basically adds the capital city data column to our main dataframe and assigns the isCapital values [0,1]. This dataframe also returns the h value to store 0s and 1s into a list.

Next we wrote the main code for BQ2. We took the country capitals from an external dataset. We merged this with the original dataset. Upon calling the function, we got the required probability.

```
## [1] 0.3448276
```

	Country_Name	Avg(Capital) > Avg(Rest)
1	Afghanistan	1
4677	Algeria	0
6461	Armenia	1
12239	Bangladesh	0
15274	Benin	0
18726	Bolivia	1
23173	Burkina Faso	0
31956	Burundi	0
35803	Cambodia	0
39763	Cameroon	0

Observations & Conclusions We made the following observations from the graph and calculation above:

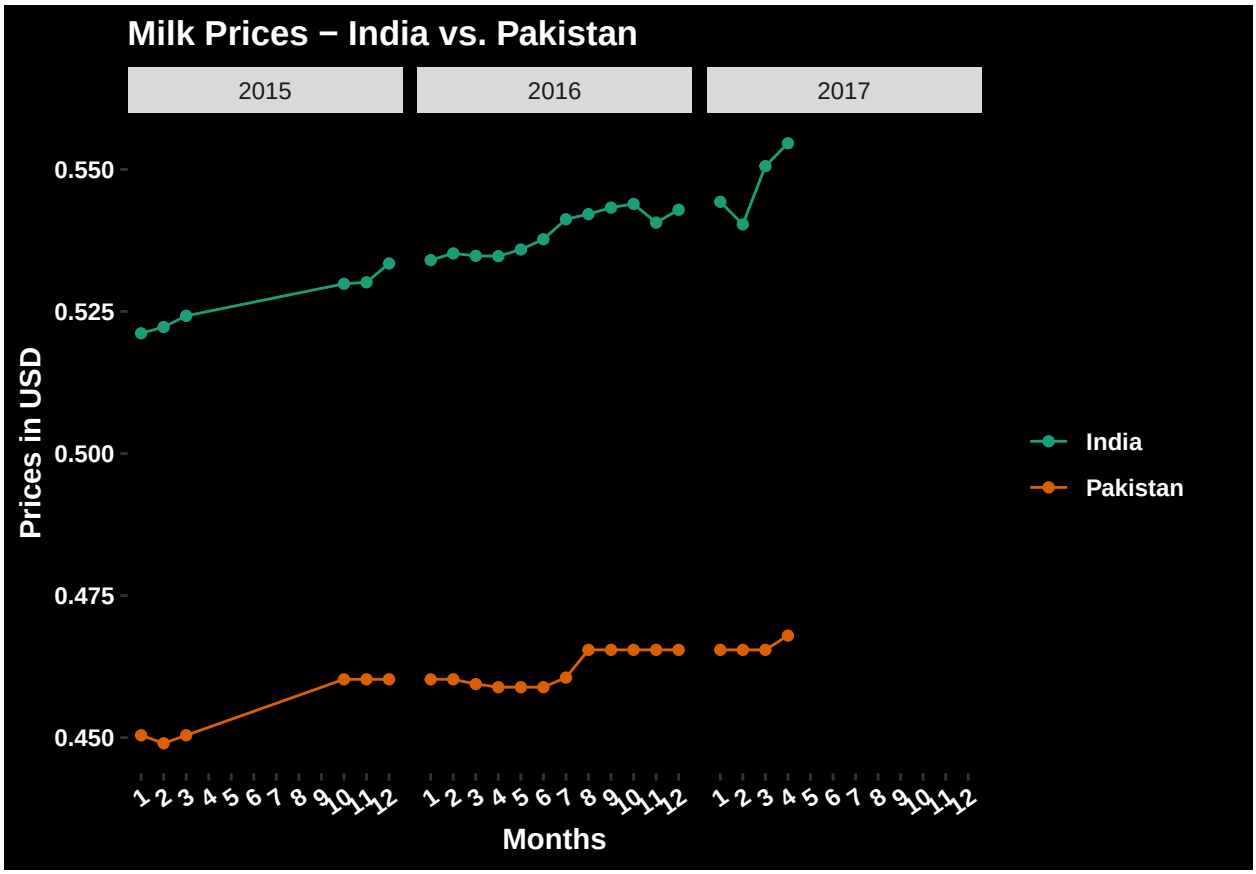
- In the US, commodities in the larger cities are slightly more expensive than in rural or suburban areas. Since most countries in our dataset are still developing, we tried to see if a similar commodity pricing trend is present.
- We found that almost 0.34 of the countries in our dataset have commodity pricing higher in the capital cities than in rural areas. This number again depends on a lot of factors (eg: geographic), so it is just a rough estimate.
- For instance, in India, New Delhi tends to average out lesser than the other areas. But in Afghanistan, prices in Kabul are much higher than in other districts.

Question 3: A commodity should be priced almost equally across the world. But this is seldom the case owing to various reasons like political and economic storms on distant continents. Milk is one such commodity that is processed, refined and traded across the globe. What is the correlation between prices of milk in different countries?

First, we found the correlation coefficient between the prices of milk in India and Pakistan over 3 years (2015-2017). We did this by filtering out only the rows corresponding to the commodity milk in the two countries, and making sure the month and year data is identical for both countries.

```
## [1] 0.9084655
```

Next, we plotted the milk price variations in the two countries by month and faceted by year, for a clearer picture.



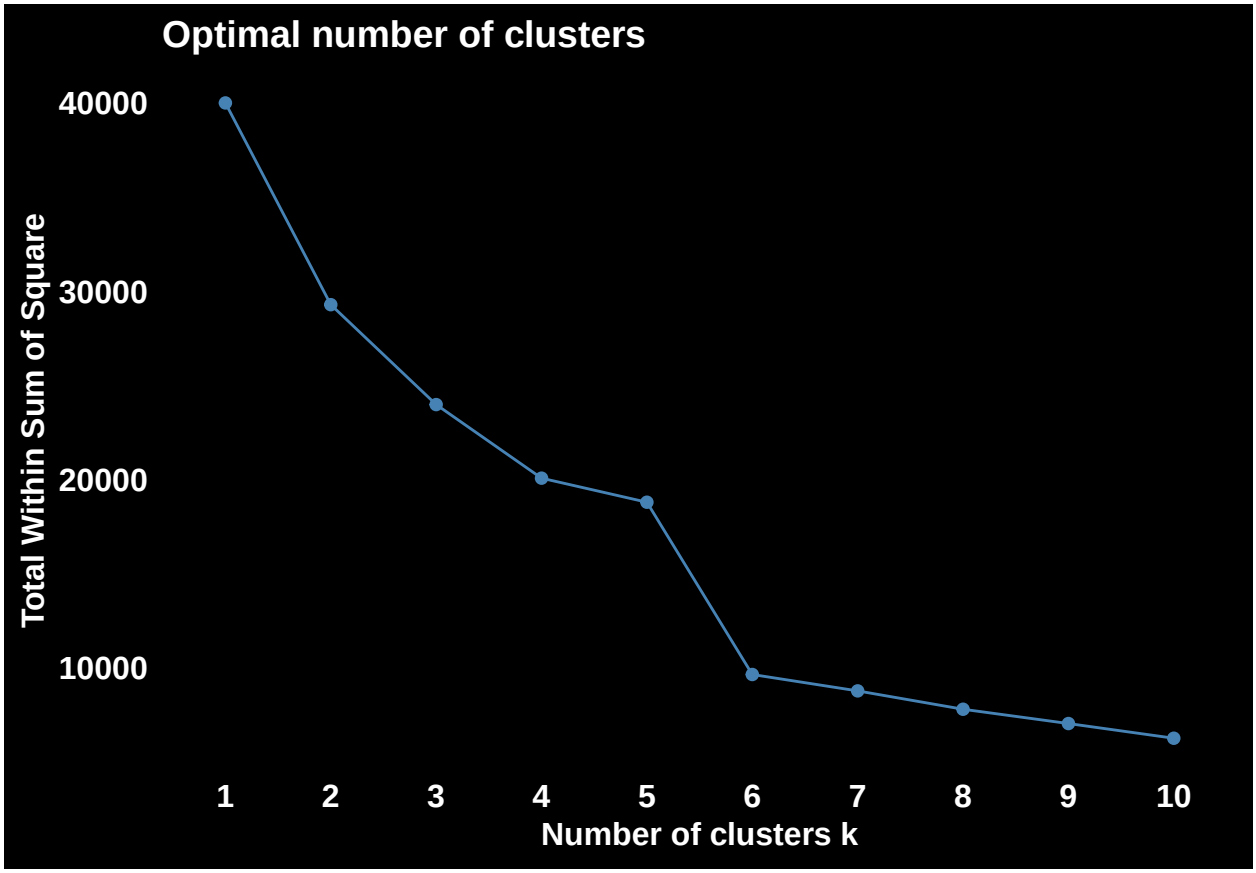
Observations & Conclusions We made the following observations from the graph above:

- India and Pakistan are neighboring countries, so we would expect the milk prices to be roughly the same. However, on plotting the prices we observed that this is not the case.
- There is a difference of 10 cents (in US currency) between the two countries, which amounts to a great deal when converted to local currencies, since milk is a daily commodity. However, despite the 10 cent difference, it appears in the plot that the price of milk tends to increase and decrease around the same time in both countries.
- Nevertheless, we found a strong positive correlation between the milk prices in the two countries. Being on the same continent and next to each other, economic and political disturbances impact the prices in a similar way. The correlation coefficient was 0.908.

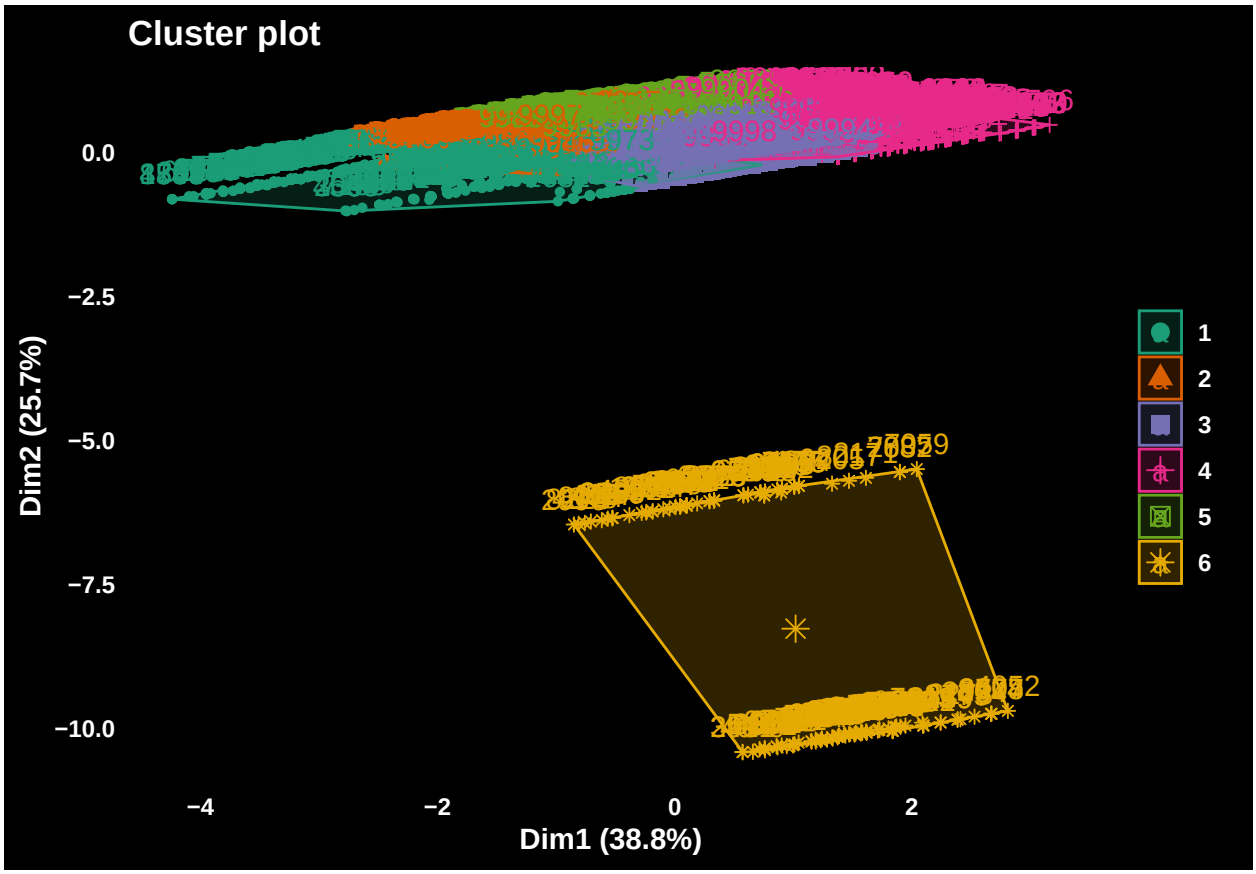
Clustering

Question 4: Oftentimes sellers and consumers would think that commodities across countries would be priced similarly, and thus clustered together. Use k-means to determine not only the optimal number of clusters, but also to determine whether there are any meaningful clusters across commodities and countries.

First we had to run fviz_nbclust to determine the optimal number of clusters. We can find this "optimal number" based on where the elbow occurs in the graph below We see that at k = 6 the total within sum of squares seems to taper off. As a result, when we run k-means we will use k = 6.



Next, we fitted the k-means clustering algorithm to the scaled data and visualized the clusters. In addition to this, we used an aggregate function to display the means of every cluster which helped us draw conclusions about the clustering.



```
##   cluster Country_ID Commodity_ID Currency_ID     Year
## 1       1   156.0019     108.4542    51.59844 2002.376
## 2       2   141.7373     102.5695    34.40906 2008.966
## 3       3   177.7805     109.4540    74.67197 2013.689
## 4       4   204.7065     345.6398    74.19159 2014.130
## 5       5   140.6133     111.8579    33.71359 2014.299
## 6       6 56833.1908     128.2824    79.39695 2011.634
```

Table 4: Countries with the Highest Country IDs

Country_ID	Country_Name
70001	South Sudan
40764	Sudan
999	State of Palestine
271	Zimbabwe
270	Zambia

Table 5: Commodities with the Highest Commodity IDs

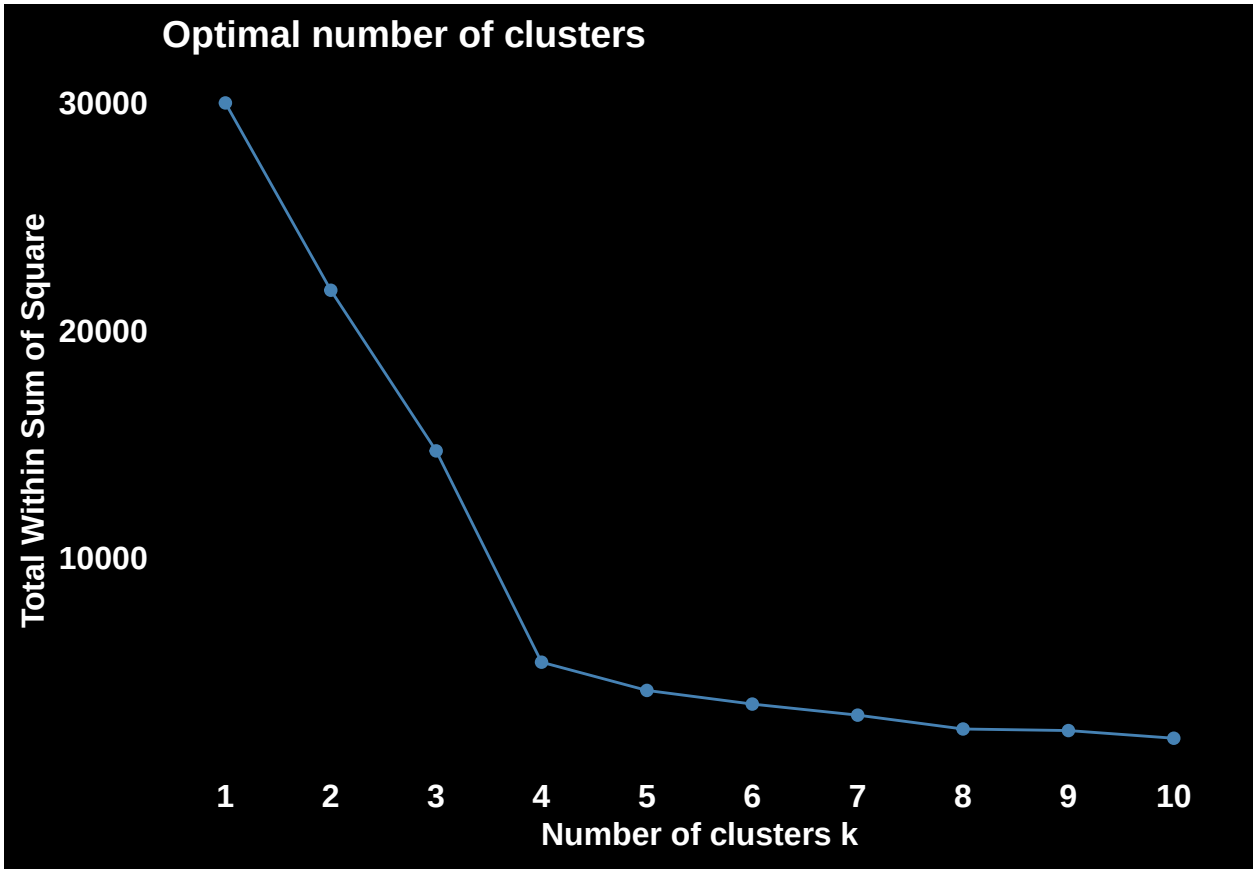
Commodity_ID	Commodity_Name
490	Beans (niebe, white)
489	Water (drinking)
488	Cotton
486	Fish (frozen)
484	Yam (Abuja)

Observations & Conclusions We made the following observations from the clustering analysis above:

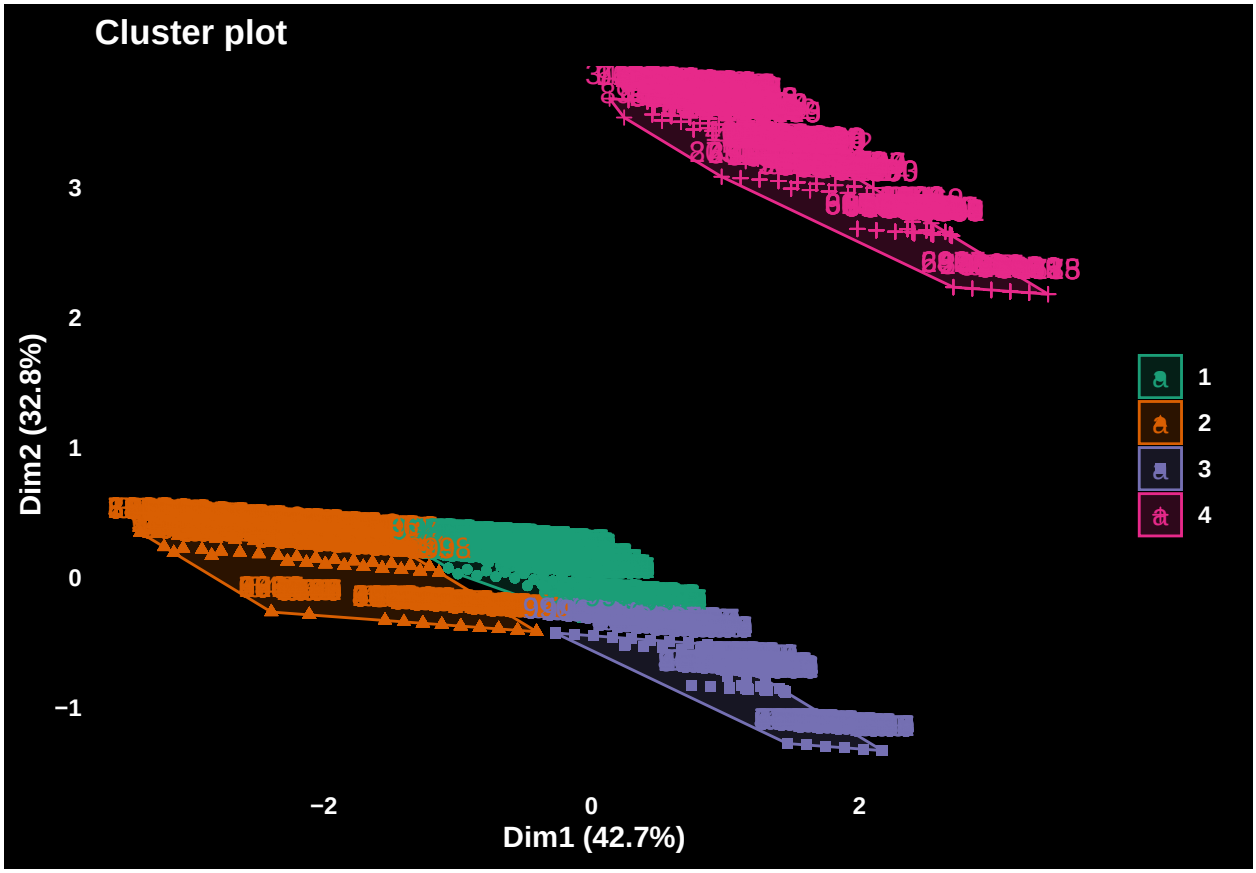
- There are 6 color coded clusters which can be seen in the cluster plot above. Cluster 6 is distinct from clusters 1, 2, 3, 4, and 5.
- Based on the aggregate function, we see that the mean Country_ID is around 50000 for one of the clusters. Based on the table Countries with the Highest Country IDs shown above, we can assume that the countries in this cluster would include South Sudan, Egypt and Sudan.
- An interesting observation about the previous point is that Egypt, Sudan, and South Sudan are all neighboring countries in Africa. It is possible that due to the close proximity that these countries have similar cultures and thus sell similar types of commodities in their markets.
- The Commodity_ID for one of the clusters is around 300, which is much higher than the other clusters. After looking at the Commodities with the Highest Commodity IDs table, we can assume that some of these are found in this cluster, which is driving up the mean commodity ID.

Question 5: When examining a country specifically, it's possible to see clusters appear by region. Using the localities and commodities found in India, perform k-medoids on the data to determine if the clusters can be identified as being regional.

Similar to the previous business question, we looked for the elbow in this graph when the total within sum of squares seems to plateau off. Based on the graph below, it seemed that the elbow happened at $k = 4$, so when we ran k-medoids, we used 4 as the optimal number of clusters.



Next run the scaled data that includes localities/regions on the k-medoids/PAM algorithm and output the means of each of the 4 clusters. Similar to above, this will help us when drawing conclusions about why the data was clustered the way it was.




```
##   cluster Locality_ID Commodity_ID     Year
## 1       1    1498.788     100.7879 2014.227
## 2       2    1498.013     109.0659 2002.060
## 3       3    1498.632     320.8858 2014.767
## 4       4   70077.855     175.8247 2014.595
```

Table 6: Indian Localities with the Highest Locality IDs

Locality_ID	Locality_Name
70082	Uttarakhand
70080	Puducherry
70078	Jharkhand
70074	Chandigarh
1511	West Bengal
1510	Uttar Pradesh
1509	Tripura
1508	Tamil Nadu
1506	Rajasthan
1505	Punjab

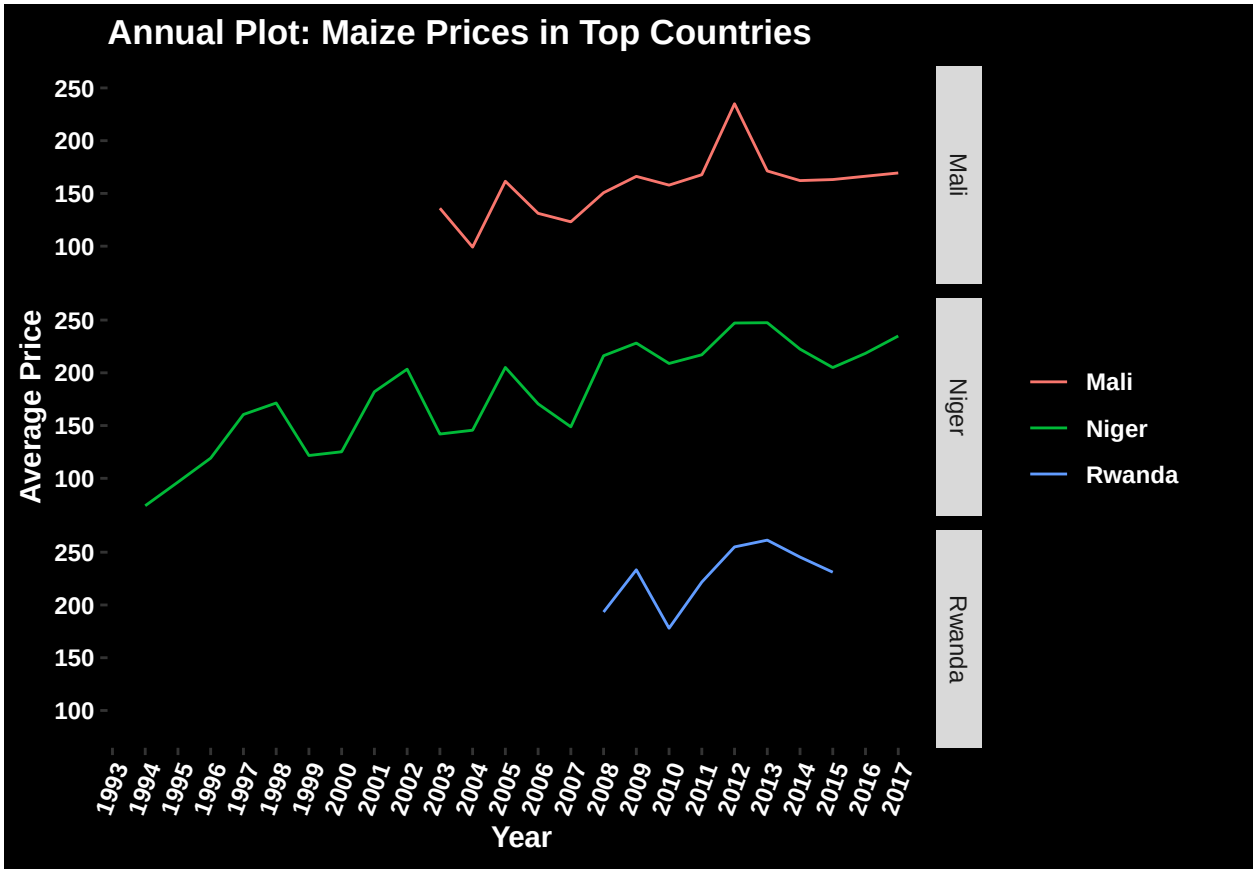
Observations & Conclusions We made the following observations from the clustering analysis above:

- There are 4 cluster plots produced by the analysis above. While each of the clusters seems to be distinct, there appears to be some overlap between clusters 1 and 3.
- The average locality ID for cluster 4 was much higher than the other three clusters. When looking at the table Indian Localities with the Highest Locality IDs, we can assume that cluster 4 may have Uttarkhand, Puducherry, Jharkhand and/or Chandigarh which is causing the average to be so high.

- After examining a map of India, it is noted that 3 of the 4 localities with the highest locality IDs are found in Northern India. It is possible that due to being in a similar region, these localities sell similar types of commodities in the markets.
- Similar to the previous business question one of the clusters, cluster 3, has a high average commodity ID. We can assume that some of the commodities in cluster 3 are found in the Commodities with the Highest Commodity IDs table.

Time Series Analysis

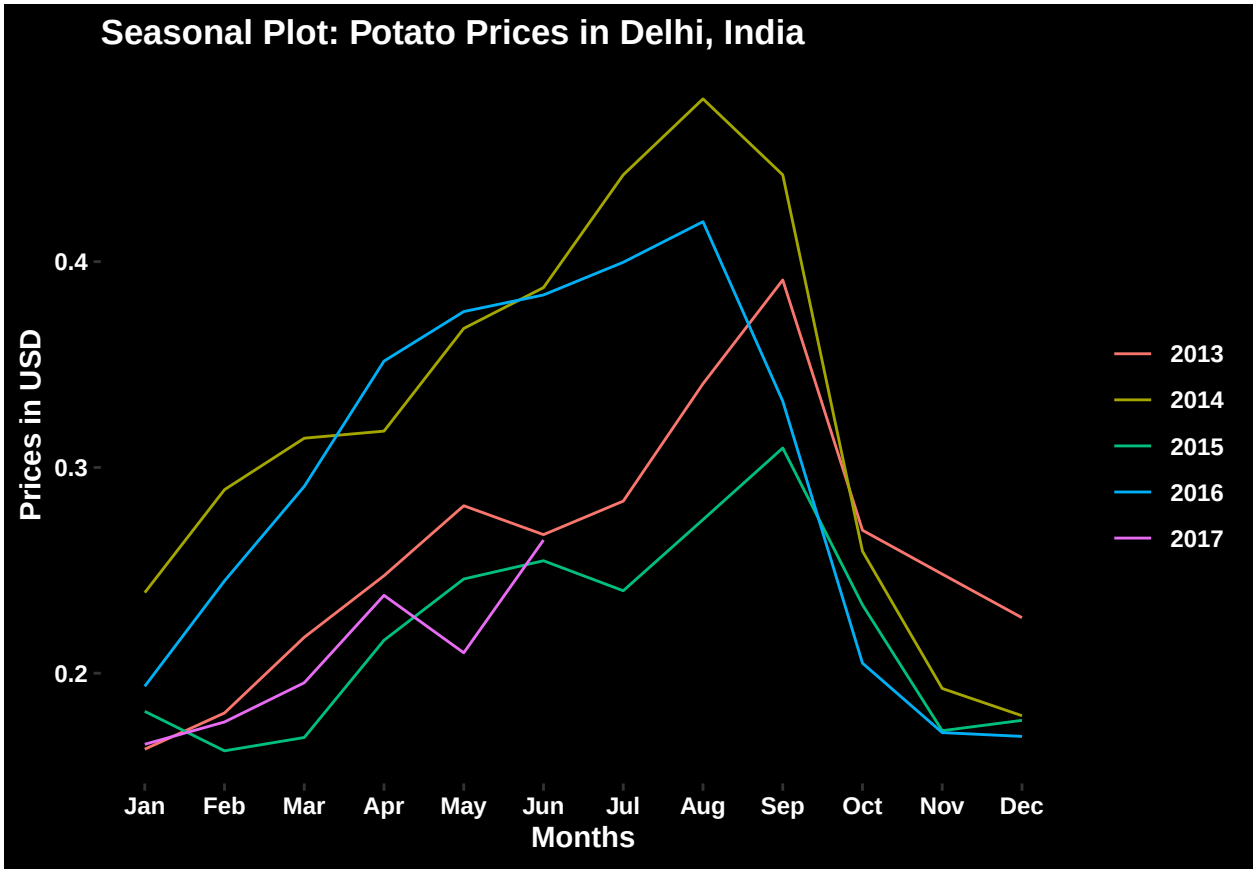
Question 6: Maize is the commodity that is seen most often in the dataset and this could mean that it is available in most of the developing countries. Plot the time series of the price of maize in the 3 countries with the greatest number of commodities listed in the dataset.



Observations & Conclusions We made the following observations from the time series line graph above:

- The overall trend of the price of maize has been increasing regardless of when the data begins.
- While the overall trend is that the price is increasing, we see that the price of maize tends to dip and spike around the same year regardless of the country. In fact, a spike for a year or two is typically followed by a drop for a year or two.
- In 2010 we see a dip in all 3 countries and in 2012 we see a spike in all 3 countries. After doing further analysis, it was discovered that a drought in the U.S. in 2012 led to a surge in global corn prices.

Question 7: Potatoes are a popular commodity found in many countries. Because it is a seasonal crop, potatoes may display fluctuations throughout a calendar year. Plot the time series by month of the price of potatoes to see if there is a specific time of year when it is most expensive in the markets.



Observations & Conclusions We made the following observations from the time series line graph above:

- Potatoes are a popular vegetable in Northern India - so we chose Delhi for our time series analysis. The time series of Potato prices in Delhi follows a similar seasonal distribution across 5 years from 2013 till 2017. The potato prices peak once in April/May, rise higher in August/September and are low at the end of the year as well as the beginning of the year.
- Potatoes require cool but frost-free weather, so in tropical regions such as India, they are mostly grown in winters. This explains the lows in the price from November to February, the "season" for potatoes.

- April/May in India is the peak summer season, scorching hot temperatures are unfavorable for cultivation of potatoes. This justifies the first peak in the prices.
- August/September is when the monsoon season kicks in. Heavy rains make the soil wet and cause the potato seeds to decay, hence throwing light on why potato prices in India peak during this time of year.

Text Analysis

Question 8: We have observed that certain commodities are more predominant in specific countries. Based on this, perform a text analysis to determine statistical measures such as term frequency. Doing this will allow consumers and sellers to determine the most common commodities in a country.

First, we found the commodities sold in India, tokenized them into one word per row, and removed the duplicates. The resulting data frame had just the commodity names.

Next, we filtered out all the news headlines related to the commodities sold in India, from the news headlines dataset. Then, we tokenized them, removed all the stop words and performed sentiment analysis using the method - "nrc" on the data frame.

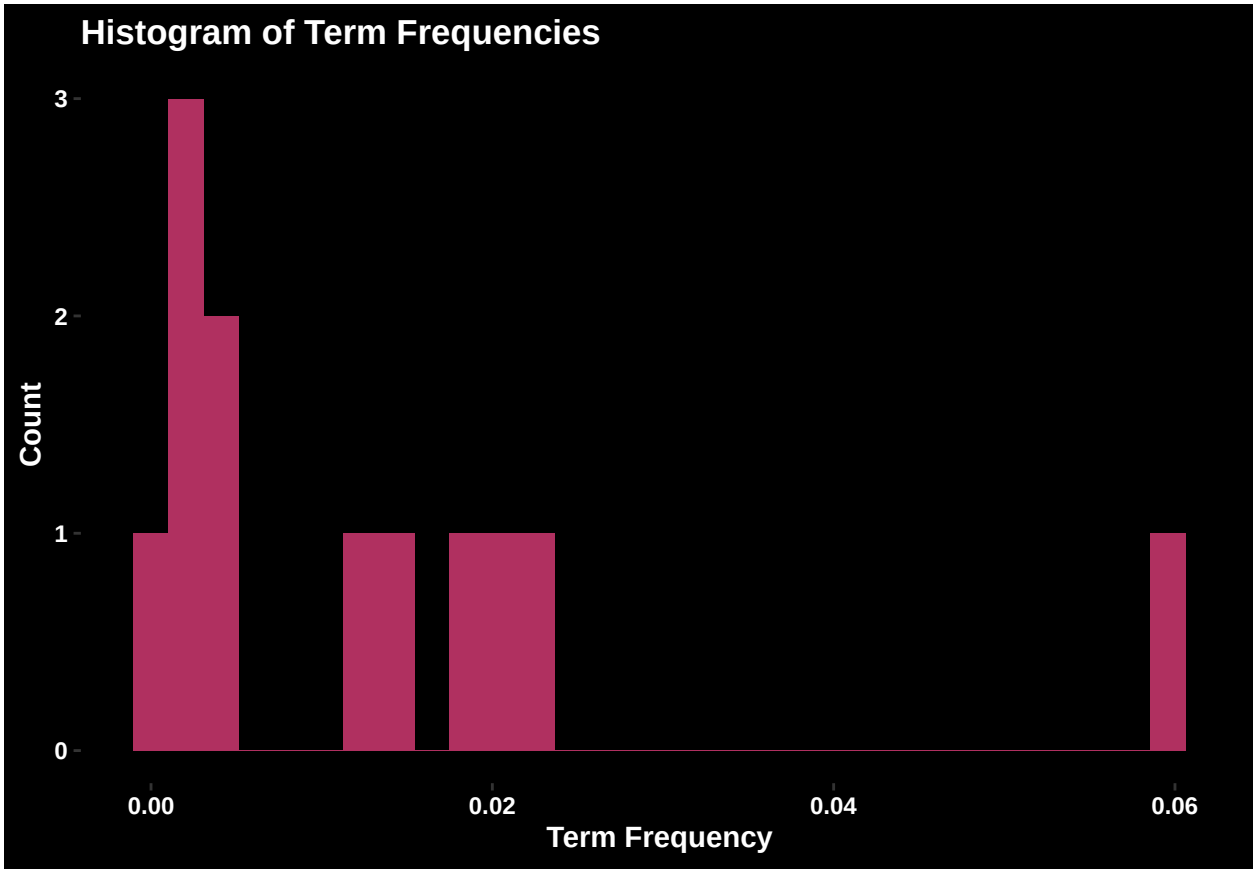
Table 7: Sentiment Analysis of Commodity News Headlines

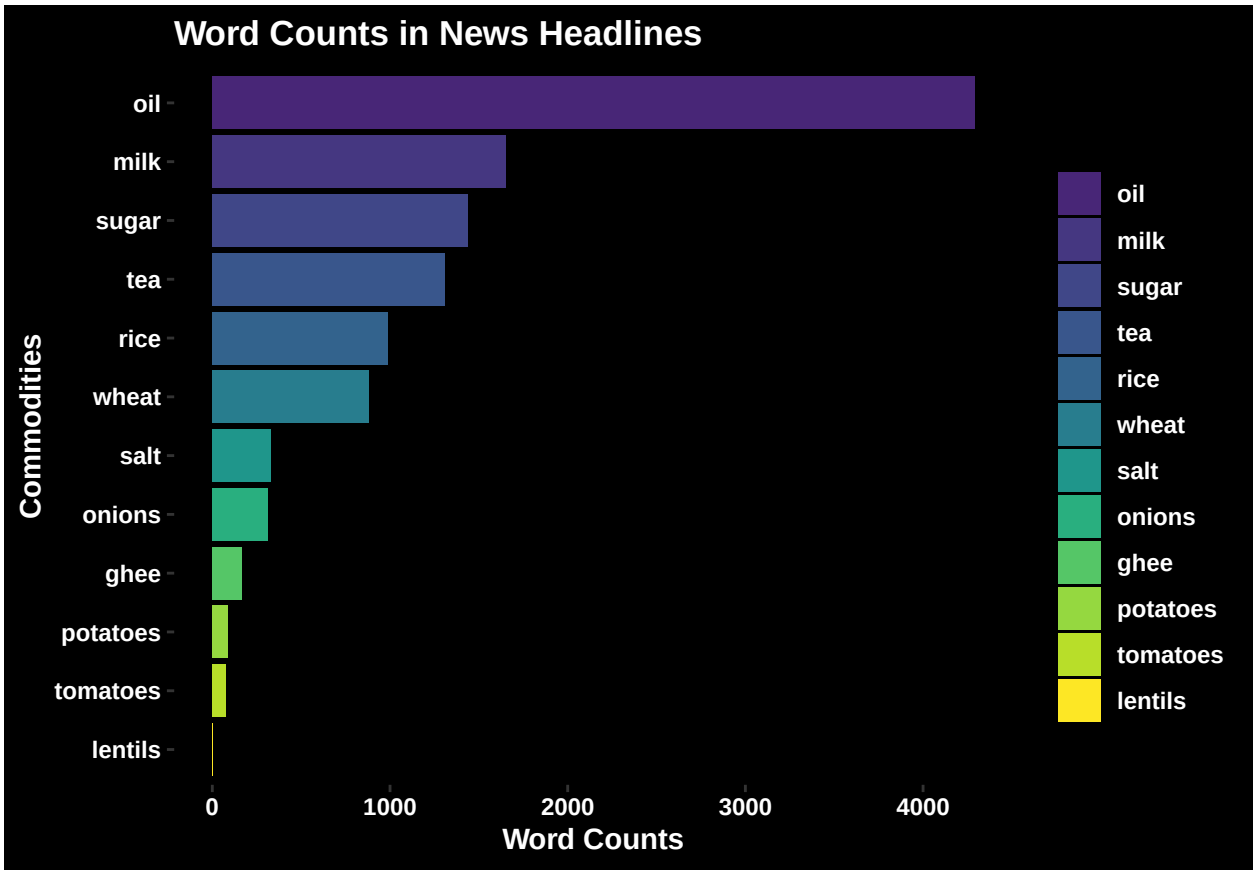
index	joy	positive	anger	negative	sadness	anticipation	trust	surprise	disgust	fear	sentiment
1	1	1	0	0	0	0	0	0	0	0	1
3	0	1	0	0	0	0	0	0	0	0	1
4	0	0	1	1	1	0	0	0	0	0	-1
5	0	2	0	0	0	1	1	0	0	0	2
9	1	1	0	0	0	1	0	1	0	0	1
10	1	1	0	0	0	1	0	1	0	0	1
11	0	1	0	1	0	1	0	0	1	0	0
12	0	0	1	1	2	0	0	1	0	0	-1
14	0	0	0	1	0	0	0	0	0	0	-1
16	0	1	0	0	0	0	0	0	0	0	1

We then calculated the term frequency of each of the commodity words in the commodity news headlines by finding the total number of words and dividing the count of each commodity word by the total. We also plotted the histogram of term frequencies and a bar chart of word counts of each commodity for visualization.

Table 8: Term Frequency of Commodity Words in Commodity News Headlines

	word	n	total	term_frequency
1	oil	4289	71941	0.0596
2	milk	1653	71941	0.0230
3	sugar	1440	71941	0.0200
4	tea	1309	71941	0.0182
5	rice	989	71941	0.0137
6	wheat	882	71941	0.0123
12	salt	332	71941	0.0046
13	onions	315	71941	0.0044
28	ghee	166	71941	0.0023
94	potatoes	85	71941	0.0012
111	tomatoes	75	71941	0.0010
8666	lentils	1	71941	0.0000





Observations & Conclusions We made the following observations from the text analysis above:

- The news headlines in India feature commodities many a times, with reference to the increasing prices and exports/imports to name a few. We found an Indian news headlines dataset and performed text analysis on this which revealed some interesting results.

- We performed Sentiment Analysis on commodity news headlines and found a blend of positive and negative news. Negative news included headlines on increasing prices, smuggling rackets and food adulteration. Positive news took into consideration headlines on arresting of smugglers, government undertakings of distribution of free commodities to the poor, best trading firms and so on.

- We found the Term Frequency of only the "commodity words" in the commodity news headlines. From this we were able to infer that the most common commodities in India that make it to the headlines are Oil, Milk, Sugar and Tea in decreasing order of occurrence. The least common commodity in the headlines was Lentils.

- We plotted a histogram of Term Frequency to see which words' appearances are equally frequent. Oil had the highest term frequency, followed by milk, sugar and tea having term frequencies close to each other. Ghee, Potatoes and Tomatoes had almost the same term frequency, which means their appearance in the headlines was equally frequent - hence, the histogram bar of value 3.

- We also plotted a bar graph of Commodity names vs. word counts for a clearer visualization of occurrences of commodities in news headlines.

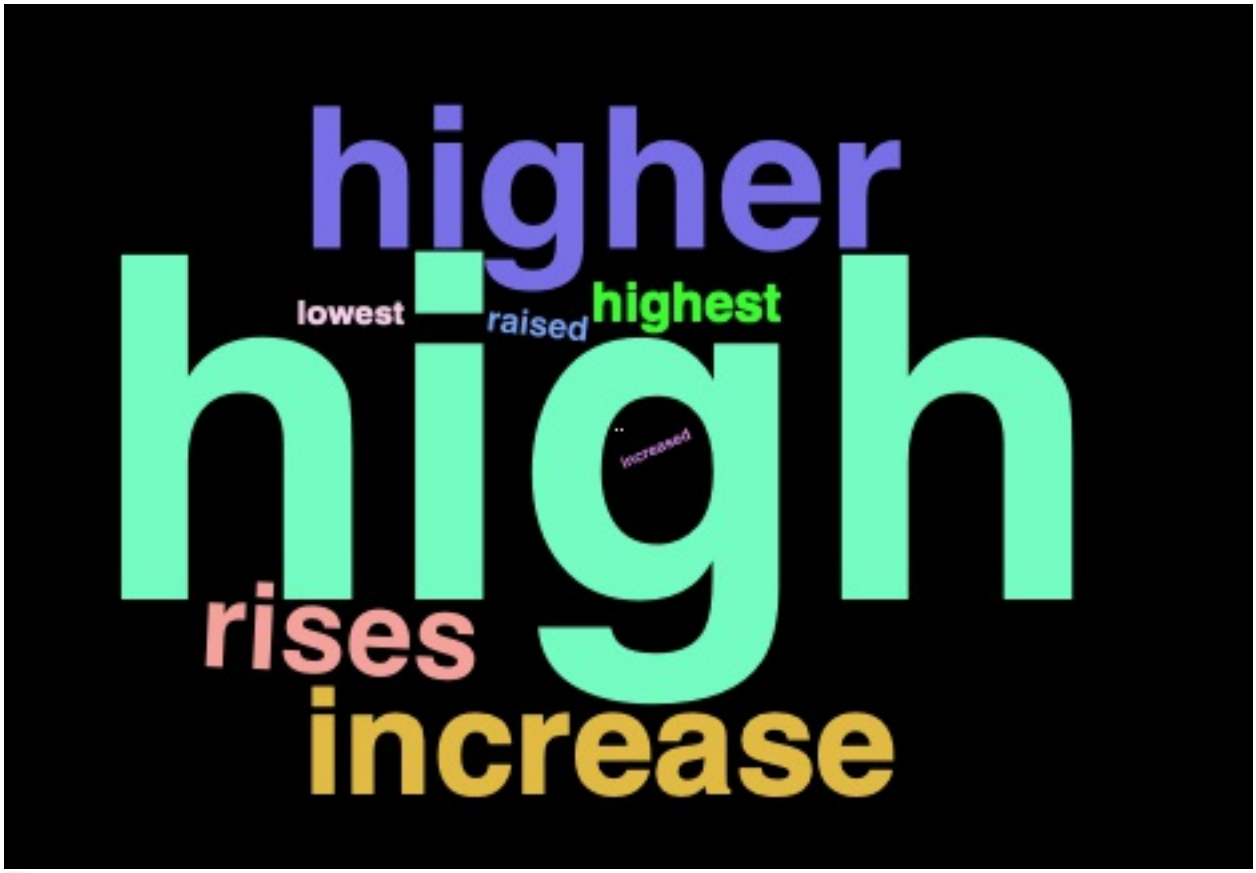
Question 9: India is one of the countries in the dataset with the most number of observations as well as news headlines. After coming across a dataset with the news headlines, determine which headlines are related to the commodities in the dataset. Of those, which headlines are related to the direction of price movements?

We first found the commodities sold in India, tokenized them into one word per row, and removed the duplicates. The resulting data frame had just the commodity names.

To find the increasing and decreasing trends, we keyed in a few popular terms that we know of to get a rough estimate and saved it into a df. We then added a new column with a 1 for increased trends, and 0 for decreased trends. We then split the bool values into a seperate column.

We plotted a wordcloud to get a visual of how the trends appear in our dataset. We also have a kable for number of increasing and decreasing trends.

Bool	count
0	79
1	271



Observations & Conclusions We made the following observations from the text analysis above:

- Based on the Kable, increasing trends are much higher than decreasing trends for Indian commodity prices.
- We can conclude that the price trends in India are usually increasing based on the popularity of words in the word-cloud.
- Since India is still not a developed country, we can expect this trend to continue for quite some time.
- Due to inflation, prices tend to go up. Once India becomes developed like the US and stabilizes, we can expect the prices to remain more or less the same.

Conclusion

Throughout this project we analyzed probabilities, where and how clusters form in the data, the presence of commodities in the news, and trends in prices over time. Consumers and sellers would be able to use our analysis to determine things like in which month potatoes would generally be most expensive to buy. While the seasonal difference in prices of commodities like potatoes may not make much of a difference to most individuals in developed countries, the difference may be significant enough for some individuals in developing countries. In addition to this, we found that countries around the world and localities in India were more likely to be clustered together if they were closer to each other. This would mean that countries and localities within one cluster had commodities that were more similarly priced. A major conclusion we drew throughout our analysis was that the general trend of commodity prices is increasing, which means inflation is high in developing countries. We saw this in the rise in prices of maize and milk, for example.