DataKind X Producers Direct DataKit 2025 - Dhanashree

Methodology

Data Source: The dataset used was provided for by the organization Producer's Direct (which they had acquired from WeFarm after their acquisition of the organization). It is a compilation of data sourced from farmers using their data sharing platform. This is a large dataset, containing 20,304,848 rows and a total of 24 columns. The dataset initially had several missing values, inconsistent data formats, and etc. Thus, the data had to be cleaned and transformed to be made suitable for analysis.

Python was used for this data analysis project. The Python packages that were used were *pandas, datetime, numpy* and *matplotlib*. No Generative AI was used throughout this analysis.

Data Cleaning & Transformation:

Due to the large nature of this dataset, a selection of a key area to guide the analysis seemed pertinent. Hence, the *topic* of the *questions* asked was used to steer the overall flow of analysis. Questions were favored over responses, as questions highlighted a need and a point of interest for the farmers. A question could have multiple responses over several topics in common. All the answers generated were still taken into account and were assumed to correspond to the frequency of the topics, and inherently the popularity of the topics being discussed.

Hence, out of the 24 columns, only a few were selected for every subsequent analysis to focus on certain aspects of the data and their relationship to the topics of the questions asked.

Columns:

```
df.columns
✓ [13] < 10 ms
Index(['question_id', 'question_user_id', 'question_language',
       'question_content', 'question_topic', 'question_sent', 'response_id',
       'response_user_id', 'response_language', 'response_content',
       'response_topic', 'response_sent', 'question_user_type',
       'question_user_status', 'question_user_country_code',
       'question_user_gender', 'question_user_dob', 'question_user_created_at',
       'response_user_type', 'response_user_status',
       'response_user_country_code', 'response_user_gender',
       'response_user_dob', 'response_user_created_at'],
      dtype='object')
```

Incorrect data types were found such as for the column “question_sent”. Instead of the data type “datetime” it was found to be an “object”.

```
df.info()
✓ [7] 17ms
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20304843 entries, 0 to 20304842
Data columns (total 24 columns):
 #   Column                    Dtype
---  ------                    -----
 0   question_id               int64
 1   question_user_id          int64
 2   question_language         object
 3   question_content          object
 4   question_topic            object
 5   question_sent             object
 6   response_id               int64
 7   response_user_id          int64
 8   response_language         object
 9   response_content          object
```

The desired columns were checked to see if they had any null values. The “question_topic” column had plenty.

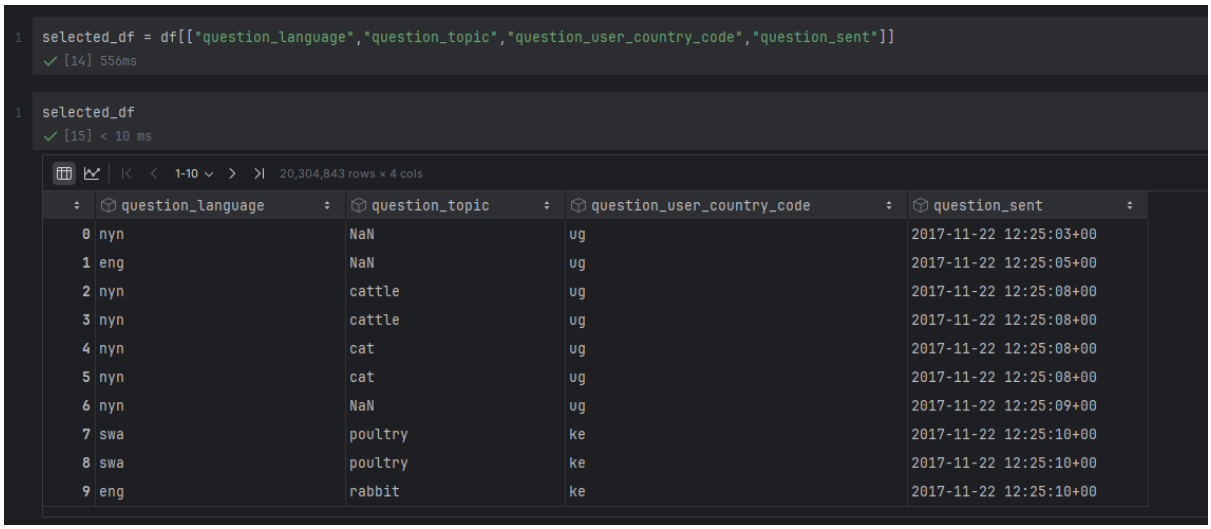
```
df["question_language"].isna().sum()
✓ [8] 568ms
np.int64(0)

df["question_topic"].isna().sum()
✓ [9] 541ms
np.int64(3537729)

df["question_sent"].isna().sum()
✓ [10] 556ms
np.int64(0)

df["question_user_country_code"].isna().sum()
✓ [11] 554ms
np.int64(0)
```

The targeted columns were then grouped into a separate dataframe.



```
selected_df = df[["question_language","question_topic","question_user_country_code","question_sent"]]
```
✓ [14] 556ms

```
selected_df
```
✓ [15] < 10 ms

1-10 · 20,304,843 rows × 4 cols

	question_language	question_topic	question_user_country_code	question_sent
0	nyn	NaN	ug	2017-11-22 12:25:03+00
1	eng	NaN	ug	2017-11-22 12:25:05+00
2	nyn	cattle	ug	2017-11-22 12:25:08+00
3	nyn	cattle	ug	2017-11-22 12:25:08+00
4	nyn	cat	ug	2017-11-22 12:25:08+00
5	nyn	cat	ug	2017-11-22 12:25:08+00
6	nyn	NaN	ug	2017-11-22 12:25:09+00
7	swa	poultry	ke	2017-11-22 12:25:10+00
8	swa	poultry	ke	2017-11-22 12:25:10+00
9	eng	rabbit	ke	2017-11-22 12:25:10+00

Any null values were dropped. The resulting dataframe was checked to see if it had any remaining null values. There were none.

```
selected_df = selected_df.dropna()
```
✓ [16] 2s 725ms

```
selected_df.isna().sum()
```
✓ [17] 1s 767ms

4 rows · 4 rows × 1 cols

	<unnamed>
question_language	0
question_topic	0
question_user_country_code	0
question_sent	0

A basic summary of the dataframe led to the targeted group having 16M+ rows, 4 unique question languages, 148 unique question topics, 4M+ unique questions, all collected from 4 countries. Moreover, the most frequent language was English, the most frequently occurring question topic was “maize” and the most questions came through from Kenya.

