TABLE OF CONTENT :

INTRODUCTION :

Data :

It is information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer. Everyday huge amount of data is being produced . It needs to be stored and analyzed to draw some insights from it and that is helpful in optimizing the performance in Businesses.

Data Analysis :

It is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision- making.

Why data analysis is important ?

It is essential as it helps businesses understand their :

- customers better
- improves sales
- improves customer targeting
- reduces costs
- allows for the creation of better problem-solving strategies.

To show the analyzed data in an understandable way for clients we need Data visualization.

Data visualization :

It is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

It helps clients in :

- Analyzing the Data in a Better Way
- Faster Decision Making
- Making Sense of Complicated Data

The dataset that I used is about sales of various toy stores in multiple countries. This dataset has data of toy vehicles like cars, trains, planes, ships etc. It has data from 2003 to 2005 , but only up to month may in 2005. It was Originally Written by María Carina Roldán, Pentaho Community Member, BI consultant (Assert Solutions), Argentina. This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License. Modified by Gus Segura June 2014. It has a total of 25 columns and 2823 rows. Columns in this dataset are :

1. ORDERNUMBER
2. QUANTITYORDERED
3. PRICEEACH
4. ORDERLINENUMBER
5. SALES
6. ORDERDATE
7. STATUS
8. QTR_ID
9. MONTH_ID
10. YEAR_ID
11. PRODUCTLINE
12. MSRP
13. PRODUCTCODE
14. CUSTOMERNAME
15. PHONE
16. ADDRESSLINE1
17. ADDRESSLINE2
18. CITY
19. STATE
20. POSTALCODE
21. COUNTRY
22. TERRITORY
23. CONTACTLASTNAME
24. CONTACTFIRSTNAME
25. DEALSIZE

Sample data of Dataset :

# ORDERNU...	# QUANTITY...	# PRICEEACH	# ORDERLIN...	# SALES	ORDERDATE	STATUS
10107	30	95.7	2	2871	2/24/2003 0:00	Shipped
10121	34	81.35	5	2765.9	5/7/2003 0:00	Shipped
10134	41	94.74	2	3884.34	7/1/2003 0:00	Shipped
10145	45	83.26	6	3746.7	8/25/2003 0:00	Shipped
10159	49	100	14	5205.27	10/10/2003 0:00	Shipped
10168	36	96.66	1	3479.76	10/28/2003 0:00	Shipped
10180	29	86.13	9	2497.77	11/11/2003 0:00	Shipped
10188	48	100	1	5512.32	11/18/2003 0:00	Shipped
10201	22	98.57	2	2168.54	12/1/2003 0:00	Shipped
10211	41	100	14	4708.44	1/15/2004 0:00	Shipped

OBJECTIVES / SCOPE OF THE ANALYSIS :

Aim of this analysis is to give results of the following analysis and give future trends :

1. Annual Revenue
2. Sales by Year
3. Sales by Country
4. Status of order
5. Sales by Product line
6. Sales by Product Code
7. To give Top 10 retailer names
8. Sales by Deal size
9. Sales in future years
10. Sales by Quantity

SOURCE OF DATASET :

I've extracted this dataset from a website called Kaggle . Dataset is named as Sample Sales Data.
Link to this dataset :
https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

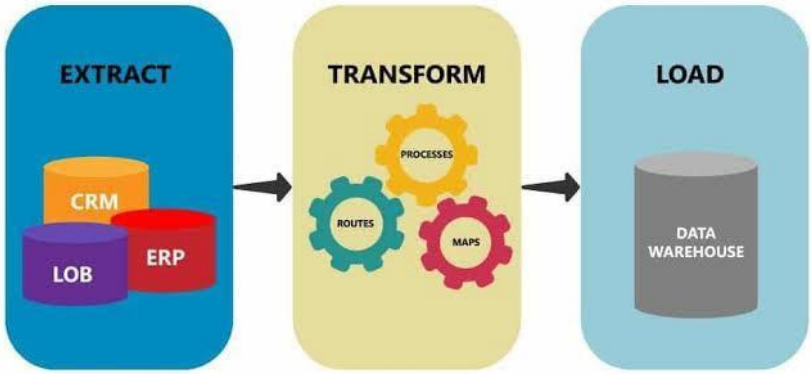
ETL PROCESS :

Extract, transform, and load (ETL) are 3 data processes, followed after data collection.

Extraction takes data, collected in data sources like flat files, databases (relational, hierarchical etc.), transactional datastores, semi-structured repositories (e.g. email systems or document libraries) with different structure and format, pre-validating extracted data and parsing valid data to destination (e.g. staging database)

Transformation takes extracted data and applies predefined rules and functions to it, including selection (e.g. ignore or remove NULLs), data cleansing, encoding (e.g. mapping "Male" to "M"), deriving (e.g. calculating designated value as a product of extracted value and predefined constant) , sorting, joining data from multiple sources (e.g. lookup or merge), aggregation (e.g. summary for each month), transposing (columns to rows or vice versa), splitting, disaggregation, lookups (e.g. validation through dictionaries), predefined validation etc. which may lead to rejection of some data. Transformed data can be stored into Data Warehouse (DW).

Load takes transformed data and places it into end target, in most cases called Data Mart (sometimes they called Data Warehouse too). Load can append, refresh or/and overwrite pre-existing data, apply constraints and execute appropriate triggers (to enforce data integrity, uniqueness, mandatory fields, provide log etc.) and may start additional processes, like data backup or replication.



ETL - Extract, Transform, Load

Tableau Prep is an ETL tool (Extract Transform and Load) that allows you to extract data from a variety of sources, transform that data, and then output that data to a Tableau Data Extract (using the new Hyper database as the extract engine) for analysis.

How Does Tableau Prep Work?

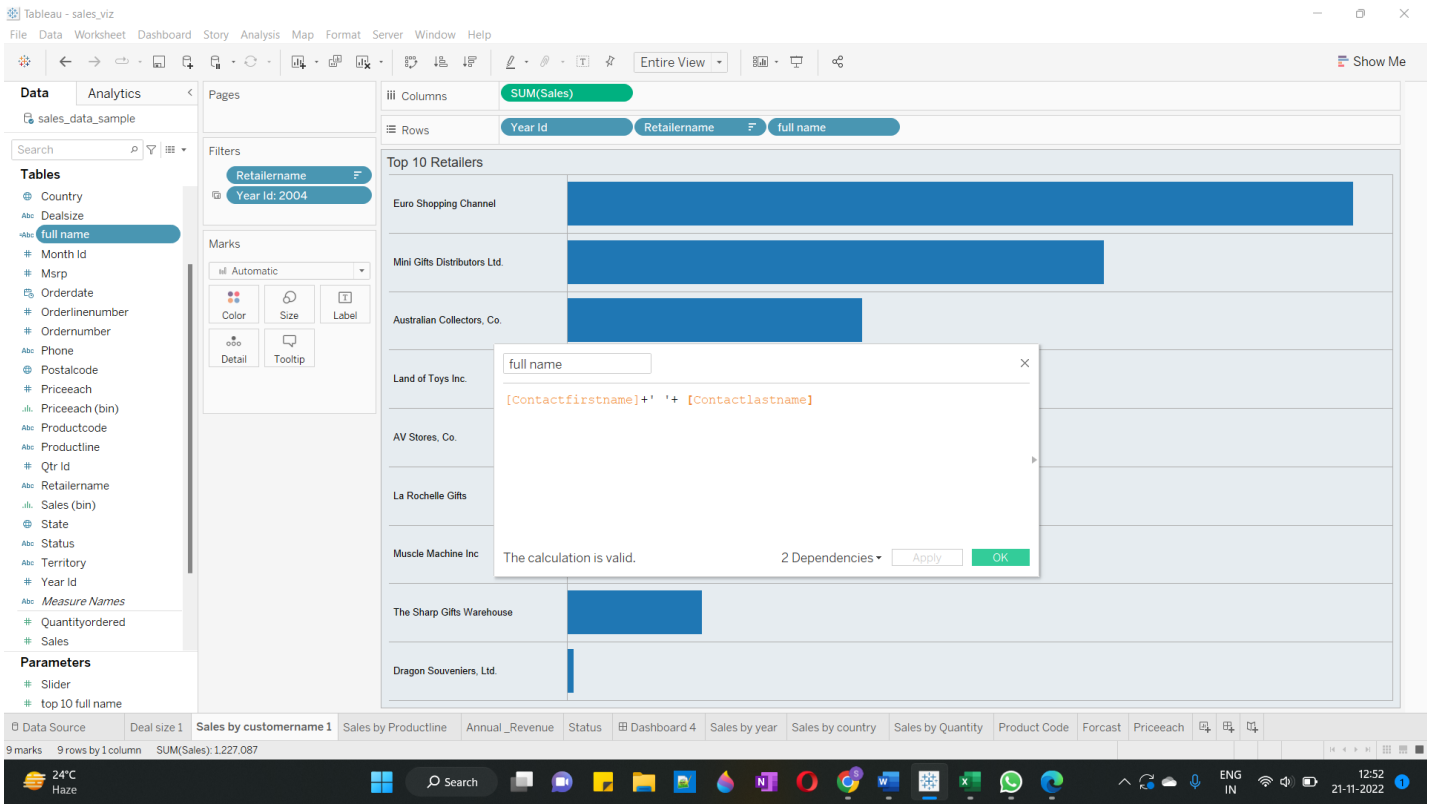
Tableau Prep helps you **examine and visualize your data**, enabling you to do the following:

1. **Connect and extract data**
2. **Understand data:**
    a. Number of columns/fields in your data
    b. Number of records
    c. Data types of fields
    d. Number of distinct values in a field
    e. Visualize how the data is distributed by field
3. **Identify issues and errors**
4. **Clean/Modify and Filter data**
    a. Rename fields
    b. Remove fields
    c. Modify/change values in a field
    d. Split fields
    e. Aggregate data
    f. Filter out data
5. **Enhance data**
    a. Add Calculated fields
    b. Join additional data
    c. Union additional data
6. **Output resulting data for use in analysis and reporting**

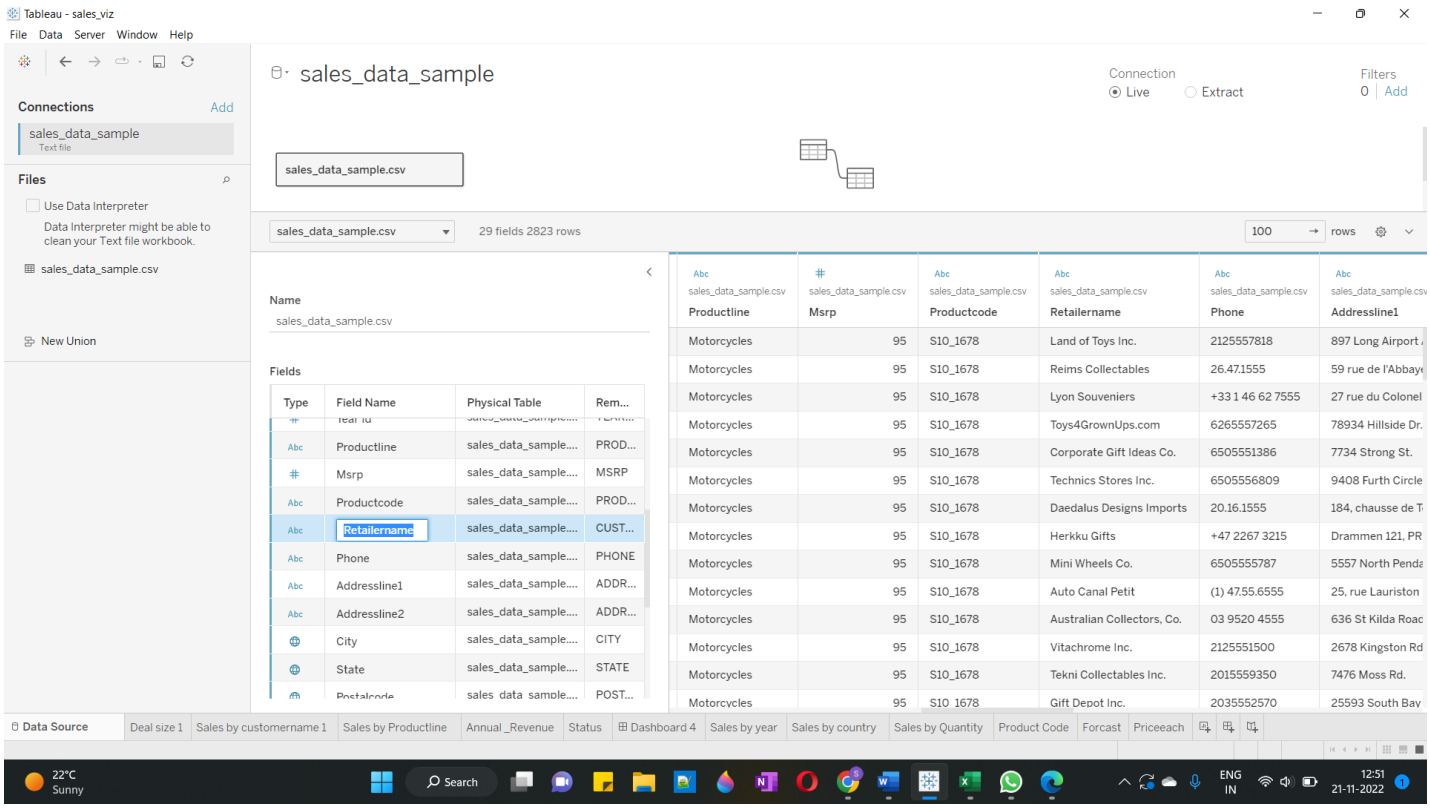
And, once you get new data (as long as it is in the same format and same field names), the ETL process **you created is reusable**. No longer will you have to repeat the process and steps necessary to transform your data each time the source data is updated, instead the ETL process flow has all the steps and logic you built. All you need to do is re-run the flow to get the new data output, resulting in **many hours saved** from data processing and cleansing, which can be used for analysis instead!

- Firstly I've extracted data from Kaggle and connected it to tableau.
- Tableau automatically gave no of fields, records and the data types of fields.
- Then I've transformed my data in following ways inorder to be make it clean and usable way :

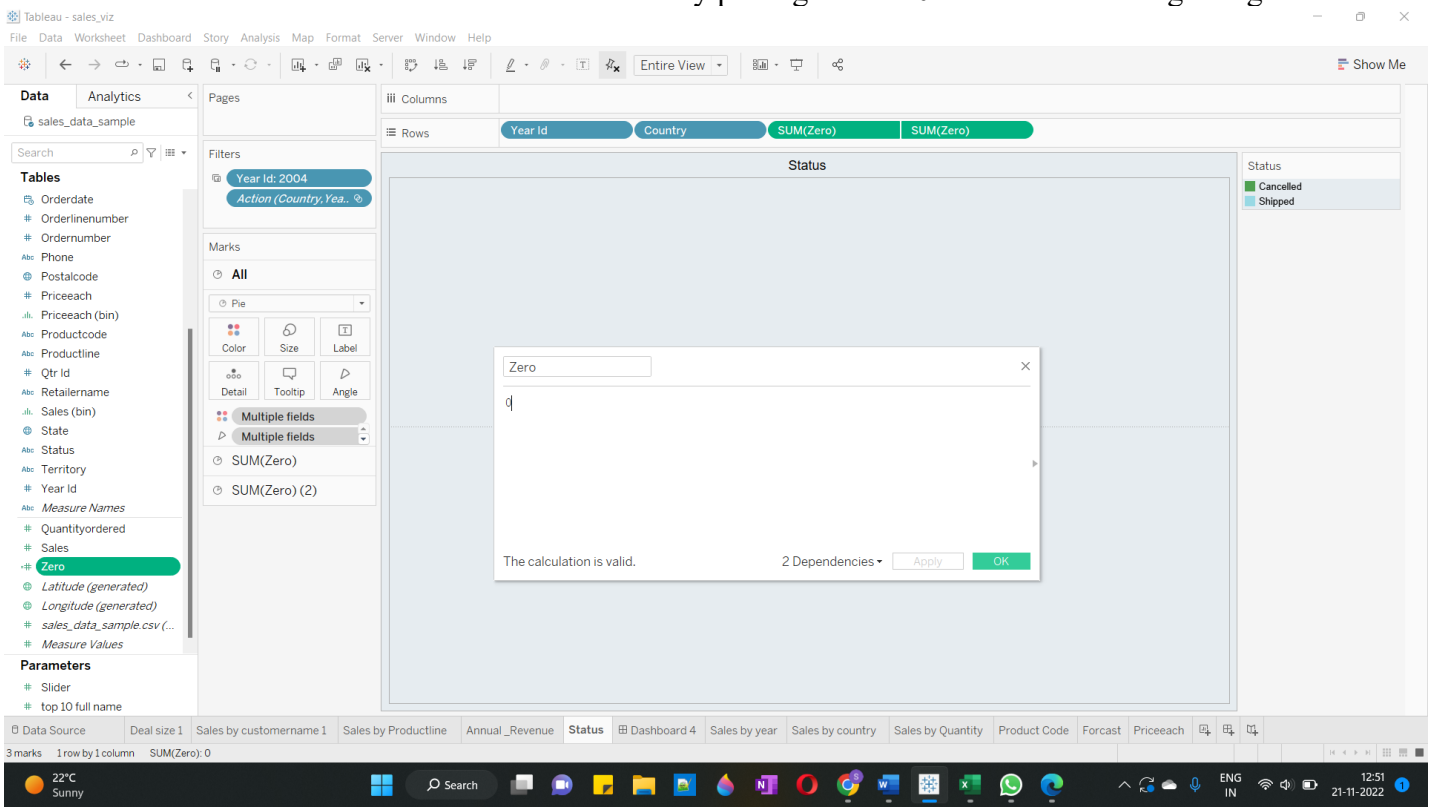
1. **Full name :**
   I've made a calculated field named "full name".
   Formula = [contactfirstname] + ' ' + [contactlastname]



2. **Retailer name :**
I've changed a field name called Customername to retailername because it should be named after retailers but had been named as Customername.

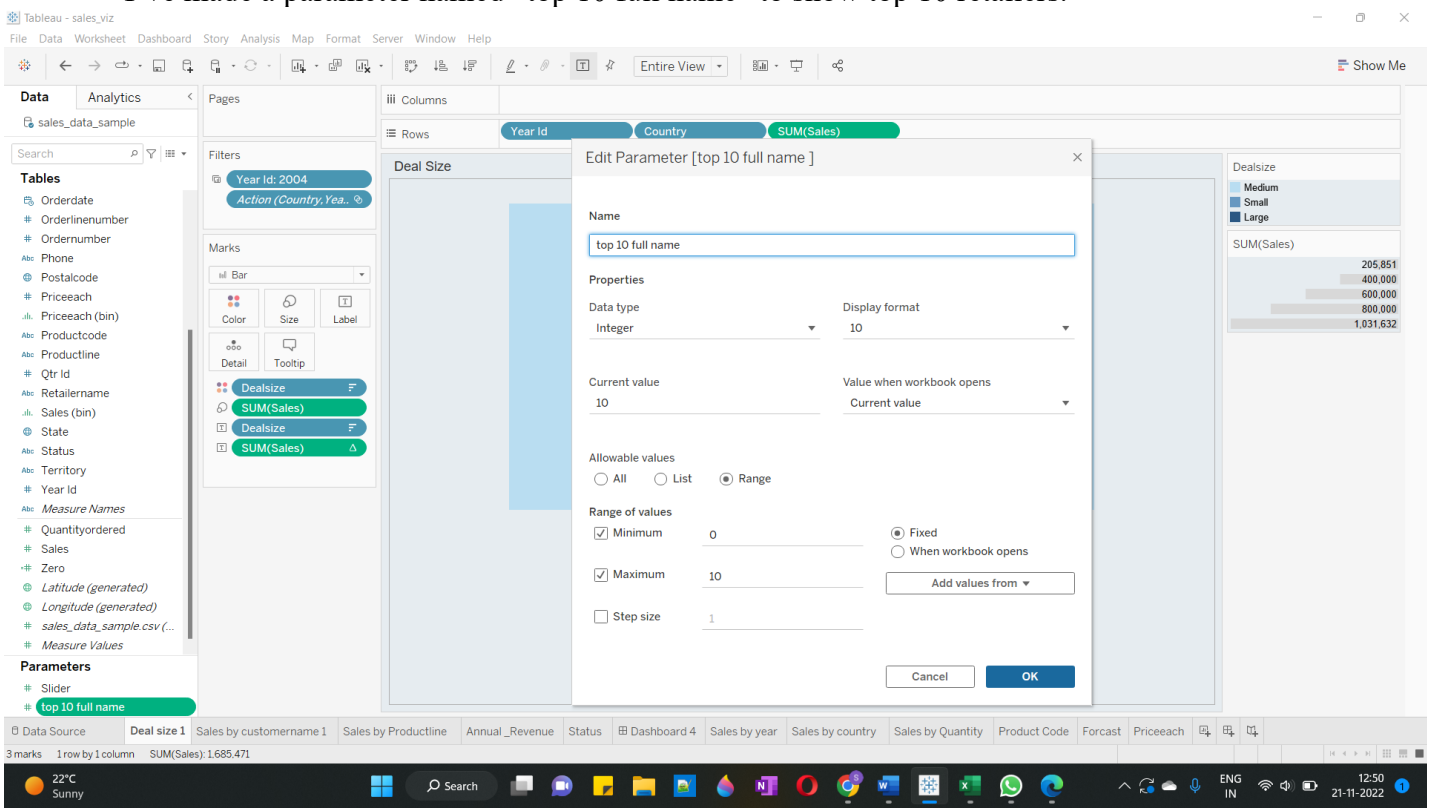


3. **Calculated field :**
I've made a calculated field named "Zero" by putting a value 0. This is for making Doughnut Chart.



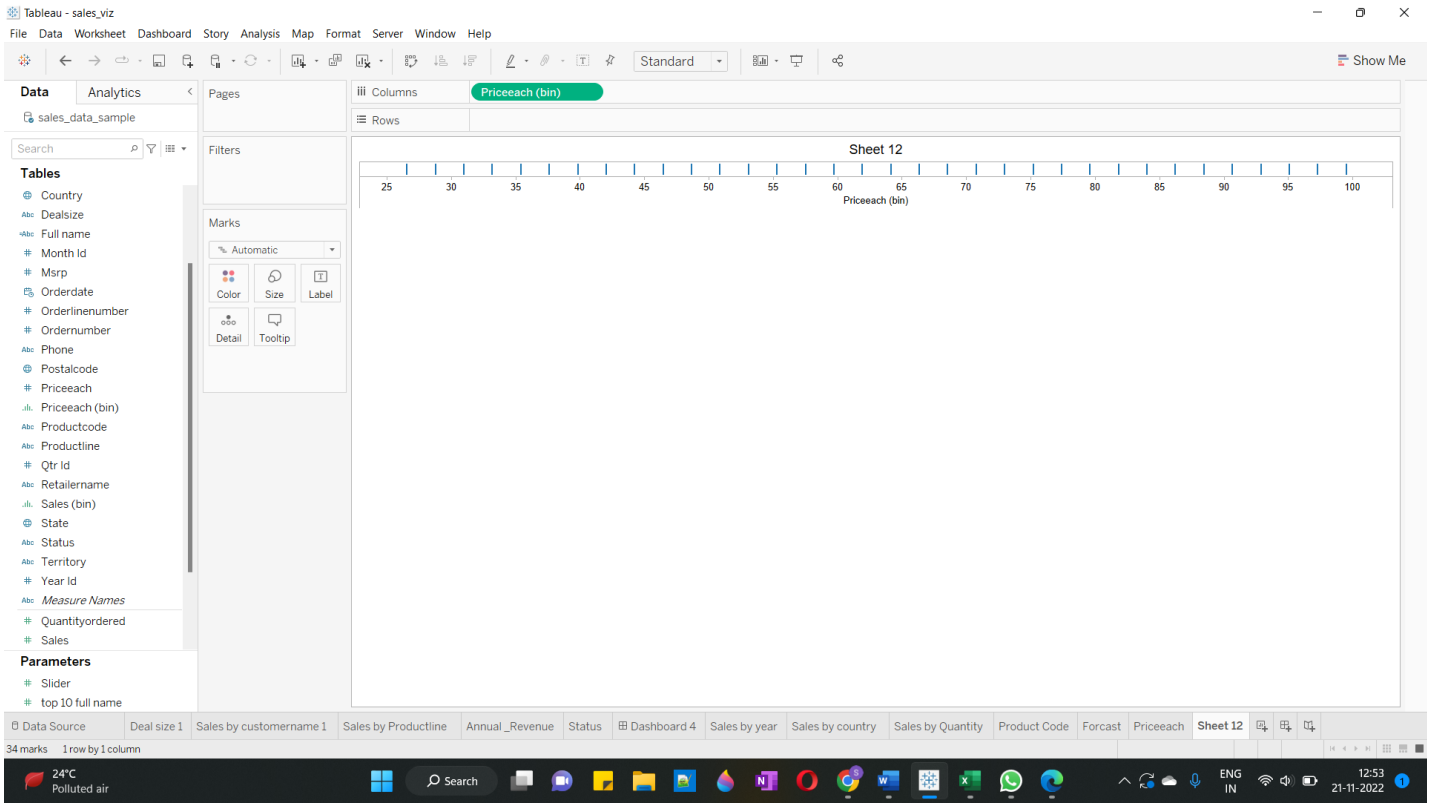
4. Parameter :

I've made a parameter named "top 10 full name" to show top 10 retailers.



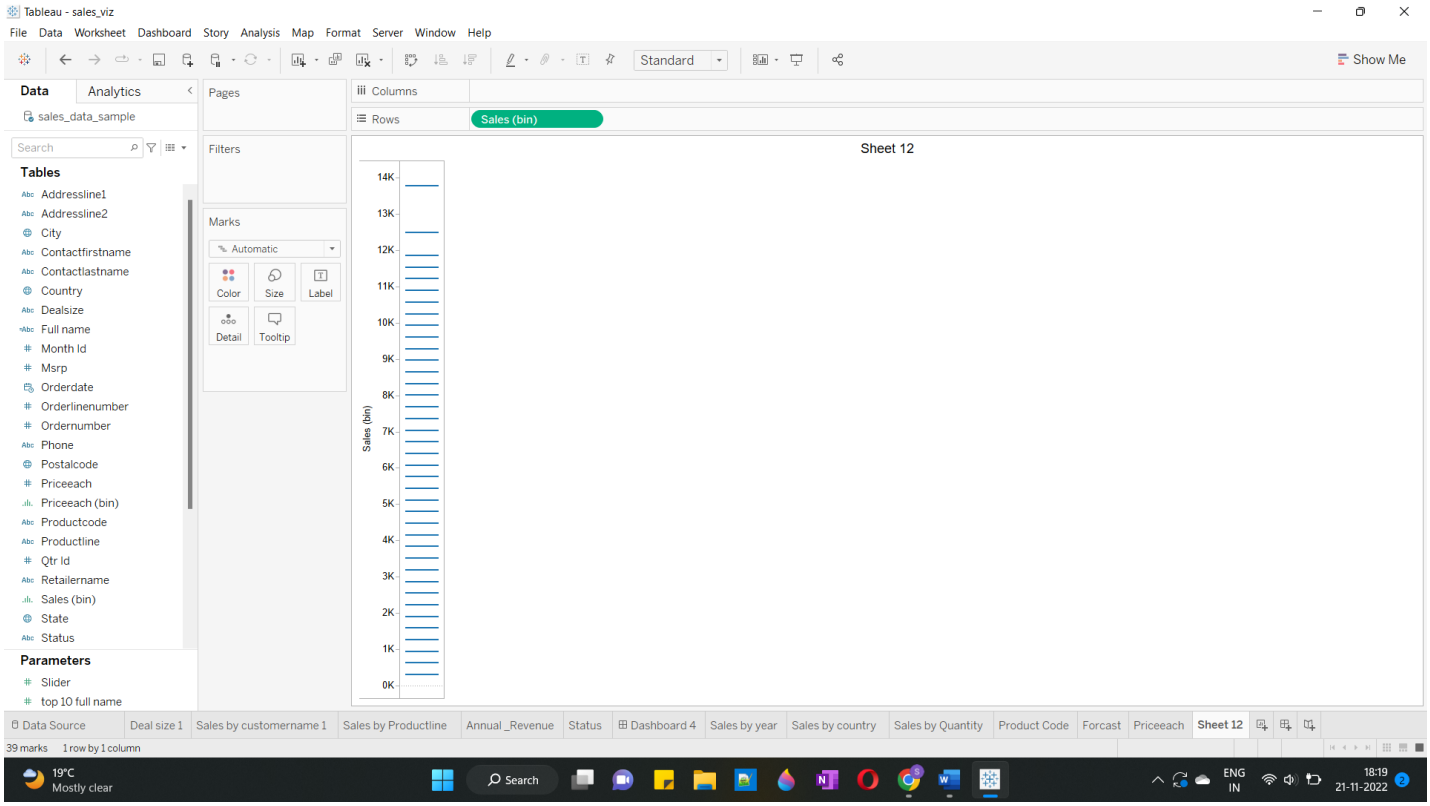
5. Priceeach bin :

Tableau automatically created a bin for Priceeach field . This helps in making Histograms.



6. Sales bin :

Tableau automatically created a bin for sales field . This helps in making Histograms.



ANALYSIS ON DATASET :

1. Deal size :

i. Introduction :

This visualization shows us the sales of a particular dealsize , like when the order is small , medium or large sized.

ii. General Description :

- Here I've dropped column Dealsize in colors in marks card.
- Columns Sales, Country and Year Id in rows.
- I've also dropped columns Dealsize and Sales in label in marks card and sales in size in marks card. I chose **Funnel chart** to show the result.

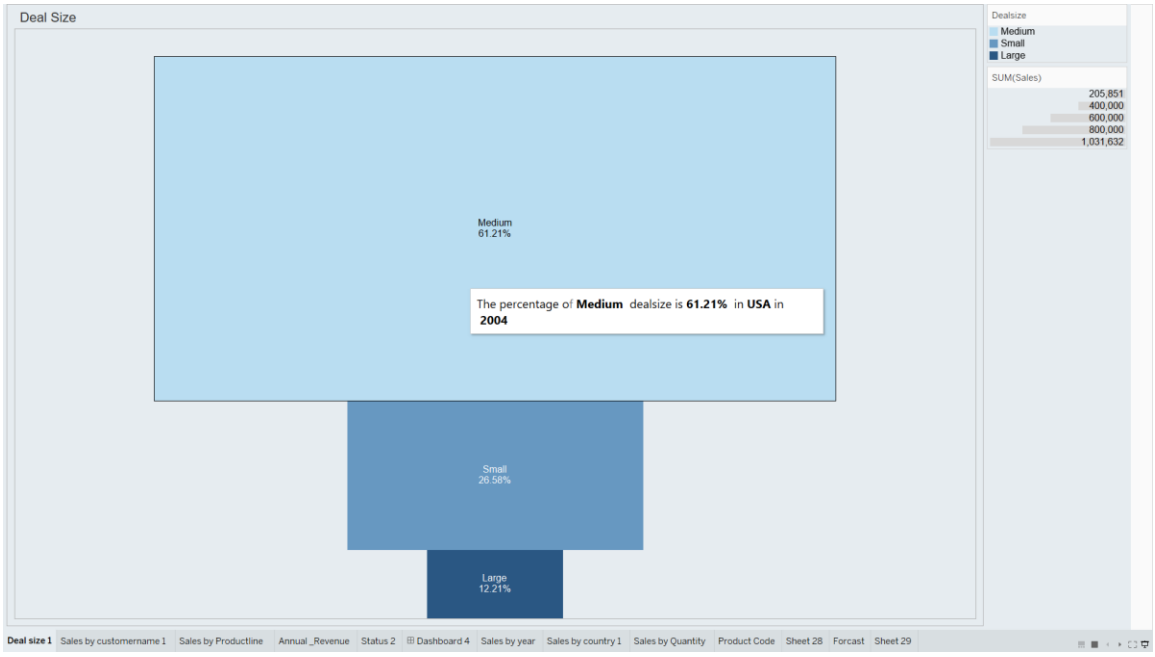
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **Percent of total** option in **Quick table calculation** to show sales in Percentage.
- I used **filters** to show results of particular country in a particular year.

iv. Analysis results :

- In all the three years **Medium dealsize** has the **most sales.**
- In **2003** and **2004 large dealsize** has **least sales** and in **2005 small dealsize** has **least sales**.
- Likewise with the filter option we can check a particular country's sales.

v. Visualization :



2. Top 10 Retailers :

i. Introduction :

This visualization shows us the sales of top 10 retailers in a particular year.

ii. General Description :

- I've dropped column Retailer name and full name in rows and sales in column.
- I edited Retailer name column's filter option , where I chose "top 10 full name" option to show up the top 10 retailers with the highest sales. I chose **Horizontal Bar Graph** to show the result.

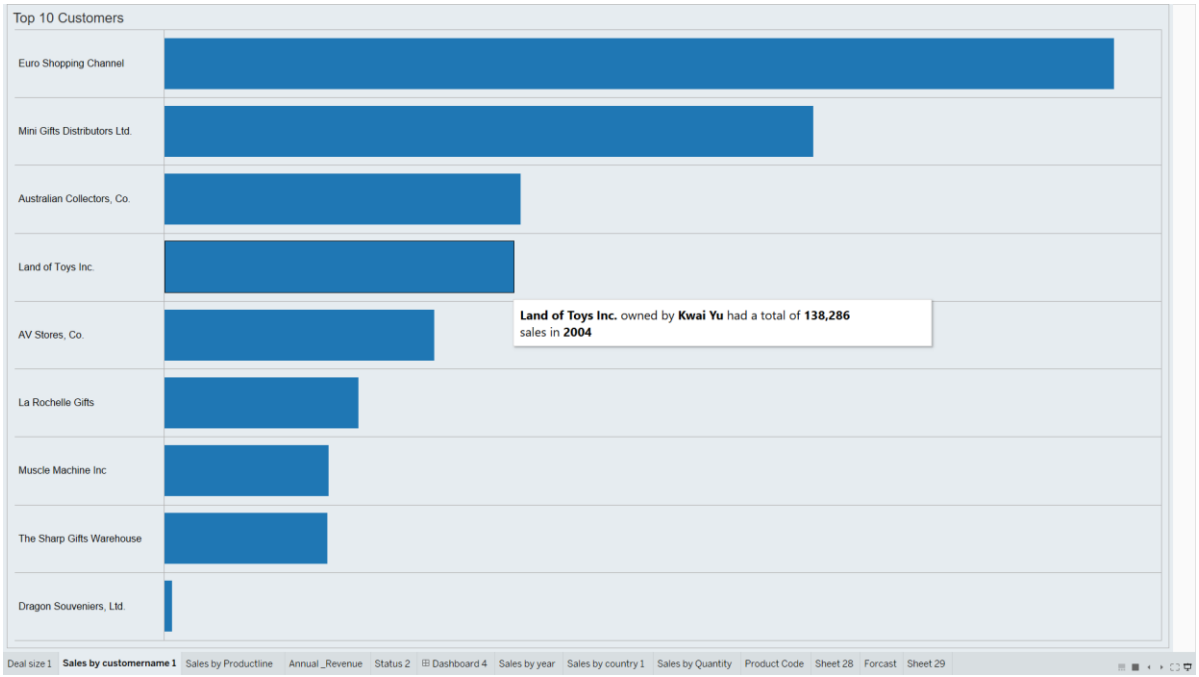
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results in a particular year.
- I used **Top** option in Filter to show top 10 names by creating a **Parameter** named top 10 full name.

iv. Analysis results :

- In all the three years **Euro Shopping Channel** owned by **Diego Freyre** and **Mini Gifts Distributors LTD.** owned by **Valarie Nelson** had **First** and **second** highest sales respectively.

v. Visualization :



3. Sales by Productline :

i. Introduction :

This chart shows us the sales of a Productline of a particular country in a particular year.

ii. General Description:

- I've dropped Productline column in color in marks card and sales in column and then I chose **Pie Chart** to show the visualization.

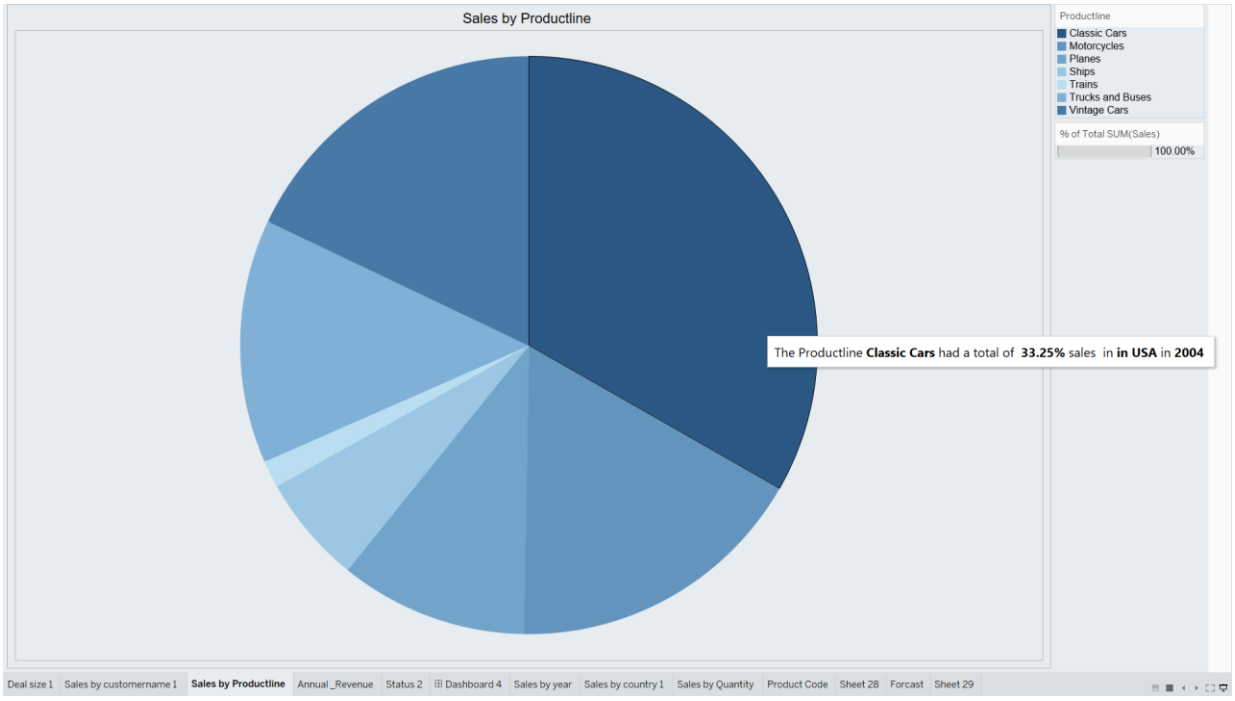
iii. Specific Requirements, functions and formulas and prediction models :

- I used **Percent of total** option in **Quick table calculation** to show sales in Percentage.
- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular country in a particular year.

iv. Analysis results :

- **Classic Cars** