Data Analysis and Visualization Project

GLOBAL TRADE ANALYSIS USING TABLEAU

By Shreyash Mehta, Anjali Dembla, Namrata Rath

Data source: https://www.kaggle.com/datasets/unitednations/global-commodity-trade-statistics

The dataset contains 10 columns and 8.225.871 rows, but we have sampled the data for 7 specific countries i.e., India, USA, UK, Canada, South Africa, Argentina, and Australia which counts to a total of 87570 rows

Objective:

The objective is to create an interactive dashboard in Tableau to study the potential trade commodities in a country, year, trade value, and quantity.

Variables used in the dataset:

Column names: Country, year, comm_code, commodity, flow, trade_usd, weight(kg), quantity, category, quantity_name

Description of variables:

- Country_or_area: Country name of record
- Year: Year in which the trade has taken place
- Comm_code: The uniharmonized coding system generally referred
- Commodity: Description of a particular commodity code
- Flow: Flow of trade i.e. export, import, others
- Trade_usd: The value of the trade provided in USD
- Weight_kg: Weight of the commodity in kilograms
- Quantity_name: Description of the quantity measurement type given the type of item (i.e. number of items, weight in, etc.)
- Quantity: Count of the quantity of a given item based on the quantity name
- Category: Category to identify commodity

The major countries in focus for global trade and logistics here are USA, Canada, India and UK.

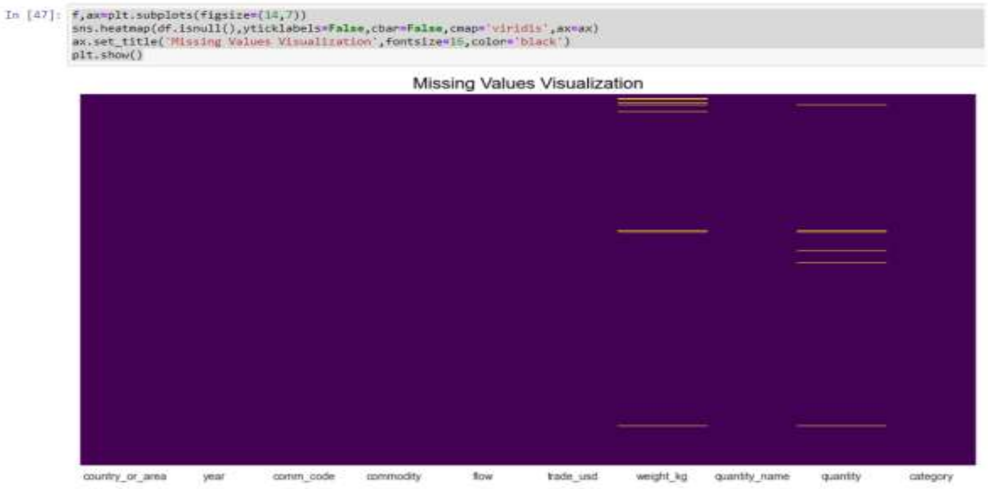
The dataset contains missing values and cleaning the dataset using Python programming language.

A line chart, Bar chart, Heat Map, Scatterplot, Bump chart etc. can be used to depict the analysis of data.

Some parameters of analysis are mentioned below:

- Top categories based on quantity
- Trade Market Analysis - Flow of the trade – Import, Export
- Demand for the commodity over the years
- Market size based on quantity
- Top commodities based on quantity

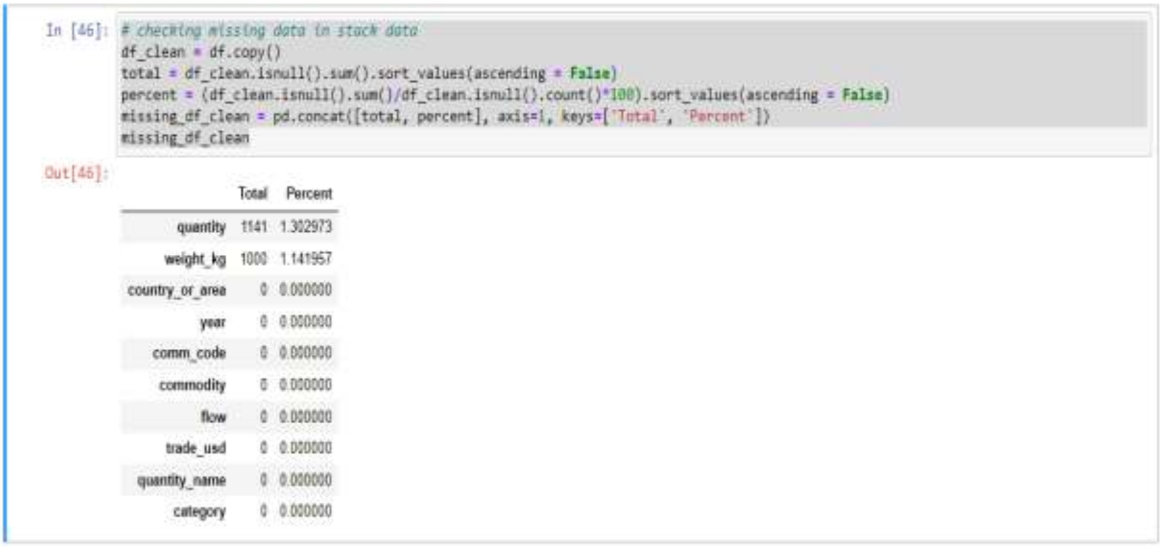
Exploratory Data Analysis:



```
In [47]: f,ax=plt.subplots(figsize=(14,7))
         sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis',ax=ax)
         ax.set_title('Missing Values Visualization',fontsize=15,color='black')
         plt.show()
```

Fig 1.

The yellow horizontal lines in a column means that there are some missing values in that column. So, the two columns, quantity and weight_kg show some missing values After some exploration, we checked that missing values only correspond to a certain type of categories where we assume data is more difficult to obtain, retain or update

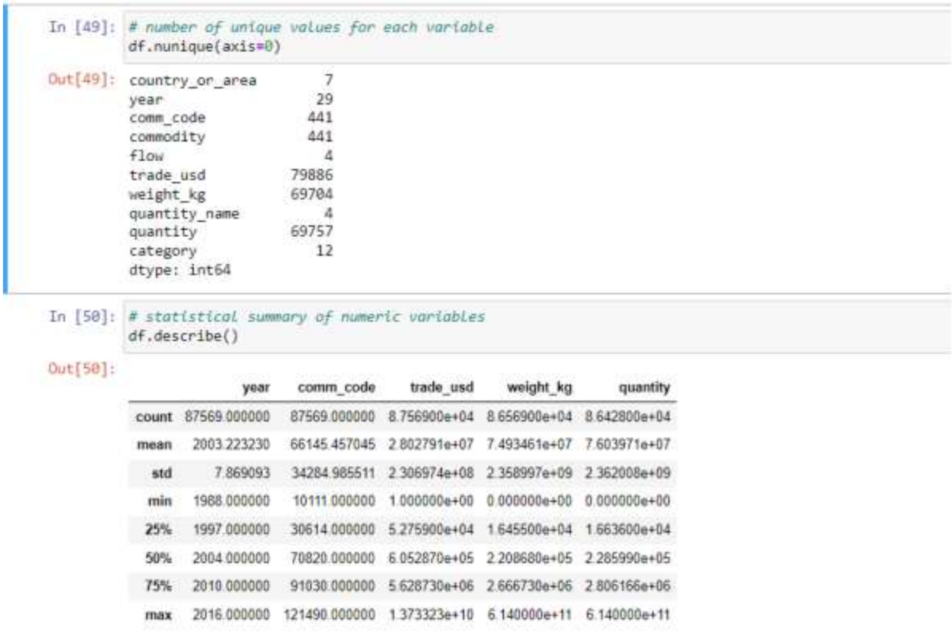


```
In [46]: # checking missing data in stack data
         df_clean = df.copy()
         total = df_clean.isnull().sum().sort_values(ascending = False)
         percent = (df_clean.isnull().sum()/df_clean.isnull().count()*100).sort_values(ascending = False)
         missing_df_clean = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
         missing_df_clean
Out[46]:
```

	Total	Percent
quantity	1141	1.302973
weight_kg	1000	1.141957
country_or_area	0	0.000000
year	0	0.000000
comm_code	0	0.000000
commodity	0	0.000000
flow	0	0.000000
trade_usd	0	0.000000
quantity_name	0	0.000000
category	0	0.000000

Fig 2.