Also, as the columns had long names, they were changed to something shorter.

```
selected_df.describe()
```
✓ [18] 9s 48ms

4 rows × 4 cols

	question_language	question_topic	question_user_country_code	question_sent
count	16767114	16767114	16767114	16767114
unique	4	148	4	4190836
top	eng	maize	ke	2021-05-09 04:42:33.942853+00
freq	10371139	2201755	8534350	2984

```
name_map = {'question_language' : 'Q_LG',
            'question_topic': 'QS_TOPIC',
            'question_user_country_code': 'COUNTRY',
            'question_sent': 'QS_SENT'}
```
✓ [19] < 10 ms

```
selected_df = selected_df.rename(columns = name_map)
```
✓ [20] 325ms

The “question_sent” column was also split into separate columns to obtain only the date which was wanted for analysis.

```
selected_df.isna().sum()
```
✓ [22] 1s 822ms

4 rows × 1 cols

	<unnamed>
Q_LG	0
QS_TOPIC	0
COUNTRY	0
QS_SENT	0

```
selected_df[["DATE", "TIMESTAMP"]] = selected_df["QS_SENT"].str.split(" ", expand = True)
```
✓ [23] 21s 961ms

```
selected_df
```
✓ [24] < 10 ms

1-10 of 16,767,114 rows × 6 cols

	Q_LG	QS_TOPIC	COUNTRY	QS_SENT	DATE	TIMESTAMP
2	nyn	cattle	ug	2017-11-22 12:25:08+00	2017-11-22	12:25:08+00
3	nyn	cattle	ug	2017-11-22 12:25:08+00	2017-11-22	12:25:08+00
4	nyn	cat	ug	2017-11-22 12:25:08+00	2017-11-22	12:25:08+00
5	nyn	cat	ug	2017-11-22 12:25:08+00	2017-11-22	12:25:08+00
7	swa	poultry	ke	2017-11-22 12:25:10+00	2017-11-22	12:25:10+00
8	swa	poultry	ke	2017-11-22 12:25:10+00	2017-11-22	12:25:10+00
9	eng	rabbit	ke	2017-11-22 12:25:10+00	2017-11-22	12:25:10+00
10	swa	poultry	ke	2017-11-22 12:25:12+00	2017-11-22	12:25:12+00
11	swa	poultry	ke	2017-11-22 12:25:12+00	2017-11-22	12:25:12+00
12	swa	poultry	ke	2017-11-22 12:25:12+00	2017-11-22	12:25:12+00

The “TIMESTAMP” column was dropped, and the “DATE” column was expanded to include the new “Month” and “Year” columns.

```
selected_df = selected_df.drop("TIMESTAMP", axis = 1)
✓ [25] 792ms
```

```
selected_df[["YEAR","MONTH","DAY"]] = selected_df["DATE"].str.split("-",expand = True)
✓ [27] 22s 874ms
```

```
selected_df["DATE"] = pd.to_datetime(selected_df["DATE"])
✓ [31] 2s 134ms
```

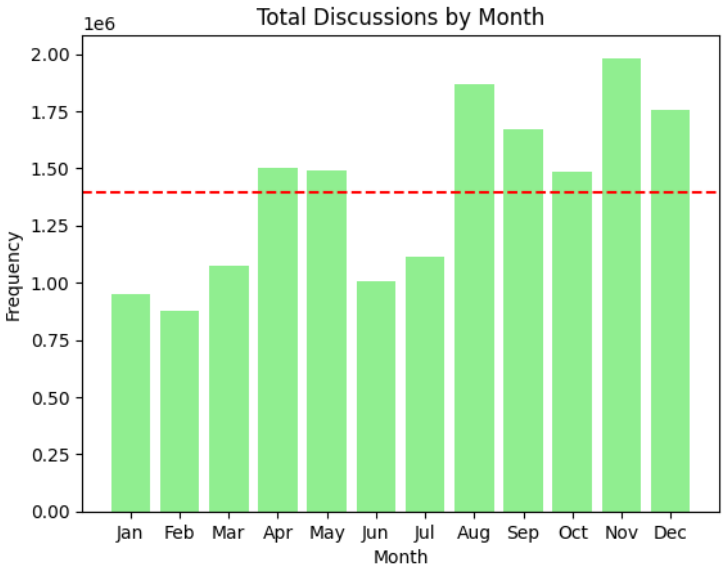
```
selected_df["Month"] = selected_df["DATE"].dt.strftime("%b")
✓ [36] 59s 206ms
```

The dataframe was then ready for some exploratory data analysis.

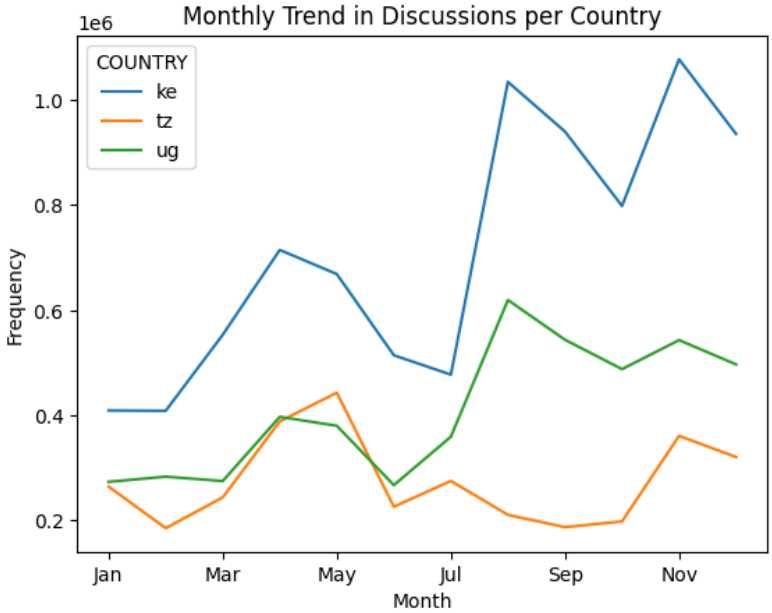
	Q_LG	QS_TOPIC	COUNTRY	QS_SENT	DATE	YEAR	MONTH	DAY	Month
2	nyn	cattle	ug	2017-11-22 12:25:08+00	2017-11-22	2017	11	22	Nov
3	nyn	cattle	ug	2017-11-22 12:25:08+00	2017-11-22	2017	11	22	Nov
4	nyn	cat	ug	2017-11-22 12:25:08+00	2017-11-22	2017	11	22	Nov
5	nyn	cat	ug	2017-11-22 12:25:08+00	2017-11-22	2017	11	22	Nov
7	swa	poultry	ke	2017-11-22 12:25:10+00	2017-11-22	2017	11	22	Nov
8	swa	poultry	ke	2017-11-22 12:25:10+00	2017-11-22	2017	11	22	Nov
9	eng	rabbit	ke	2017-11-22 12:25:10+00	2017-11-22	2017	11	22	Nov
10	swa	poultry	ke	2017-11-22 12:25:12+00	2017-11-22	2017	11	22	Nov
11	swa	poultry	ke	2017-11-22 12:25:12+00	2017-11-22	2017	11	22	Nov
12	swa	poultry	ke	2017-11-22 12:25:12+00	2017-11-22	2017	11	22	Nov

Preliminary Results

An overall graph for total discussions per month was plotted. An average value of the total counts was calculated and is denoted by a red, dashed line. The months April, May, August, September, October, November and December all recorded counts that were *above average*.



A line plot where the total counts for each country versus months was visualised. The plot did seem to show some variation in between the countries charactersied by the rise and fall of counts for certain months. Great Britain was omiited from the analysis as it had very few total counts in comparison to these three countries, so only Kenya, Uganda and Tanzania were included.



Kenya showed the highest frequency trend, followed by Uganda and Tanzania. It peaked in discussions for the months April, August and November while the months with low counts were January, February, June and July. Uganda showed a similar trend to Kenya, with prominent peaks at April, August and November. The months with low counts were January, February, March and June. In contrast, Tanzania showed a dip in counts from January to February, and a slow rise in counts with a sharp peak for the month of May. The months from June to October recorded rather low counts. There was another substantial peak in November and a dip again in December.

```
year_counts = selected_df["YEAR"].value_counts()
year_counts = year_counts.sort_index()
year_counts = year_counts.head(5)
year_counts
```

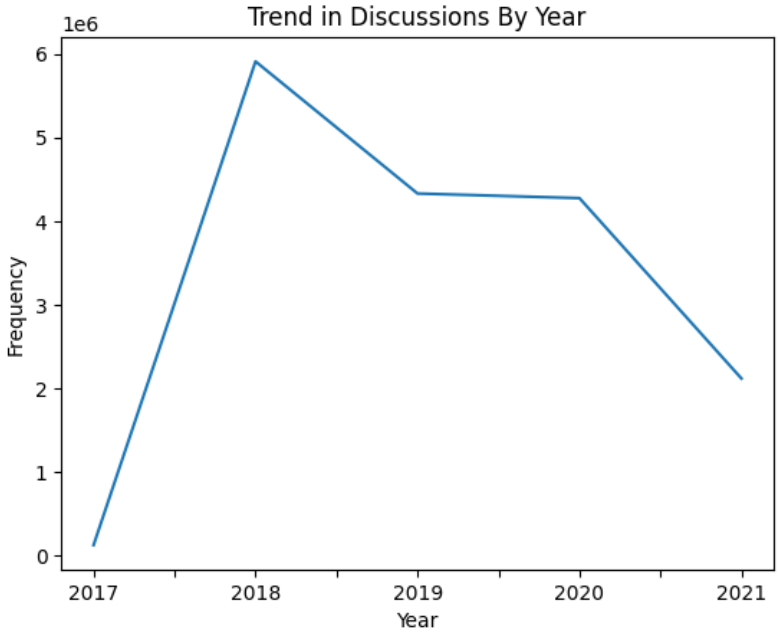
✓ [59] 785ms

5 rows × 5 rows × 1 cols

YEAR	count
2017	121760
2018	5910405
2019	4329753
2020	4274293
2021	2115769

The overall counts for each year was counted for all the countries and visualised.

There was a sharp rise in counts for the years 2017 to 2018, with 2018 showing the highest frequency of discussions recorded. This was marked by a fall in 2019. For the years 2019 to 2020, there was not much of a difference in counts, but in 2021 there was prominent decrease in the frequency of discussions. 2022 was omitted as there were very few discussion counts overall.



```
year_country = selected_df.groupby("YEAR")["COUNTRY"].value_counts()
year_country = year_country.reset_index()
year_country = year_country.head(20)
```

[61] 1s 804ms

```
year_country
```

[62] < 10 ms

1-10 20 rows × 3 cols

	YEAR	COUNTRY	count
0	2017	ke	77205
1	2017	ug	44549
2	2017	gb	4
3	2017	tz	2
4	2018	ke	4198314
5	2018	ug	1710580
6	2018	tz	1470
7	2018	gb	41
8	2019	ke	2251270
9	2019	ug	1307880

```
year_country_pivot = year_country.pivot(index = "YEAR", columns = "COUNTRY", values = "count")
year_country_pivot
```

[63] < 10 ms

5 rows 5 rows × 4 cols

COUNTRY	gb	ke	tz	ug
2017	4	77205	2	44549
2018	41	4198314	1470	1710580
2019	62	2251270	770541	1307880
2020	57	1489098	1451480	1333658
2021	7	503329	1081128	531305

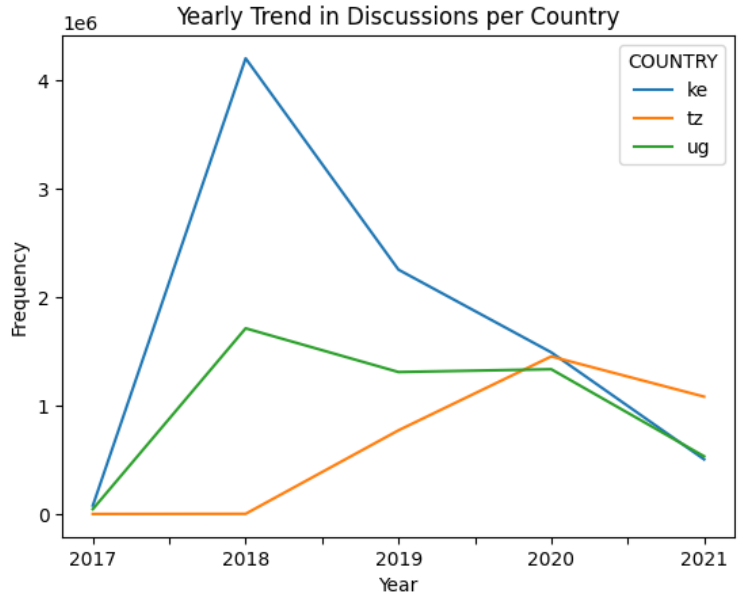
```
year_country_pivot = year_country_pivot[["ke","tz","ug"]]
year_country_pivot
```

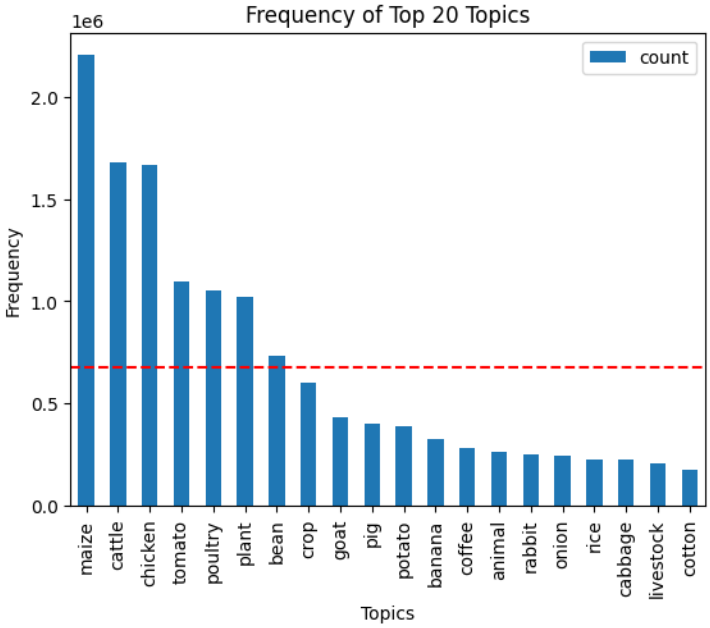
[64] < 10 ms

5 rows 5 rows × 3 cols

COUNTRY	ke	tz	ug
2017	77205	2	44549
2018	4198314	1470	1710580
2019	2251270	770541	1307880
2020	1489098	1451480	1333658
2021	503329	1081128	531305

For a closer look, the trends for discussions on a yearly basis were compared amongst the three countries. For both Kenya and Uganda, there were strong peaks in 2018, while Tanzania only showed a strong peak in 2020. Kenya has showed a steady decline in frequency since then, while Uganda showed a small peak in 2020, and a decline in 2021. Tanzania showed higher frequency in discussions compared to both Uganda and Kenya in 2021.





The overall frequency of topics was visualised. The top three topics were *maize*, *cattle* and *chicken* respectively. Tomato, poultry, plant and bean were the other topics with above average frequency among the farmers.

The top topics that were found among all three countries were selected to further observe any variations that occurred in between the countries.

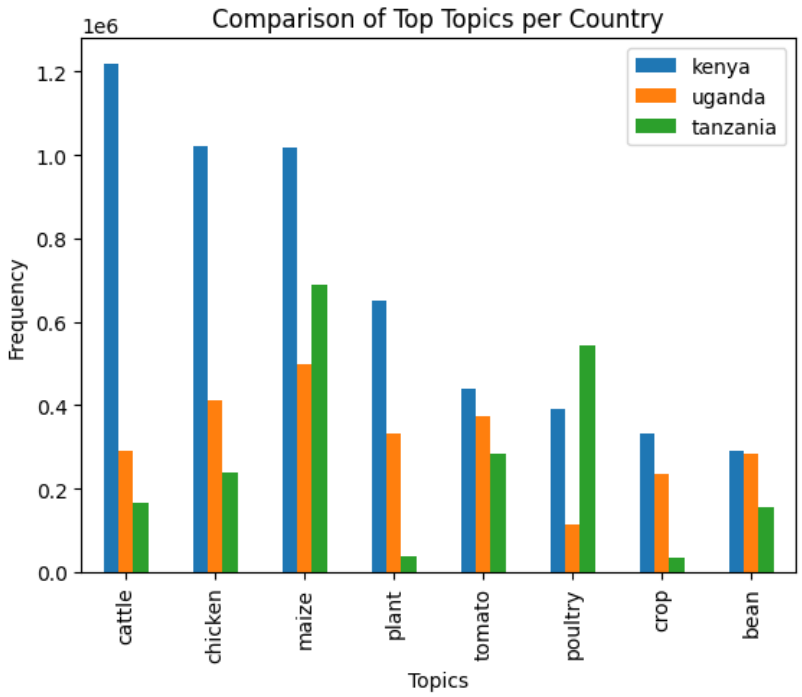
```
new_topic_plus = new_topic_plus.rename(columns = {"count_x" : "kenya", "count_y": "uganda", "count":"tanzania"})
```
✓ [87] < 10 ms

```
new_topic_plus_8 = new_topic_plus.head(8)
new_topic_plus_8
```
✓ [88] < 10 ms

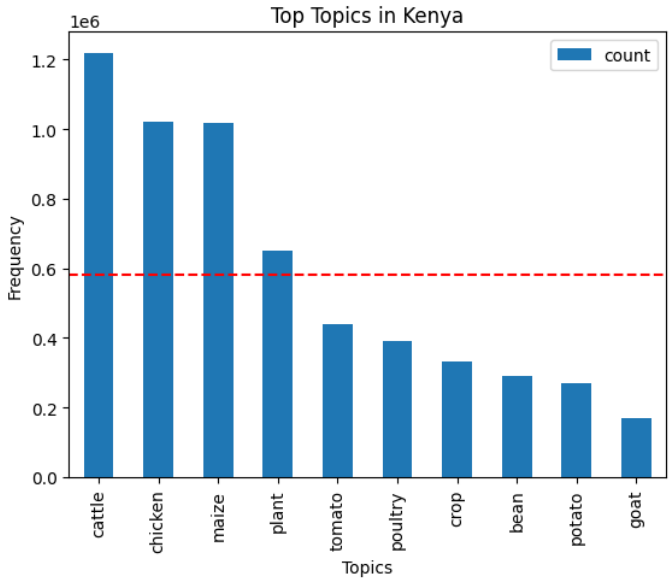
8 rows × 7 cols

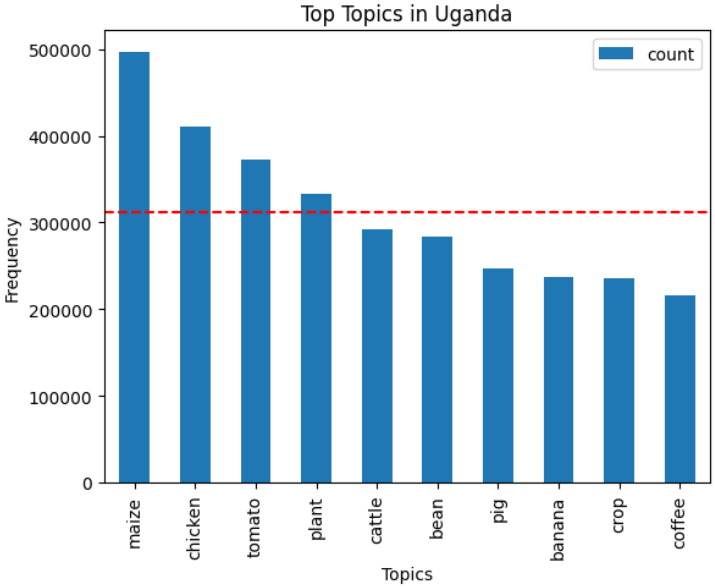
	COUNTRY_x	QS_TOPIC	kenya	COUNTRY_y	uganda	COUNTRY	tanzania
0	ke	cattle	1218676	ug	291854.0	tz	164936.0
1	ke	chicken	1019592	ug	410347.0	tz	237295.0
2	ke	maize	1016368	ug	496842.0	tz	688528.0
3	ke	plant	652196	ug	333585.0	tz	38686.0
4	ke	tomato	439970	ug	372093.0	tz	284292.0
5	ke	poultry	392805	ug	114764.0	tz	544228.0
6	ke	crop	333066	ug	235157.0	tz	34690.0
7	ke	bean	291827	ug	283567.0	tz	156713.0

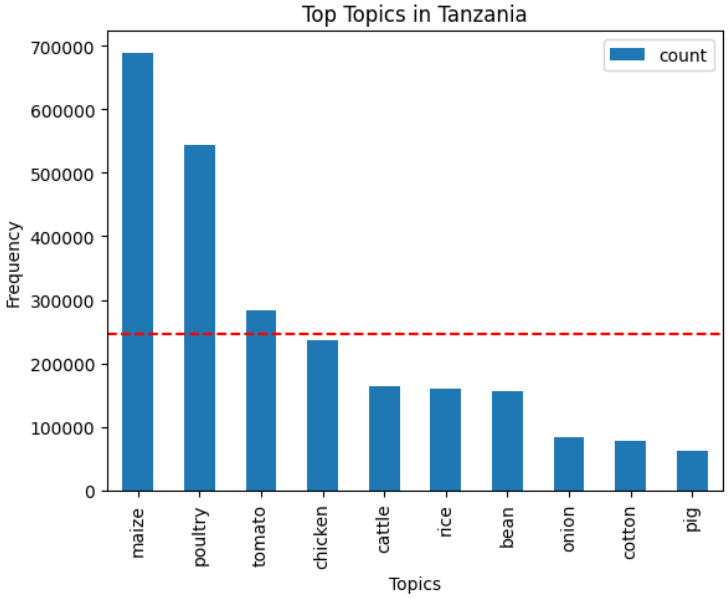
A plot was visualised to show this comparison. For Kenya, the top topic discussed was cattle, followed by chicken and maize. For Uganda, the top topics were maize, chicken and tomato respectively. For Tanzania, the top topic discussed was maize, followed by the general group of poultry, and tomato. Poultry could include chickens, turkey and any other bird related discussions which could be further analyzed through natual language processing algorithms to identify what was their particular interest. Cattle is a more popular topic in Kenya than in Uganda and Tanzania. Maize was also talked about more in Tanzania than in Uganda. Tomato is also very popular among all three countries.



The following are topic versus frequency plots for each country. These numbers were visualised to have a closer look at the topics that were frequently featured in farmer discussions within each country to observe any variations or patterns.

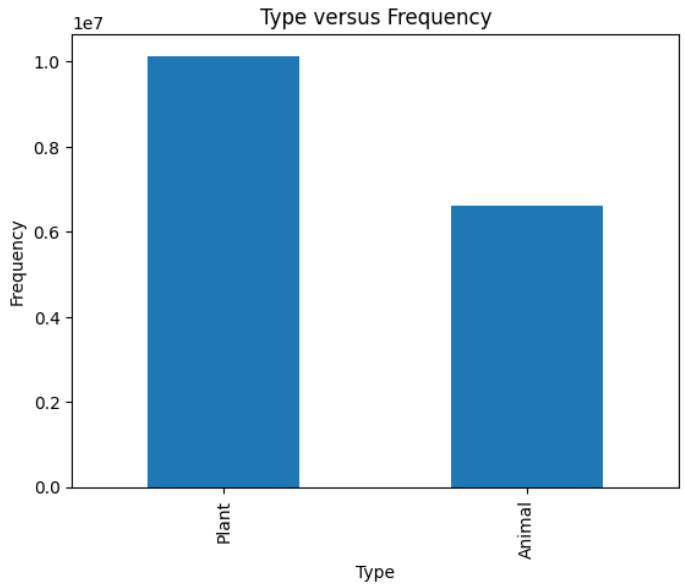




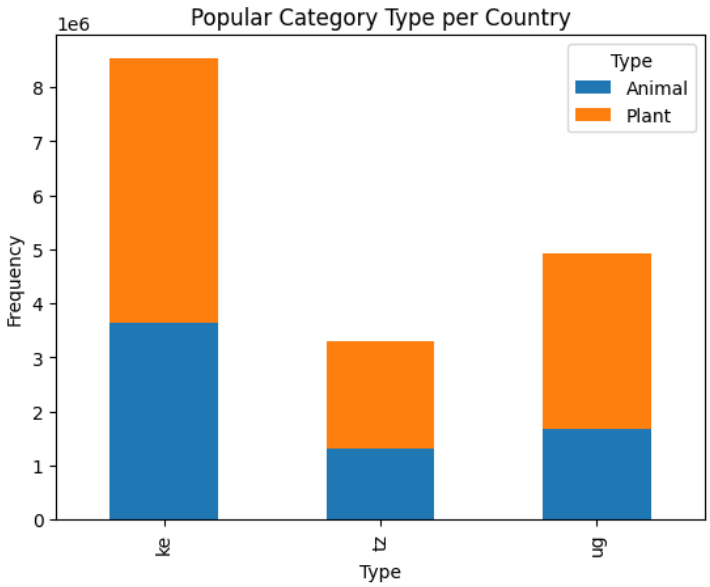


In Kenya, the farmers talked mainly about cattle and chicken. The non-animal related topics talked about here were maize and other ambiguous “plant” topics. Maize was the most talked about topic in the plants category among all three countries. “Goat” and “Potato” are some notable mentions as they are not featured in the top list for the other two countries. Tomato is not as frequently talked about in Kenya like in Uganda and Tanzania. In Uganda, cattle was not a top topic, but maize was, followed by chicken, tomato and “plant”. Here, other topics that were also featured were banana and coffee. In Tanzania, maize was the top topic, followed by “poultry” and tomato. Pigs were more popular in Uganda, but not as much in Tanzania and barely in Kenya, where it did not make it to the top 10 list. Certains crops in Tanzania that were only featured in their top list were rice, onion and cotton.

There were a lot more “Plant” topics discussed compared to “Animal” topics.



Subsequently, there were a lot more Plant topics discussed than Animal topics within each country as well. However, from a general overview, there was a heavier incline towards plant topics in Uganda and Kenya compared to Tanzania.



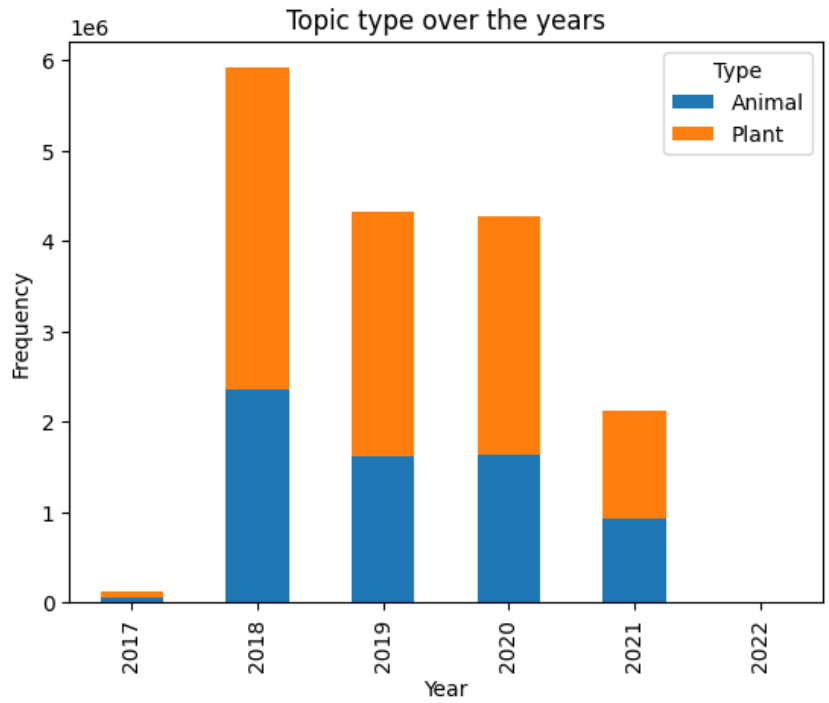
```
year_type_country_pivot = year_type_country.pivot_table(values = "count",index = "YEAR", columns = "Type")
✓ [112] < 10 ms

year_type_country_pivot
✓ [113] < 10 ms
```

