Exploratory Data Analysis

Dataset: Agriculture and rural development

This dataset contains statistics on agriculture and rural development from 1960 to 2020 of different countries and groups of countries. Columns are based on country name, year along with different areas of rural development and agricultural development.

Initial Analysis Questions

1. Visualize electricity access development in the rural area of Bangladesh over the last three decades.
2. Situation of Rural population growth in Bangladesh and the World?
3. Agricultural raw material export in different income categorical countries.
4. Agricultural raw material import in different income categorical countries.
5. Greenhouse issues by agricultural development getting higher or lower?

Questions found while exploring:

6. Which country is more focused on agricultural development (Judge by employment growth)
7. Which sub continental region has the highest rural population growth.
8. How is agricultural raw material export and import in bangladesh?
9. Which continent has the highest rate of production and which continent has the least?
10. How fertilizer Consumption relates to crop production?

Discoveries & Insights

Our analysis starts with basic details. Later plots of individual variables to assess distributions and data quality. As we progress, we build up multi-dimensional views for our analysis questions.

Overview of the shape & structure of your dataset

Shape of Dataset: Dataset contains 16163 rows and 45 columns.
Structure of Dataset: Record data category.
Variables used:

- Categorical column: 2 (as object data type)
- Numerical Column: 43 (Data type for ‘Year’ column is int64 and other numerical columns contain float64 datatype)

How are they distributed?

We don't necessarily need to see categorical values here. So, I am skipping this and focusing on the distributions on numerical variables (except year). Used histogram to understand the distribution. Or in the other words the frequency of data. So, what I have found from this is that almost all the data has skewed distribution (mostly right skewed and few left skewed), with few fairly normal and a bimodal distribution.

Data quality issues

1. Data type issues with float64 and int64 which uses more memory and are actually not required here. Rural population column should be integer value rather than float.
2. Column name with a common string that unnecessarily makes it larger.
3. Another thing to notice is the "Country Name" column. While observing the column I found that not all names are country names. There are groups of countries too as country names. Such as continents, economically categorized country groups, sub regional groups, country organizations and also the whole world itself. This data certainly impacts on the distribution at a high level in some columns.
4. When we talk abou data quality issues the most common issue we find is null values. More null values, more issues. In this dataset there are almost 281425 Null values which is almost 38.7% of the whole dataset. So it is a big issue to deal with.

Manipulation

- Rename column names by just removing "average_value_" from each column name.
- Data Types bit adjustment (float64 to float32, 'Year' column int64 to int16, categorical columns set to category data type)
- Adding 'Country Type' columns because of different categorical country values.
- Sort data by Country Type, Country Name and Year
- Shortening dataset according to needs. (Targeting questions above)
    1. Removing unnecessary rows (Such as individual country rows except Bangladesh)
    2. Omitting unnecessary columns
- Dealing with Null Values:
    1. Removing columns that have over 80% null values.
    2. Linear interpolation in both directions to fill the nulls.

Manipulated dataset has 29 columns, 854 rows.

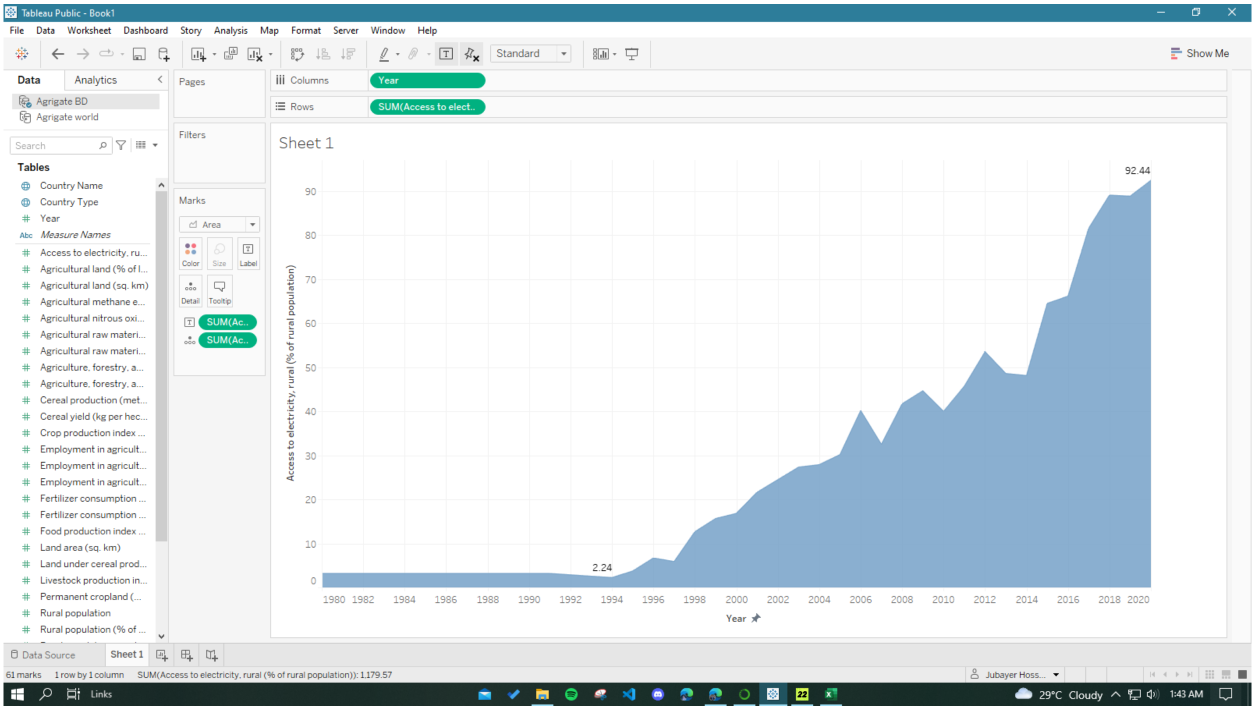
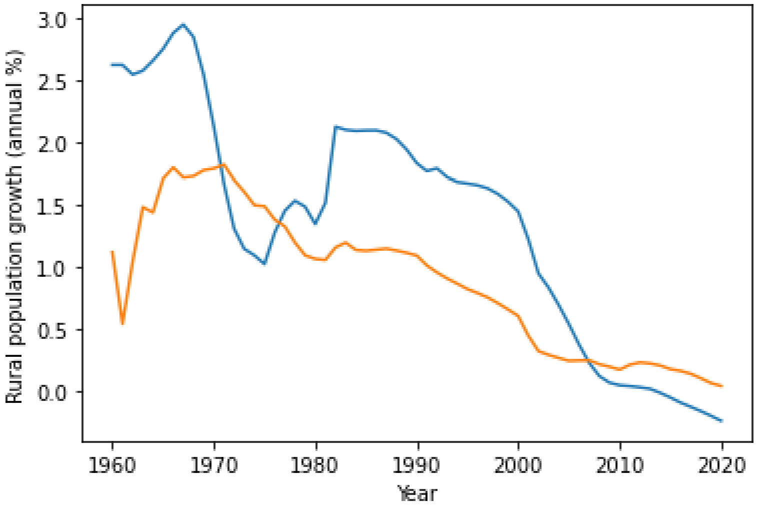


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
Agrigate BD
Agrigate world
Search
Tables
Country Name
Country Type
Year
Measure Names
Access to electricity, ru...
Agricultural land (% of l...
Agricultural land (sq. km)
Agricultural methane e...
Agricultural nitrous oxi...
Agricultural raw materi...
Agricultural raw materi...
Agriculture, forestry, a...
Agriculture, forestry, a...
Cereal production (met...
Cereal yield (kg per hec...
Crop production index ...
Employment in agricult...
Employment in agricult...
Employment in agricult...
Fertilizer consumption ...
Fertilizer consumption ...
Food production index ...
Land area (sq. km)
Land under cereal prod...
Livestock production in...
Permanent cropland (...
Rural population
Rural population (% of ...
Pages
Filters
Marks
Area
Color
Size
Label
Detail
Tooltip
SUM(Ac..
SUM(Ac..
Columns
Year
Rows
SUM(Access to elect..
Sheet 1
92.44
2.24
Access to electricity, rural (% of rural population)
0
10
20
30
40
50
60
70
80
90
1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Year
Data Source
Sheet 1
61 marks 1 row by 1 column SUM(Access to electricity, rural (% of rural population)): 1,179.57
Jubayer Hoss...
Links
29°C Cloudy
1:43 AM