Moreover, since the total percentage of missing values of both 'weight_kg' and 'quantity' is lower than 2% and taking into account we cannot discover or assume the data (i.e. we cannot randomly write the amount of kilogram that was sent 20 years ago), we will just remove this rows in the Data Cleaning process.



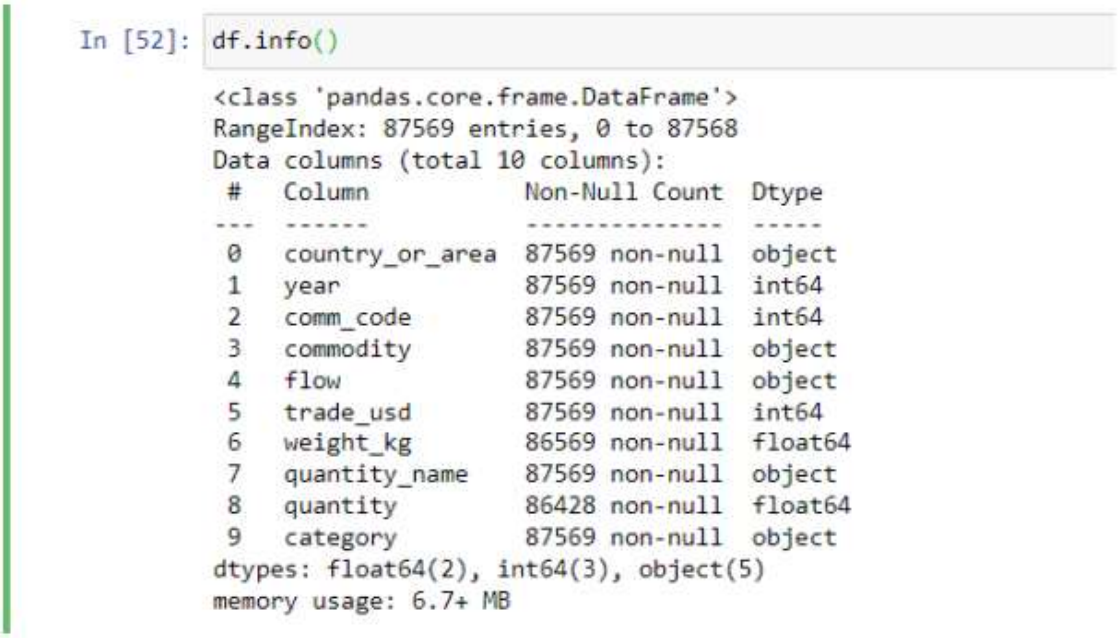
```
In [49]: # number of unique values for each variable
         df.nunique(axis=0)
Out[49]: country_or_area        7
         year                  29
         comm_code            441
         commodity            441
         flow                   4
         trade_usd          79886
         weight_kg          69704
         quantity_name          4
         quantity           69757
         category              12
         dtype: int64

In [50]: # statistical summary of numeric variables
         df.describe()
Out[50]:
```

	year	comm_code	trade_usd	weight_kg	quantity
count	87569.000000	87569.000000	8.756900e+04	8.656900e+04	8.642800e+04
mean	2003.223230	66145.457045	2.802791e+07	7.493461e+07	7.603971e+07
std	7.869093	34284.985511	2.306974e+08	2.358997e+09	2.362008e+09
min	1988.000000	10111.000000	1.000000e+00	0.000000e+00	0.000000e+00
25%	1997.000000	30614.000000	5.275900e+04	1.645500e+04	1.663600e+04
50%	2004.000000	70820.000000	6.052870e+05	2.208680e+05	2.285990e+05
75%	2010.000000	91030.000000	5.628730e+06	2.666730e+06	2.806166e+06
max	2016.000000	121490.000000	1.373323e+10	6.140000e+11	6.140000e+11

Fig 3.

As per the screenshot above, there are 12 categories that can be identified in the data. The year ranges from 1988 to 2016(29 values(years)as depicted in the screenshot).



```
In [52]: df.info()
```

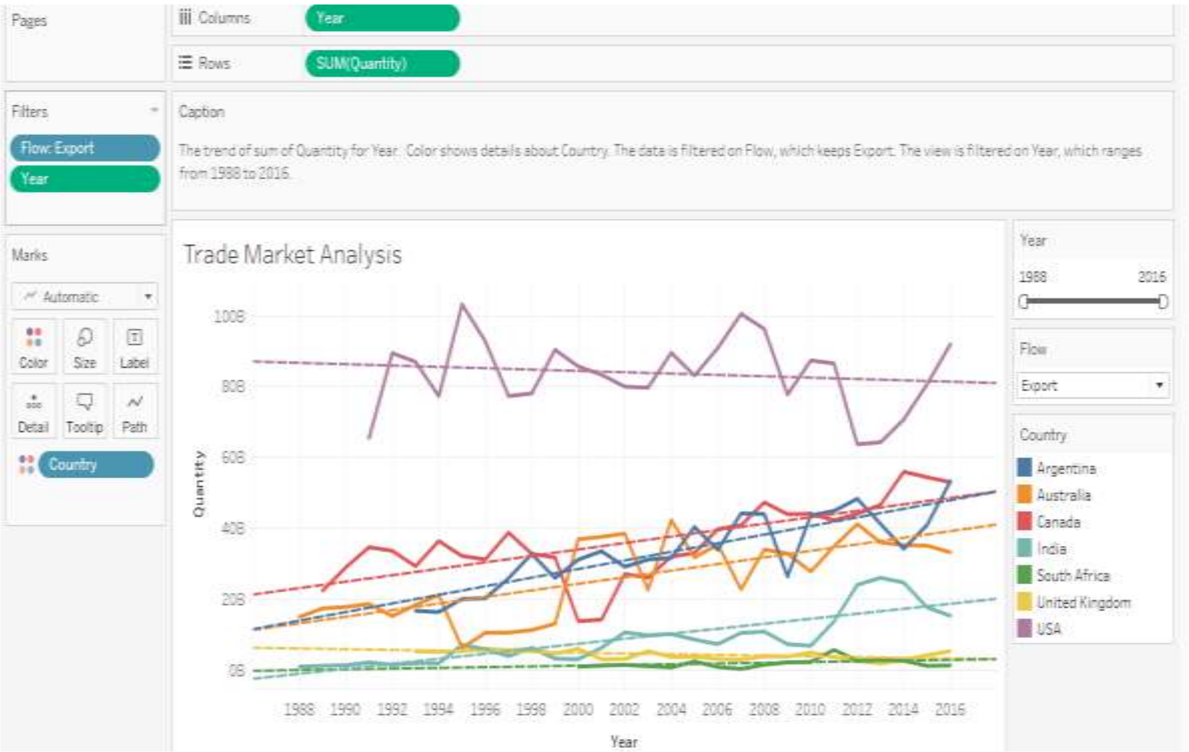
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87569 entries, 0 to 87568
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   country_or_area  87569 non-null  object 
 1   year             87569 non-null  int64  
 2   comm_code        87569 non-null  int64  
 3   commodity        87569 non-null  object 
 4   flow             87569 non-null  object 
 5   trade_usd        87569 non-null  int64  
 6   weight_kg        86569 non-null  float64
 7   quantity_name    87569 non-null  object 
 8   quantity         86428 non-null  float64
 9   category         87569 non-null  object 
dtypes: float64(2), int64(3), object(5)
memory usage: 6.7+ MB
```

Fig 4.

The dataset contains 10 columns consisting of datatypes like integer, float etc.

Listed below are few of the charts made in Tableau:

1. Line Chart



Pages

Columns: Year

Rows: SUM(Quantity)

Filters: Flow: Export, Year

Caption

The trend of sum of Quantity for Year. Color shows details about Country. The data is filtered on Flow, which keeps Export. The view is filtered on Year, which ranges from 1988 to 2016.

Marks: Automatic

Country

Year: 1988 – 2016

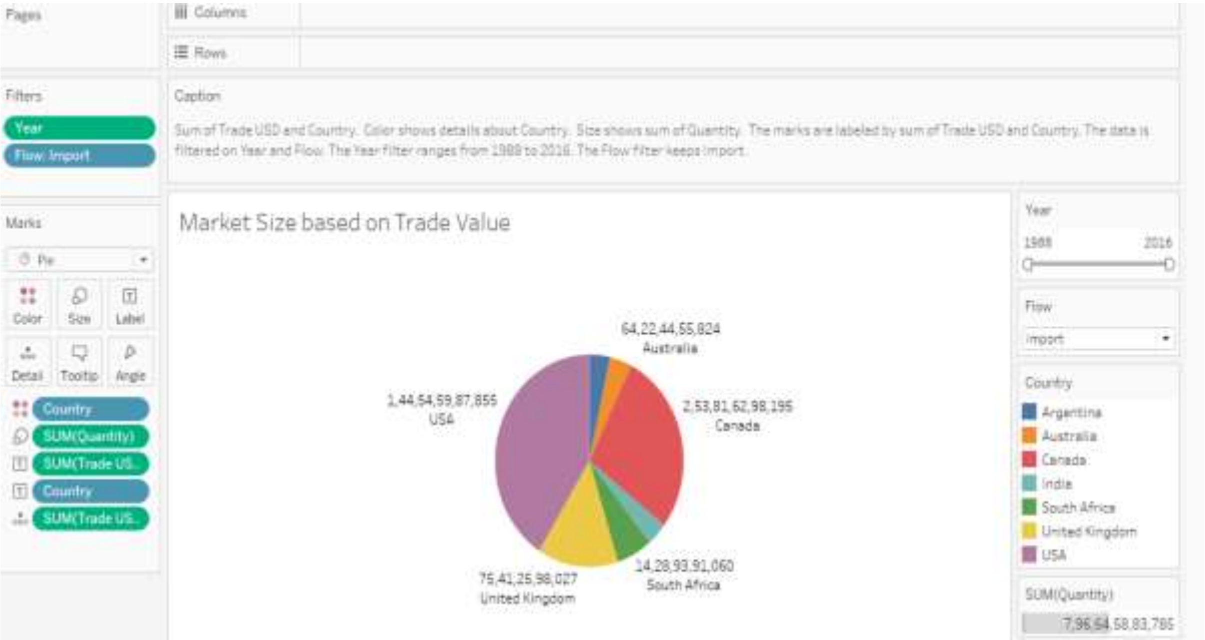
Flow: Export

Fig 5.

The line chart above depicts the quantity of commodities exported over the years for different countries. The graph can be used to do a trend analysis. Similar analysis can be done for “Import”. The range of years can be changed as per the requirement. Filters can be applied on the flow of commodities.

As per the chart, South Africa shows consistent quantities over the years, 2000 to 2016. On the other hand, USA shows varying trend in the year range of 1991 to 2016. The quantity is over 65B in the year 1991 and over 91B in the year 2016.

2. Pie Chart



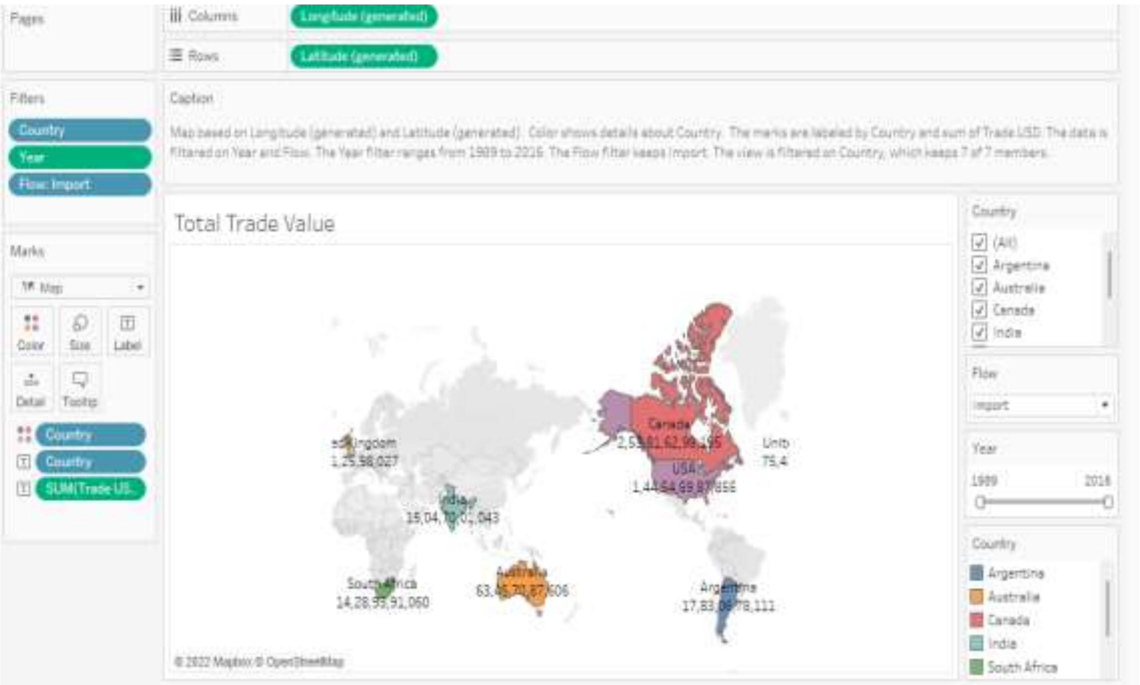
Caption

Sum of Trade USD and Country. Color shows details about Country. Size shows sum of Quantity. The marks are labeled by sum of Trade USD and Country. The data is filtered on Year and Flow. The Year filter ranges from 1988 to 2016. The Flow filter keeps Import.

Fig 6.

The above pie chart shows the distribution of Trade USD for different countries. The chart can be filtered based on flow(Import/Export). The time range can be changed as per requirement. As per the figures displayed in the chart, USA has the highest Import value and South Africa has the least value.

3. Geospatial Chart



Columns: Longitude (generated)

Rows: Latitude (generated)

Filters: Country, Year, Flow: Import

Marks: Country, Country, SUM(Trade US..

Caption

Map based on Longitude (generated) and Latitude (generated). Color shows details about Country. The marks are labeled by Country and sum of Trade USD. The data is filtered on Year and Flow. The Year filter ranges from 1989 to 2016. The Flow filter keeps Import. The view is filtered on Country, which keeps 7 of 7 members.

Country
☑ (All)
☑ Argentina
☑ Australia
☑ Canada
☑ India

Flow: Import

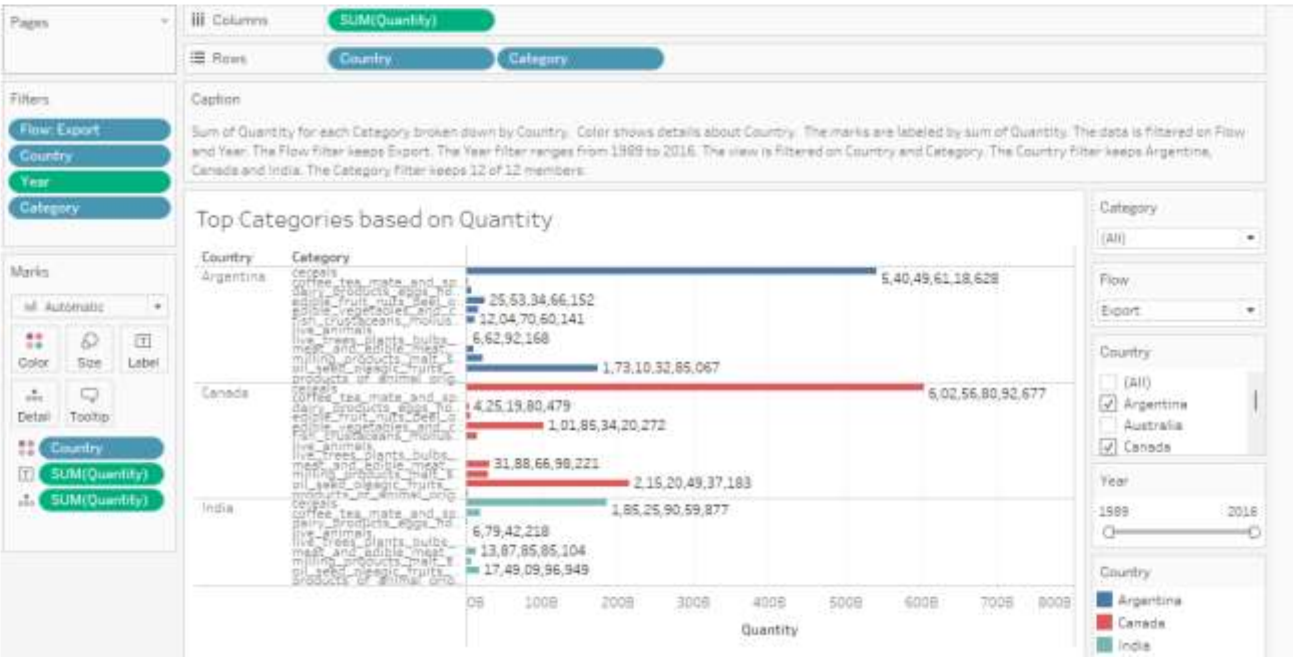
Year: 1989 – 2016

Country
- 🟦 Argentina
- 🟧 Australia
- 🟥 Canada
- 🟦 India
- 🟩 South Africa

Fig 7.

Geospatial chart is another way for representing the trade value. Various countries are displayed in the map here along with the Import trade values here. The countries can be chosen as per requirement. Also, the time range of data that needs to be displayed can be changed. The flow can be filtered on “Import” or “Export”.

4. Horizontal Bar Chart



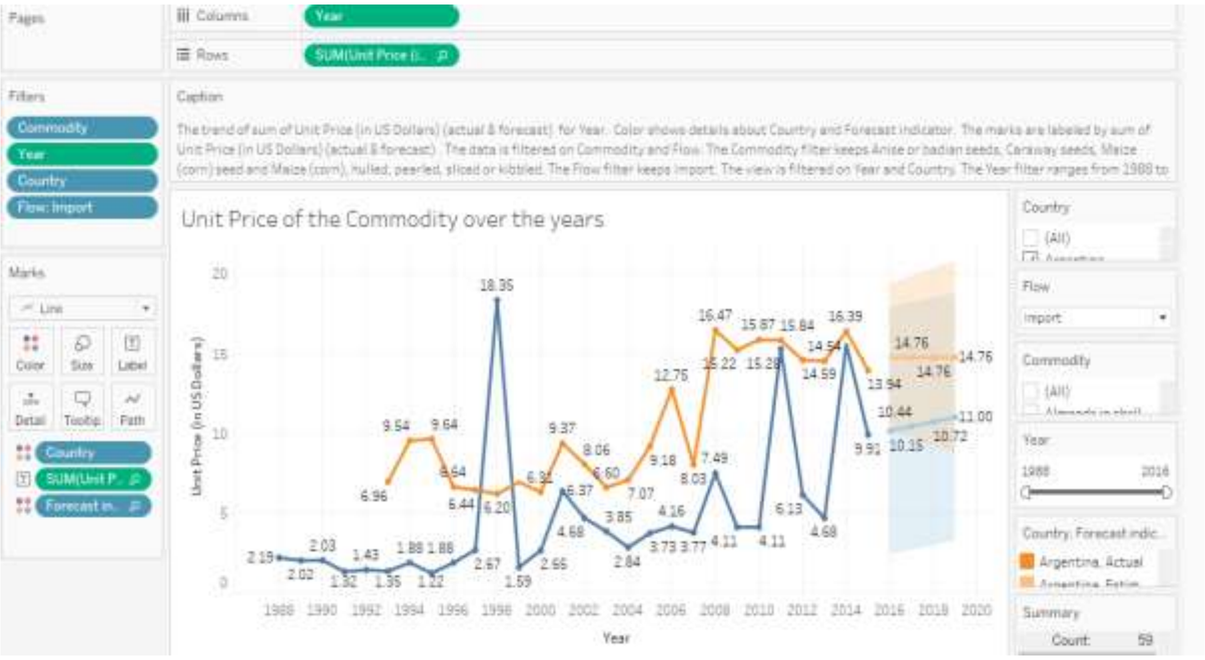
Caption

Sum of Quantity for each Category broken down by Country. Color shows details about Country. The marks are labeled by sum of Quantity. The data is filtered on Flow and Year. The Flow filter keeps Export. The Year filter ranges from 1989 to 2016. The view is filtered on Country and Category. The Country filter keeps Argentina, Canada and India. The Category filter keeps 12 of 12 members.

Fig 8.

The chart above shows the top categories based on Quantity for different countries. The countries can be chosen as per the data requirements. The Category can also be filtered upon. As per the data that needs to be shown, the flow can be filtered on Export and Import. As per the chart, out of Canada, India and Argentina, Canada has maximum export of cereal while India has the minimum value.

5. Line Chart with Forecast



Pages

Columns: Year

Rows: SUM(Unit Price (i...

Filters: Commodity, Year, Country, Flow: Import

Caption

The trend of sum of Unit Price (in US Dollars) (actual & forecast) for Year. Color shows details about Country and Forecast indicator. The marks are labeled by sum of Unit Price (in US Dollars) (actual & forecast). The data is filtered on Commodity and Flow. The Commodity filter keeps Anise or badian seeds, Caraway seeds, Maize (corn) seed and Maize (corn), hulled, pearled, sliced or kibbled. The Flow filter keeps Import. The view is filtered on Year and Country. The Year filter ranges from 1988 to

Marks: Line; Country, SUM(Unit P..., Forecast in...

Country: (All)

Flow: Import

Commodity: (All)

Year: 1988 – 2016

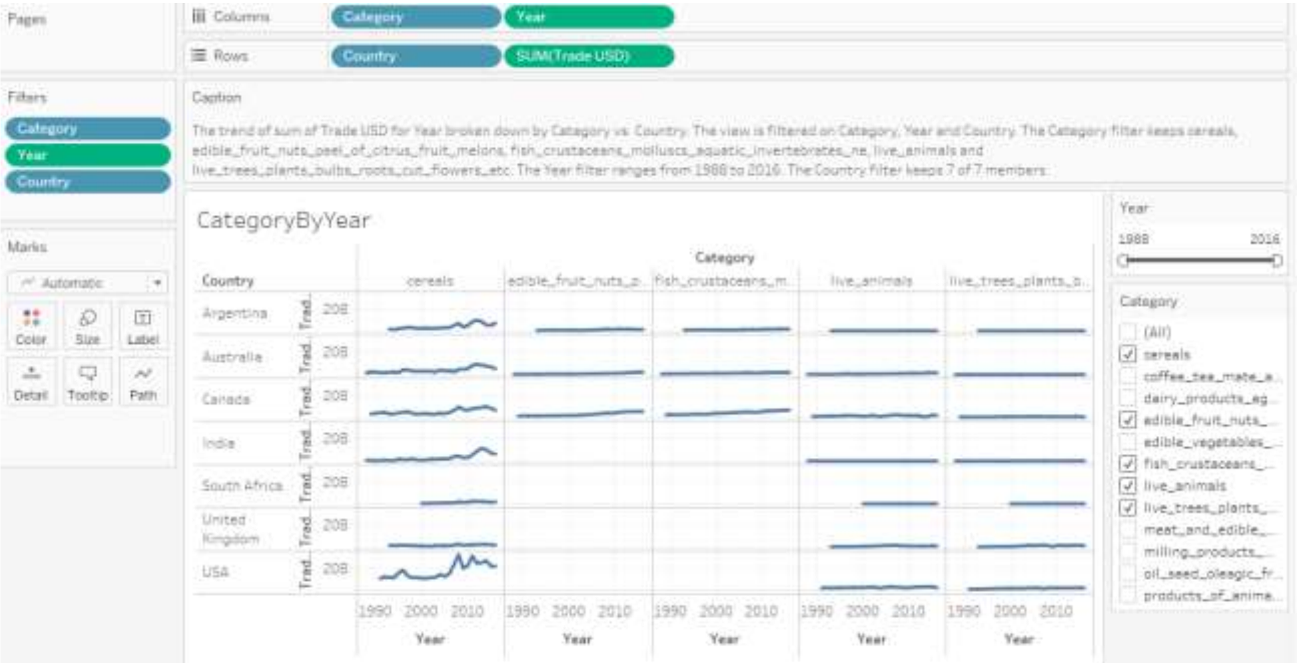
Country, Forecast indic...: Argentina, Actual; Argentina, Estim...

Summary: Count: 59

Fig 9.

The line chart shows the forecast for the Unit Prices of Argentina and India for the next 5 years. The countries can be filtered upon as per the data required. The time range can be changed too. Currently in the above chart, the flow is filtered on Import. It can be changed to Export as per the data that needs to be displayed. Also the graph above shows the unit price of commodities, Maize corn seed, Caraway seeds, Anise/badian seeds. This can be filtered too. As per the chart, the Unit Price is forecasted to be constant with the value, $14.46 for the next 5 years for Argentina. For India, the Unit price is estimated to be $10.15 in 2016, $10.44 in 2017, $10.72 in 2018 and $11 in 2019.

5. Correlation Chart



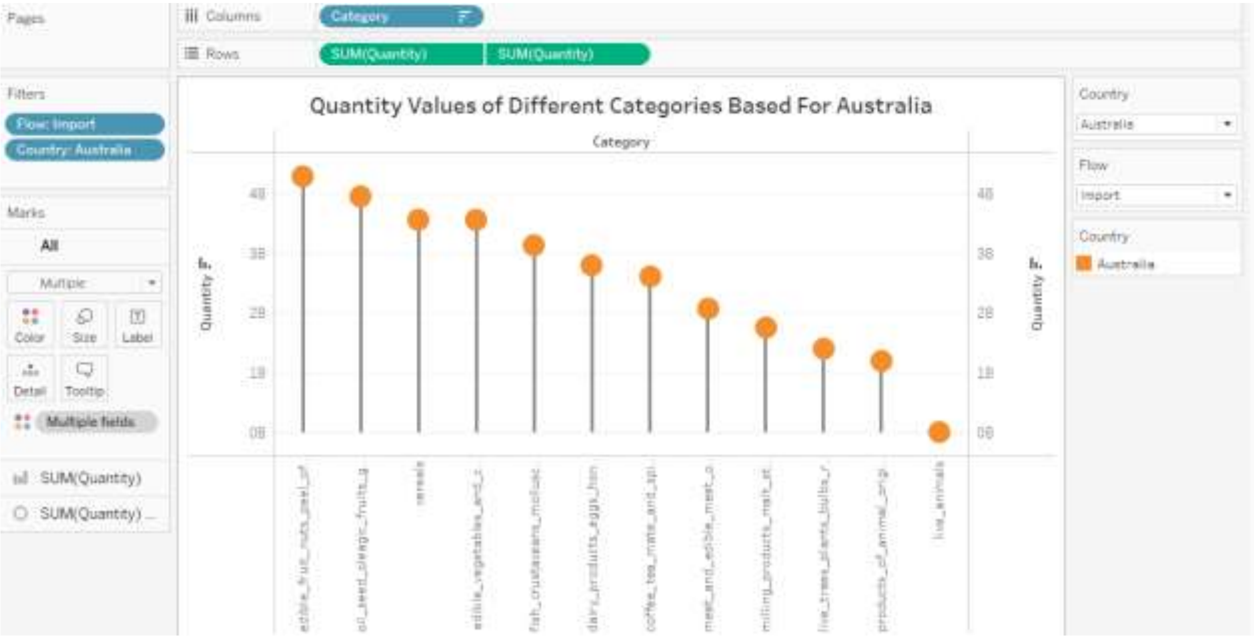
Caption

The trend of sum of Trade USD for Year broken down by Category vs. Country. The view is filtered on Category, Year and Country. The Category filter keeps cereals, edible_fruit_nuts_peel_of_citrus_fruit_melons, fish_crustaceans_molluscs_aquatic_invertebrates_ne, live_animals and live_trees_plants_bulbs_roots_cut_flowers_etc. The Year filter ranges from 1988 to 2016. The Country filter keeps 7 of 7 members.

Fig 10.

The above charts show the trend of Trade USD value per country and category. Filters can be applied based on Category and Year. For example, the above chart shows that cereals have the trade value USD constant for some time and increases to about 7B dollars in the year 2016 in Argentina. Fruits and nuts, fish, live animals and trees have constant trade values with no sudden change over the years 1993 to 2016.

6 (a) Lollipop Chart



Columns: Category

Rows: SUM(Quantity) SUM(Quantity)

Filters: Flow: Import, Country: Australia

Marks: All, Multiple fields, SUM(Quantity), SUM(Quantity) ...

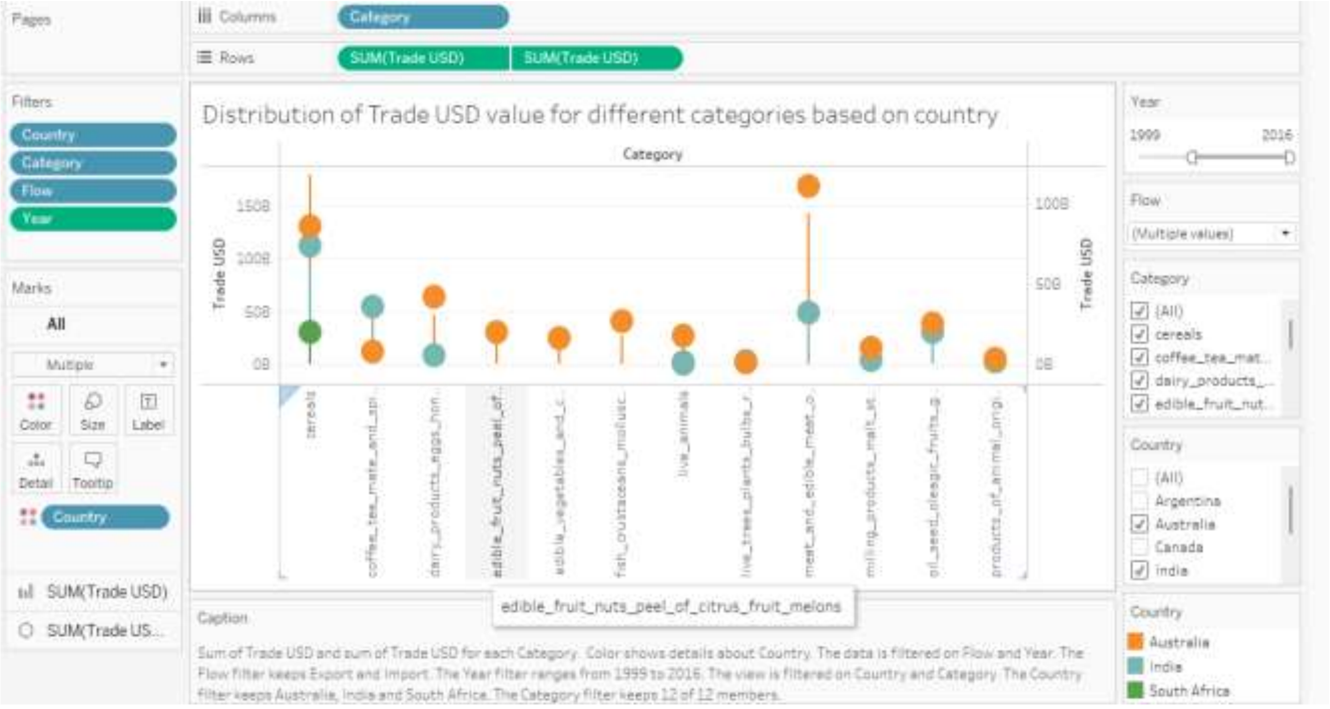
Country: Australia

Flow: Import

Fig 11.

The above chart shows the comparison of quantities of different categories. The flow can be filtered on “Import” and “Export”. Also the chart can be filtered based on country. For eg. The chart above shows the Import quantities for different categories in Australia. The chart has been sorted in descending order. Edible fruits/nuts has the highest quantity of imports and live animals has the least imports in Australia.

(b)Lollipop Chart



Caption

Sum of Trade USD and sum of Trade USD for each Category. Color shows details about Country. The data is filtered on Flow and Year. The Flow filter keeps Export and Import. The Year filter ranges from 1999 to 2016. The view is filtered on Country and Category. The Country filter keeps Australia, India and South Africa. The Category filter keeps 12 of 12 members.

Fig 12.

Trade USD Value can also be studied using the Lollipop chart. The graph shows the distribution of trade value of different categories based on different countries. Country wise filters are available. The year span can be changed as per the data required. Category specific filters can be applied along with filters on Import or Export. In the above chart, for the commodity, coffee/tea, India has a higher trade value (about $35B) as compared to Australia($7B). For the commodity, meat and edible meat, India has lower trade value and Australia has a way more trade value of about $110B.

7. Waterfall Chart

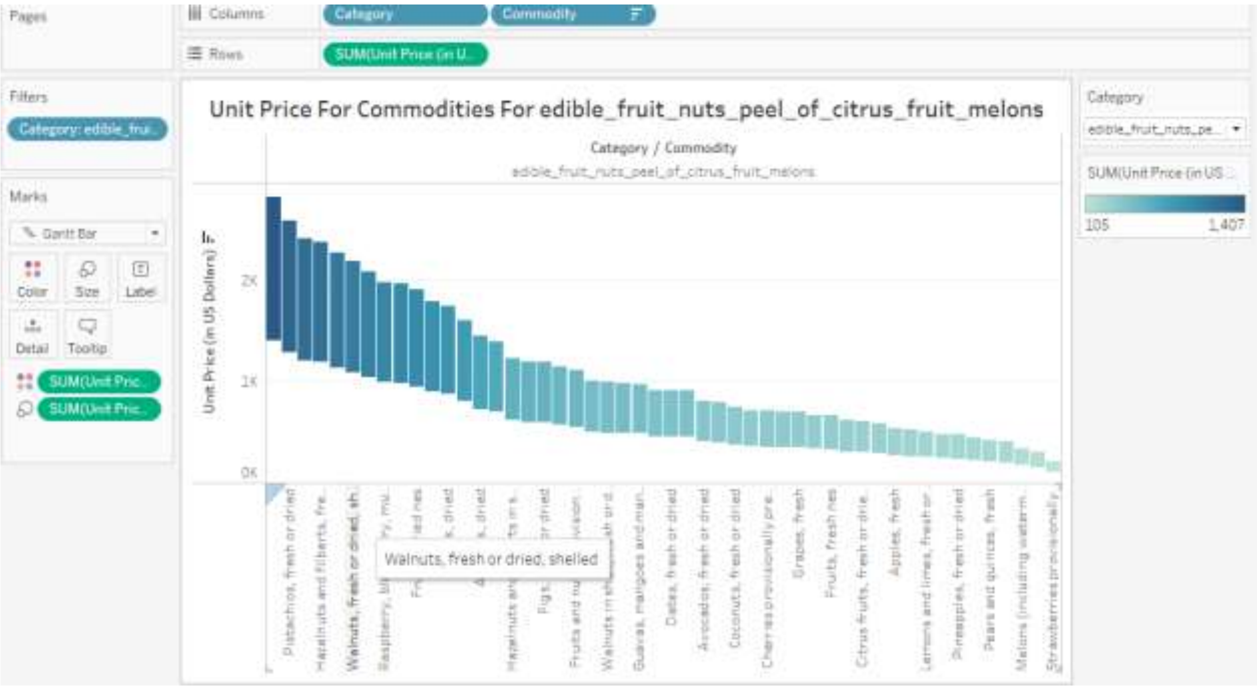


Fig 13.

The waterfall chart above shows the unit price for commodities with respect to different countries. Category related filter is available. The values have been sorted in the chart above to show the decreasing order in the Unit Price USD. As depicted in the graph above, Pistachios have the highest unit price of $1407 and Strawberries have the lowest unit price of $105. The unit price for the edible nut commodity ranges from $105 to $1407.

8. Bump Chart

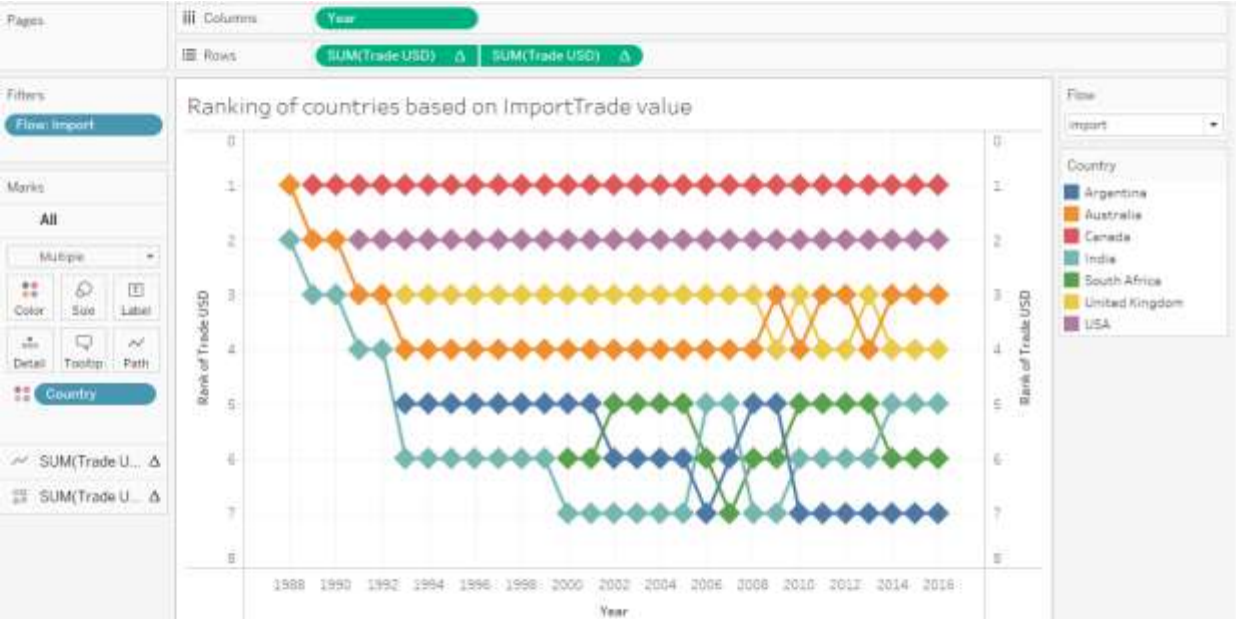


Fig 14.

The bump chart above shows the ranking of different countries based on the import and export trade value. Filters based on flow can be applied. Year ranges from 1988 to 2016. In the graph above, in the year 1988, Australia is ranked as 1 and India is ranked as 2 as per the Trade USD values. In the year 1990, Canada is ranked the highest , followed by Australia and India.

9. Donut chart

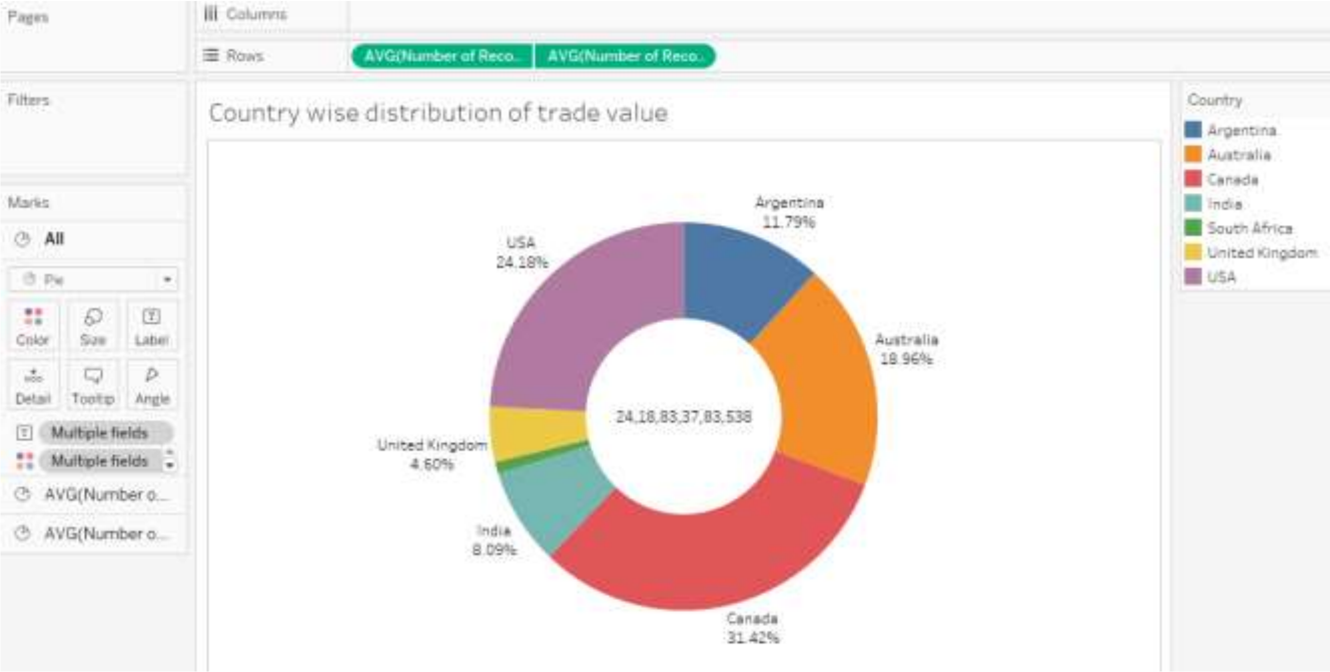
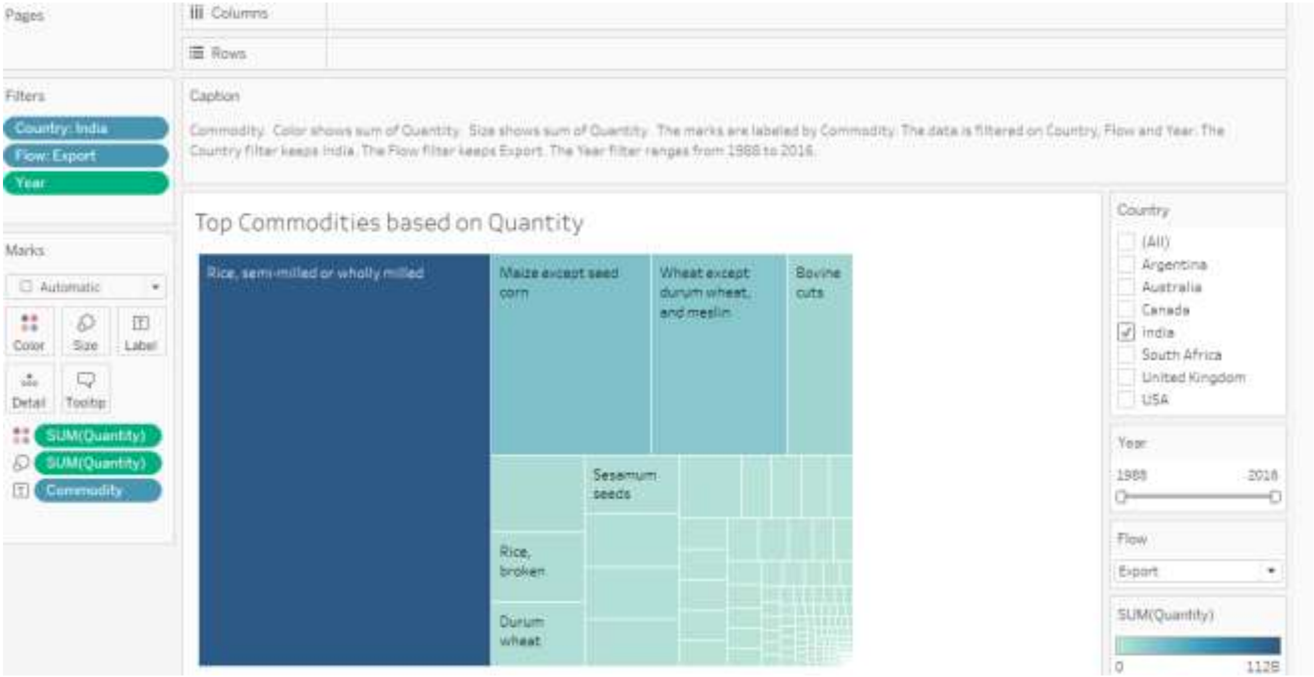


Fig 15.

The above chart shows the percentage distribution of trade value for different countries. Canada shows the highest distribution of trade USD value. South Africa has the lowest trade USD value. Sum of trade values is represented in the centre. This chart helps in explaining the split in trade values.

10. Heat Map



Filters: Country: India, Flow: Export, Year

Marks: SUM(Quantity), SUM(Quantity), Commodity

Caption

Commodity. Color shows sum of Quantity. Size shows sum of Quantity. The marks are labeled by Commodity. The data is filtered on Country, Flow and Year. The Country filter keeps India. The Flow filter keeps Export. The Year filter ranges from 1988 to 2016.

Country: ☐ (All) ☐ Argentina ☐ Australia ☐ Canada ☑ India ☐ South Africa ☐ United Kingdom ☐ USA

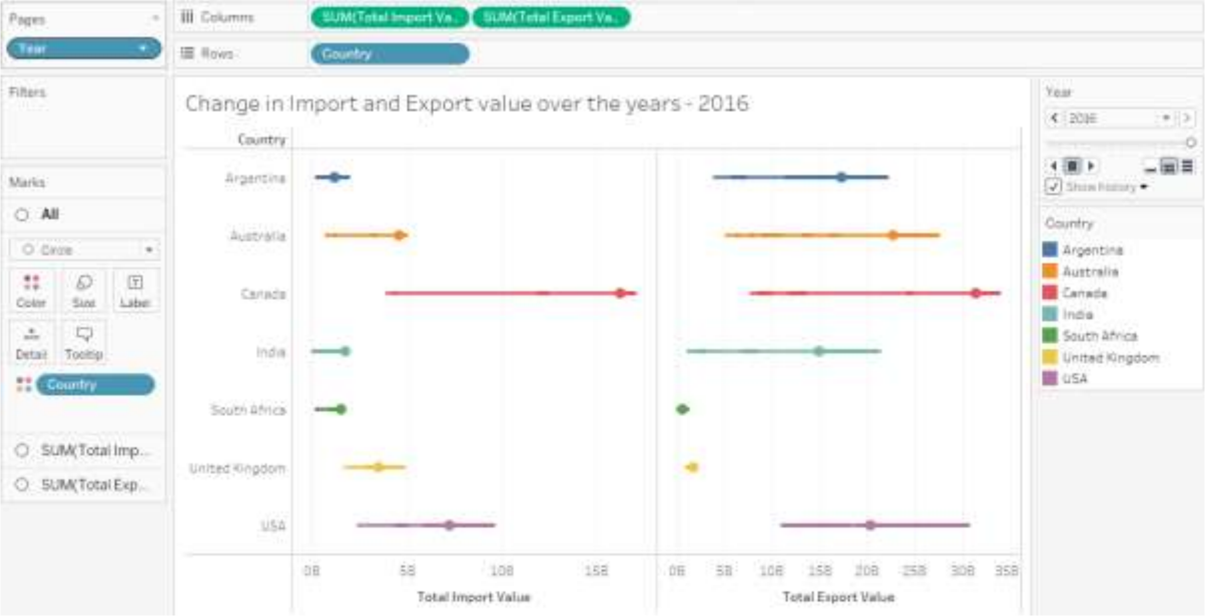
Year: 1988 – 2016

Flow: Export

Fig 16.

The heat map above shows the top commodities for exports. Country level filters can be applied. Year range can be changed as per the requirement. As per the values depicted in the graph above, India has the highest exports of rice, semi milled or wholly milled. Mushroom spawn is not exported from India. Hence it has the quantity as zero.

11. Connected Scatterplot



Pages: Year

Columns: SUM(Total Import Va.. SUM(Total Export Va..

Rows: Country

Marks: Circle; Country; SUM(Total Imp... SUM(Total Exp...

Year: 2016

Fig 17.

The Chart above shows the change that has occurred in the import and export values over the years. The time period can be selected as required. This chart has a live feature hence the changes can be seen as and when it happens. For eg: In the year 2016, USA has export value of $20B and import value of $7B.

Insights from the data

- Correlation chart between the measure values:

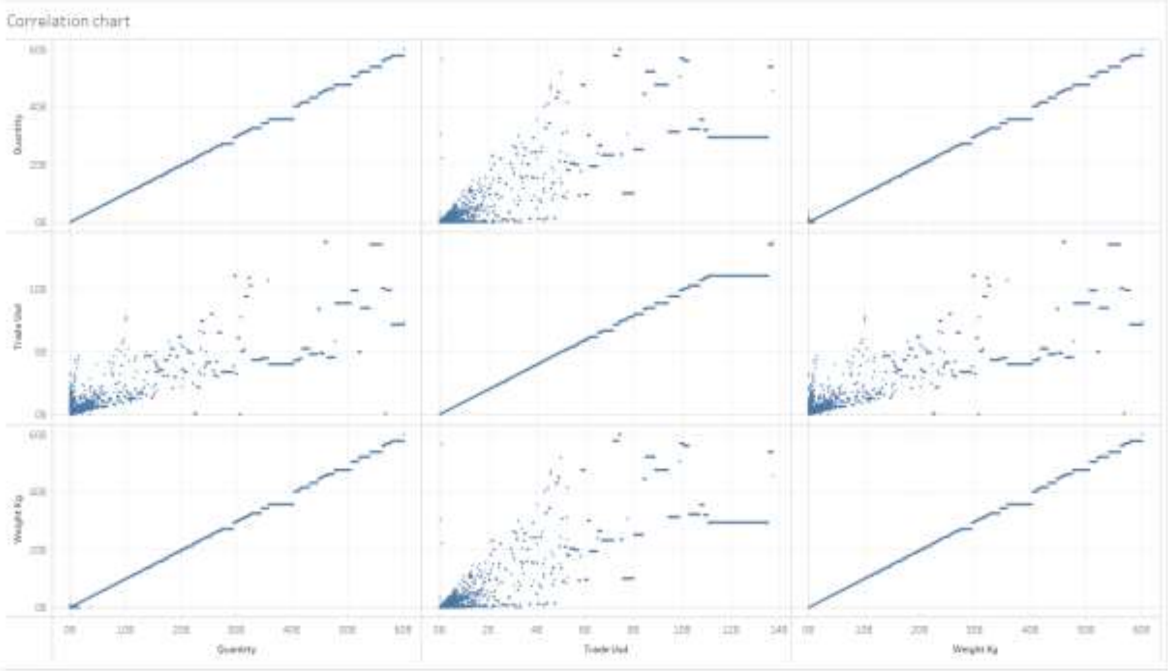


Fig 18.

- From 1988 to 2016, the USA has been the largest exporter and importer of commodities followed by Canada at 2nd position.

- Top categories were identified for each country.

- Market size of each country based on the quantity

- After identifying the potential commodities, we can observe their quantity and unit price over the past years and it is different for each country. Further, we can use forecasting techniques to get the estimate of the unit prices for future years.