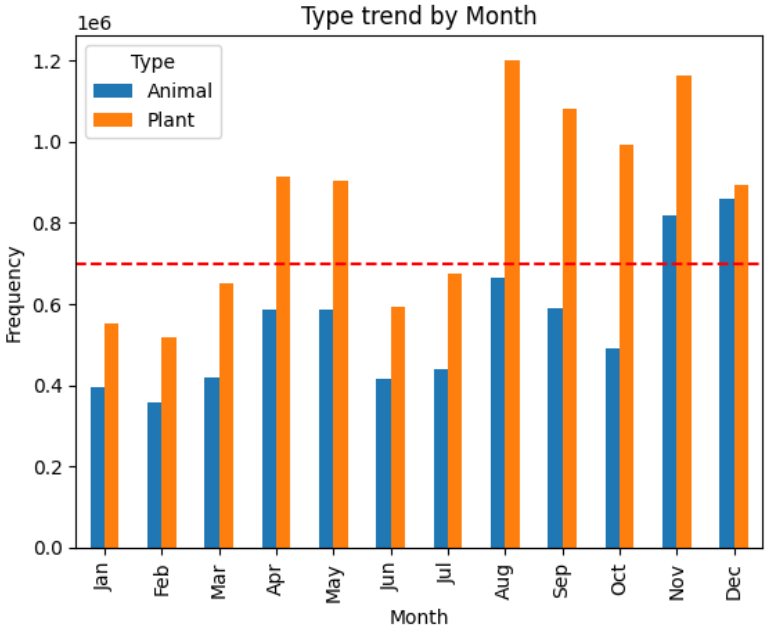
6 rows 6 rows × 2 cols

Type	Animal	Plant
2017	56064.0	65696.0
2018	2364011.0	3546394.0
2019	1625620.0	2704133.0
2020	1636462.0	2637831.0
2021	936565.0	1179204.0
2022	8020.0	7114.0

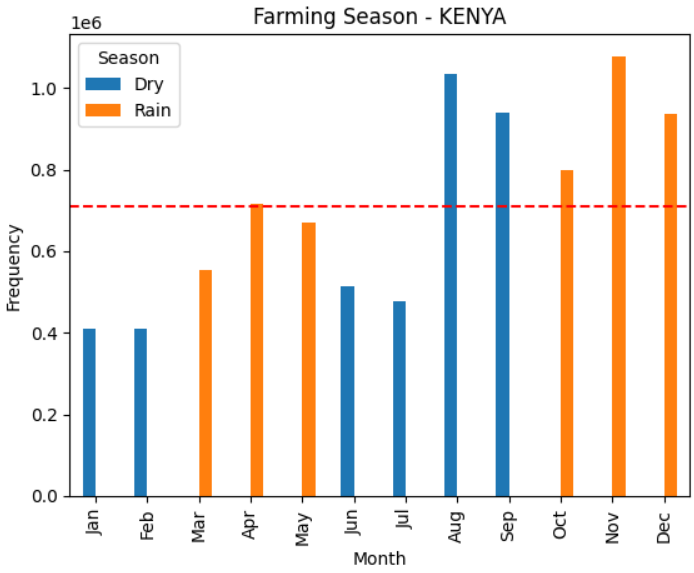
Over the years, there are more Plant based topics being discussed until 2022. In 2021, the difference between Animal and Plant topics being discussed reduces, and is of similar frequency. In 2022, the overall counts are low. More data could be collected to show if there is any different trend taking place.



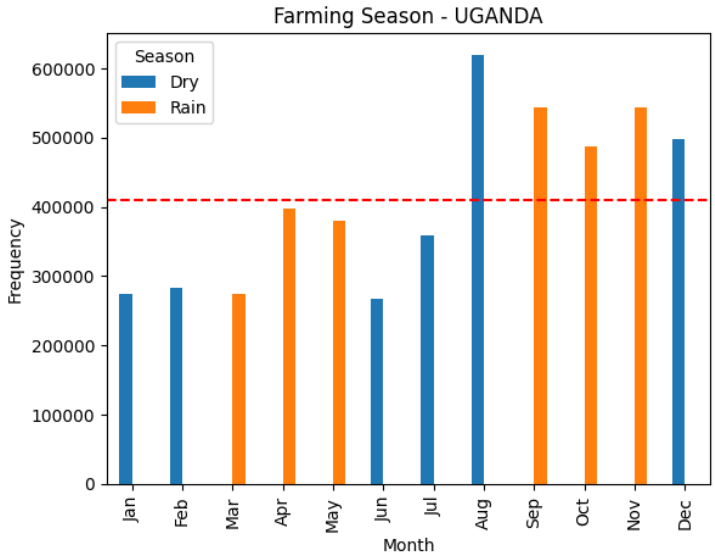
For the months April to May, there is a strong rise in topics, especially for Plant based topics. This is continued for the months August to December.



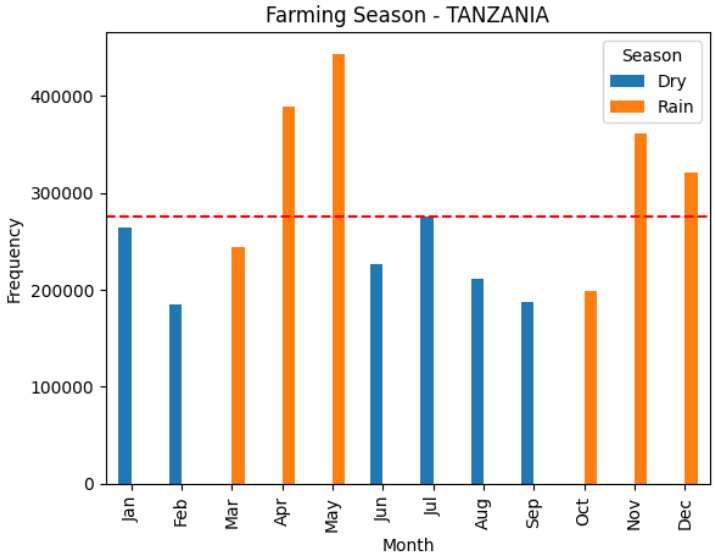
The rainy season months for each country and any relation between the frequency of discussions and those seasonal months were explored. As it was mentioned in online literature that the three countries generally relied on traditional rain-based farming, the months which conincided with the rainy seasons were deduced to be the farming months, where planting or growing would be the main activities. The dry months would typically be harvesting months, with some months like August and September being very popular as it could denote the month right before the Vulli rains. This could mean that people were trying to prepare for the farming season, and perhaps planting early to maximise on the rains.



In Kenya, among the 'Long rain' months, April was popular for discussions. In the second 'Shorter rain' season, November recorded the highest number of discussions overall, with October and December being high in frequency. August and September were popular in discussions during the dry months.



For Uganda, the 'Short' rainy season started early in September and ended in November. August was the most popular month for discussions. December still ranked high in discussion for a dry month. Further analysis through NLP methods could provide more insights. Some suggestions could be that pests, predation in animal farming or farmers prepping for rain-resistant crops during the harvest months could possibly contribute to the high count in discussions during the non-rainy months.



In Tanzania, the 'Long rain' seasons were typically March to May, and the 'Short rain' seasons were from October to December. There was a general rise in discussions for those rainy months. April and May were the most popular months. This was followed in the next rainy season with November and December. Unlike in Kenya and Uganda, the most popular dry/harvest month in Tanzania was July.

During the dry months, there was no significant rise in discussion (i.e. the dry months did not rise above the average discussion count). This could mean that Tanzania was striking a balance with a blend of animal and plant based farming to compensate for the shifting climate. In contrast, in Kenya and Uganda, the compensation for the shifting climate change could be on making changes to their preparation techniques and perhaps finding ways to plant more crops in the off-rainy seasons. Further analysis could be done through NLP methods to possibly explore this hypothesis.

The table below shows the most talked about topics, ordered according to their total counts, for each month for all the three countries. This table hopefully would help in providing an overall topic to look into for further research using NLP methods in the future.

Monthly Topic Table – Most talked about topic per month for each country

Month/Country	*Kenya*	*Uganda*	*Tanzania*
January	**Chicken**, Cattle, Maize, Plant	**Chicken**, Maize, Tomato, Plant	**Maize**, Poultry, Tomato, Chicken
February	**Cattle**, Chicken, Maize, Plant	**Maize**, Chicken, Cattle, Plant	**Maize**, Poultry, Tomato, Chicken
March	**Maize**, Cattle, Chicken, Plant	**Maize**, Chicken, Tomato, Plant	**Maize**, Poultry, Tomato, Chicken
April	**Maize**, Cattle, Chicken, Plant	**Maize**, Plant, Tomato, Chicken	**Poultry**, Maize, Tomato, Chicken
May	**Cattle**, Maize, Chicken, Plant	**Chicken**, Tomato, Maize, Cattle	**Poultry**, Maize, Tomato, Chicken
June	**Cattle**, Chicken, Maize, Plant	**Tomato**, Chicken, Maize, Plant	**Poultry**, Maize, Tomato, Chicken
July	**Cattle**, Chicken, Maize, Plant	**Maize**, Tomato, Chicken, Plant	**Poultry**, Maize, Tomato, Chicken
August	**Tomato**, Cattle, Chicken, Plant	**Tomato**, Maize, Plant, Chicken	**Maize**, Poultry, Tomato, Chicken
September	**Maize**, Cattle, Chicken, Plant	**Maize**, Tomato, Chicken, Plant	**Maize**, Poultry, Tomato, Chicken
October	**Maize**, Cattle, Chicken, Plant	**Maize**, Chicken, Tomato, Plant	**Maize**, Poultry, Tomato, Chicken
November	**Chicken**, Cattle, Maize, Plant	**Maize**, Banana, Chicken, Plant	**Maize**, Poultry, Tomato, Chicken
December	**Cattle**, Chicken, Maize, Plant	**Chicken**, Cattle, Banana, Tomato	**Maize**, Poultry, Tomato, Chicken

Short – Summary of Insights

Overall:

- There are more Plant topics discussed than Animal topics
- There is a general trend followed where there are high discussion counts during the rainy season, and low counts during the dry season, except for some variations for the months July, August and September within each country.
- Maize and tomato are the most talked about crops. Cattle and chicken are most talked about in the animal category.
- 2018 was the year where most discussions occurred. It is followed by a dip and a plateau from 2019 to 2020. It reduces from 2021 onwards.
- There is a large gap between animal and plant topics discussed from 2018 to 2020, which reduces to a large extent in 2021. More data has to be collected to see if a change in trend could be taking place.

Kenya:

- Most discussions are about cattle, chicken and maize.
- However, overall they tend to have more discussions on plant topics than animal topics.
- Plant topics generally peak a month or so before the rainy seasons occur or in the first months of the start of the rains.

Uganda:

- Most discussions are about maize, chicken and tomato.
- However, overall they tend to have more discussions on plant topics than animal topics.
- Maize and tomato seem to be the favored plant topics among the farmers here.
- During the rainy seasons (except for May), maize is the most talked about crop.

Tanzania:

- Most discussions are about maize, poultry and tomato.
- Though there is a tilt to more plant topics than animal topics, the gap between the two categories is remarkably lesser than the other two countries.
- Most discussions peak during the rainy seasons. However, in April and May, where the counts were the highest, the most talked about topic was poultry. It is the most popular topic until July. From August to March, the most talked about topic is maize.
- Throughout the year, maize and poultry interchange for most talked about topic, but tomato and chicken remain consistent in their ranking throughout.