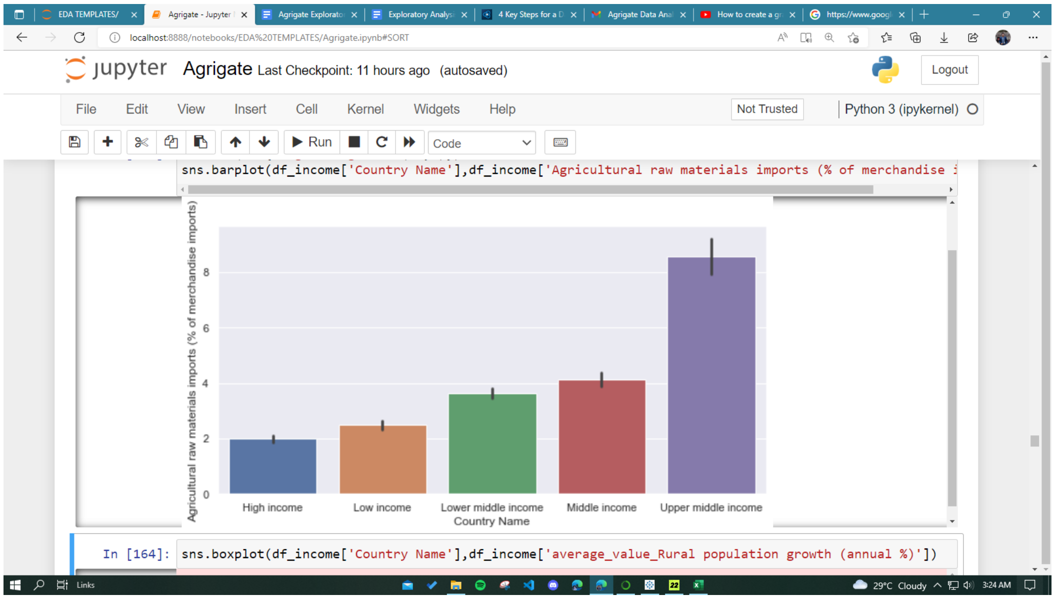
Electricity access in rural Bangladesh: This scatter plot shows the growth of electricity access in the rural Bangladesh area. Electricity access wasn't available much before 1990. In the last 3 decades availability of electricity in the rural area of Bangladesh had a linear and drastic improvement. The rural population getting access to electricity has changed.From 2.24% in 1994 to 92.44% in 2020 of the rural population getting the access.

```
In [412]: sns.lineplot(df_bd['Year'],df_bd['Rural population growth (annual
          sns.lineplot(df_world['Year'],df_world['Rural population growth (a

Out[412]: <AxesSubplot:xlabel='Year', ylabel='Rural population growth (annua
```

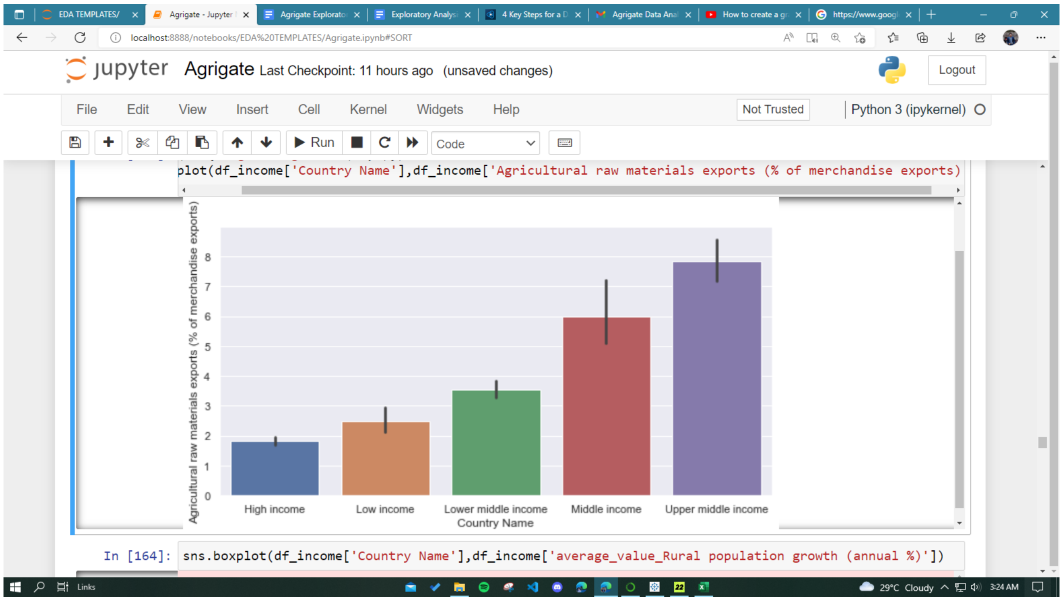


Rural Population Growth Downwards: This line graph is clearly showing us that the growth of rural population is going down. The orange line is for Bangladesh and the other one is for the whole world. How population growth in rural areas has changed over time. We can assume that people from rural areas mostly came to urban and central areas. Urbanization may have caused this graph to go down.



```
sns.barplot(df_income['Country Name'],df_income['Agricultural raw materials imports (% of merchandise i
```

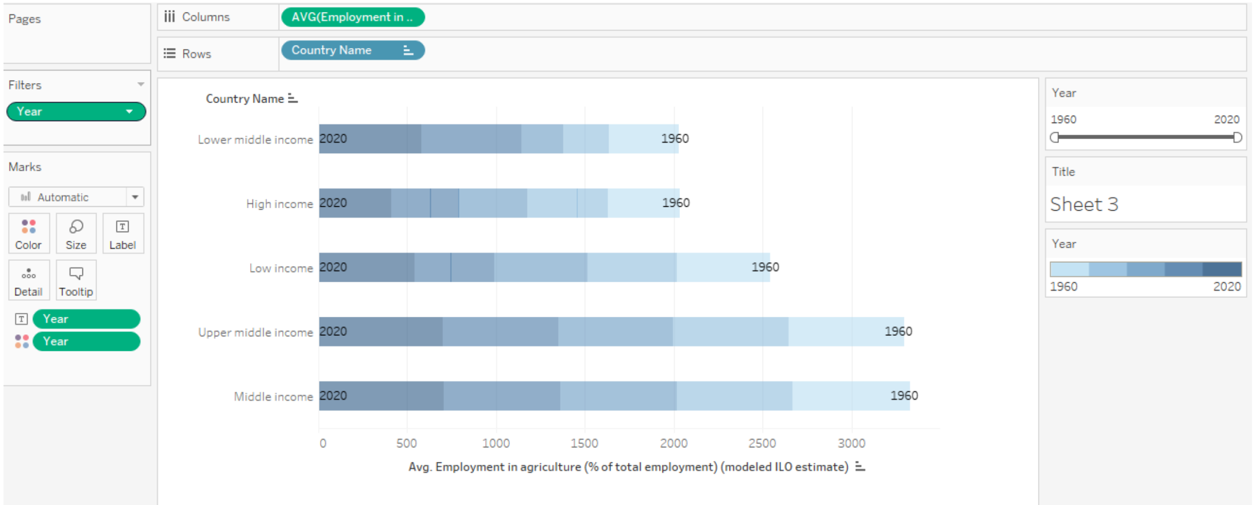
```
In [164]: sns.boxplot(df_income['Country Name'],df_income['average_value_Rural population growth (annual %)'])
```



```
plot(df_income['Country Name'],df_income['Agricultural raw materials exports (% of merchandise exports)
```

```
In [164]: sns.boxplot(df_income['Country Name'],df_income['average_value_Rural population growth (annual %)'])
```

Agricultural raw materials export and import (Income based country type): There are two barcharts on agricultural raw material export and import which is on income based country name. In the last 60 years. Upper middle income class countries have the most import and export and high income class countries are the lowest. This also gives an insight that upper middle and middle class countries are mostly agriculture centric which is opposite in high income countries.



Columns: AVG(Employment in ..

Rows: Country Name

Filters: Year

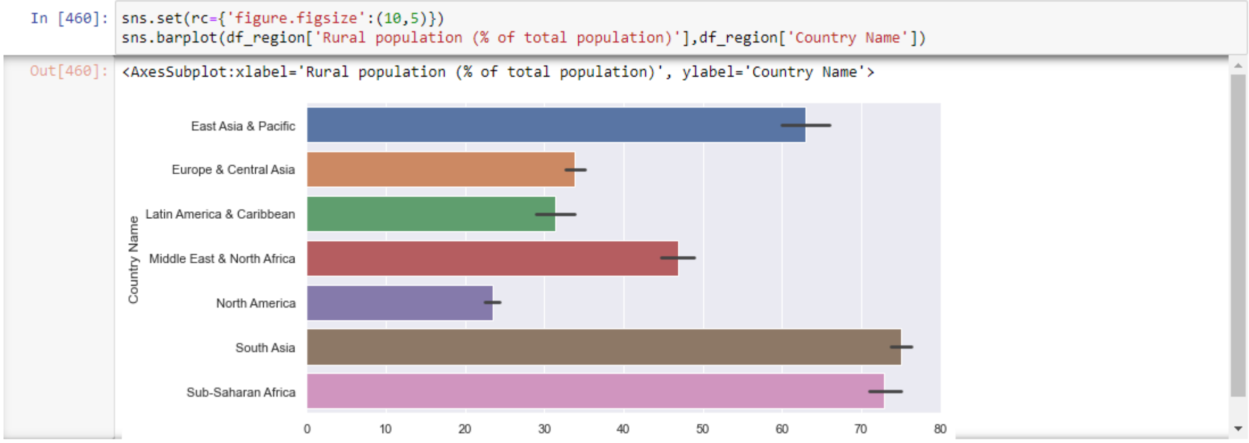
Marks: Automatic

Year

Year

Title: Sheet 3

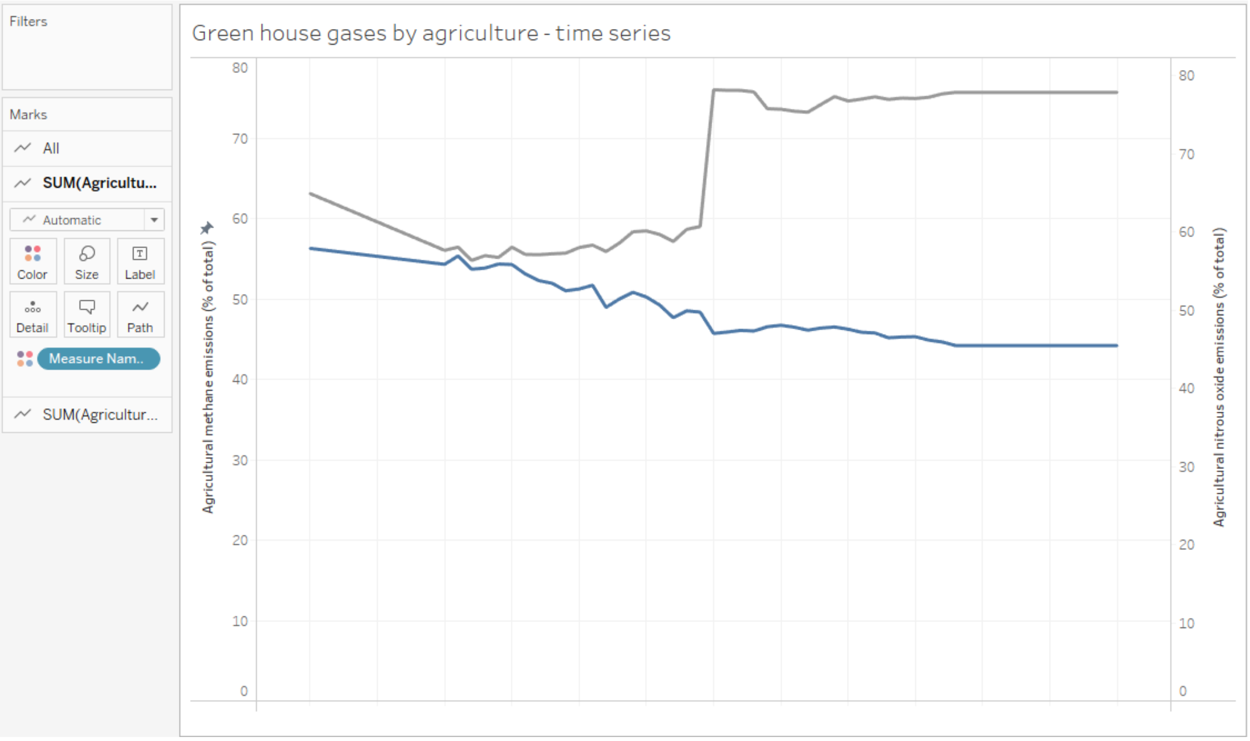
Which income category country is more focused in agriculture: This horizontal bar chart shows the sum of employment in agriculture which is the percentage of total employment in different income categorical countries. Middle income country category is highest in agricultural centric employment. Upper middle income countries are very close to it. This again gives us the idea of which income based countries are mostly focused in agriculture.



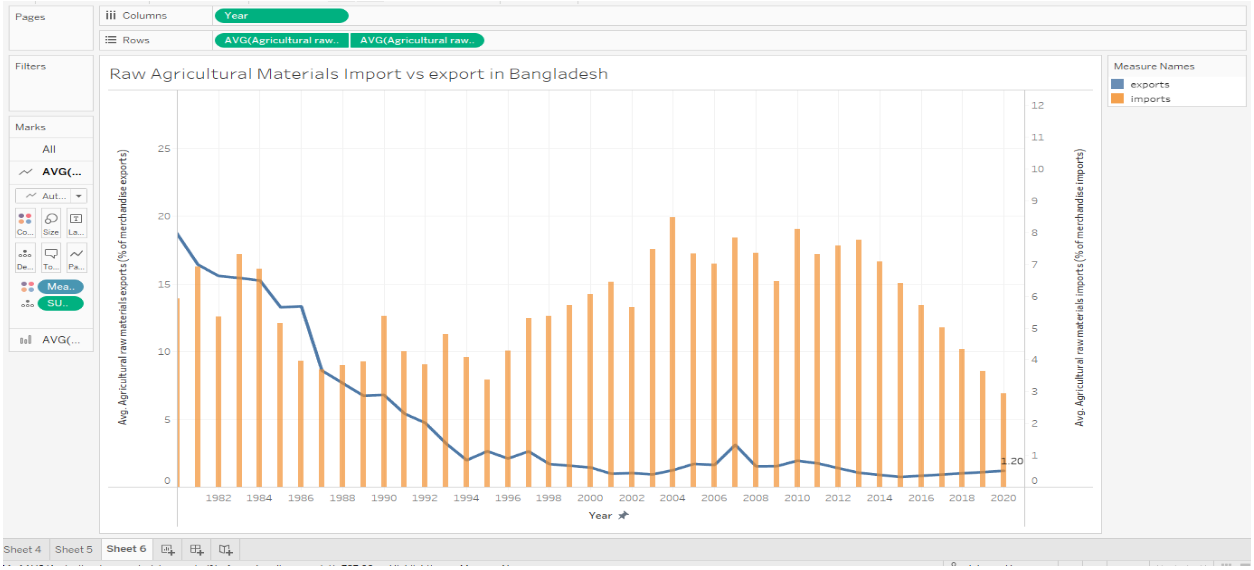
```
In [460]: sns.set(rc={'figure.figsize':(10,5)})
          sns.barplot(df_region['Rural population (% of total population)'],df_region['Country Name'])

Out[460]: <AxesSubplot:xlabel='Rural population (% of total population)', ylabel='Country Name'>
```

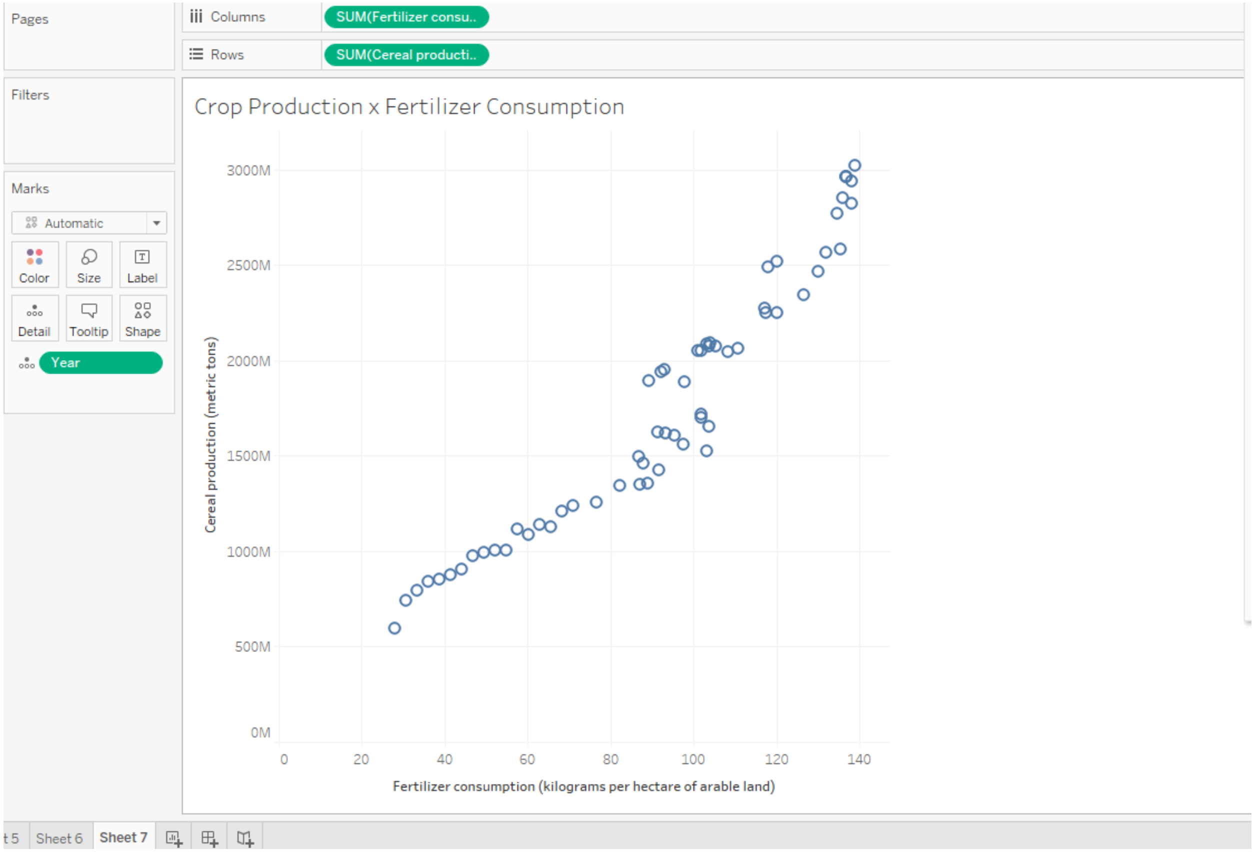
Which region has the highest rural population growth: This bar chart describes sub continental regions with 60 years of rural population growth by the percentage of total population. South Asia region has the highest growth of population in rural areas in the total population perspective followed by Sub-saharan Africa region. And East Asia Pacific is close to it. If we merge the graphs above with it, it may show us which income based countries in which region have a great impact on agriculture in their economy.


Filters
Marks
All
SUM(Agricultu...
Automatic
Color
Size
Label
Detail
Tooltip
Path
Measure Nam..
SUM(Agricultur...
Green house gases by agriculture - time series
Agricultural methane emissions (% of total)
Agricultural nitrous oxide emissions (% of total)
80
70
60
50
40
30
20
10
0

Greenhouse gas world effect by agriculture: This is a line graph, two lines showing how two different greenhouse gasses are generated by agricultural development over time (1960 to 2020). Gray line is for methane gas and the other one represents nitrous oxide gas. Methane gas is going down which is a good sign but nitrous oxide is rising high. Global warming risk developed from agricultural development.



Agricultural Material Export vs Import in Bangladesh: This is a bar and line graph that represents agricultural raw material export and import insight on the last 40 years in Bangladesh. Orange bar represents importing over time and the blue line shows exporting over time. It is clear enough to judge that exporting is gradually getting down and importing is mostly above it.



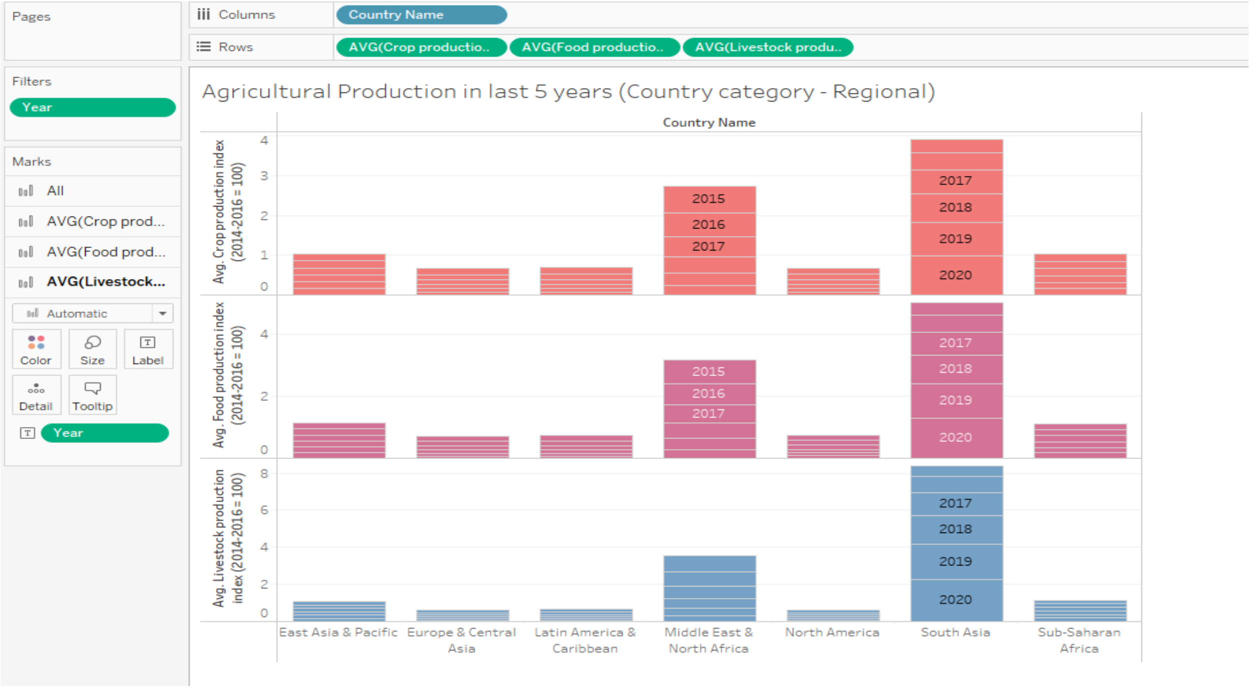
Columns: SUM(Fertilizer consu..

Rows: SUM(Cereal producti..

Marks: Automatic, Year

Pages
Filters
Marks
Color
Size
Label
Detail
Tooltip
Shape
Crop Production x Fertilizer Consumption
3000M
2500M
2000M
1500M
1000M
500M
0M
Cereal production (metric tons)
0
20
40
60
80
100
120
140
Fertilizer consumption (kilograms per hectare of arable land)
t 5
Sheet 6
Sheet 7

Scatter Plot on Fertilizer Consumption x Crop Production index: There is an interesting scatter plot showing how fertilizer consumption and crop production index is positively correlated. The more fertilizer is consumed, the crop production index went higher with it.



Columns: Country Name

Rows: AVG(Crop productio.. AVG(Food productio.. AVG(Livestock produ..

Filters: Year

Marks: All, AVG(Crop prod..., AVG(Food prod..., AVG(Livestock..., Automatic, Year

Pages
Filters
Marks
Color
Size
Label
Detail
Tooltip
Agricultural Production in last 5 years (Country category - Regional)
Country Name
Avg. Crop production index (2014-2016 = 100)
4
3
2
1
0
2015
2016
2017
2017
2018
2019
2020
Avg. Food production index (2014-2016 = 100)
4
2
0
2015
2016
2017
2017
2018
2019
2020
Avg. Livestock production index (2014-2016 = 100)
8
6
4
2
0
2017
2018
2019
2020
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa

Agricultural Production in the last 5 years (Country category - Regional): 3 different bar charts showing agricultural productions in different sub continental regions. Where south Asia is a clear winner in agricultural food production. (crop or livestock). Middle east & north africa following it.

Summary

Take away from this document: we can get insights on how rural population growth changed over time, agricultural raw materials export-import scenario in different categories, food production insight with interesting correlation with fertilizer consumption data, greenhouse gas from agriculture etc. This has Visual understanding with captions describing it.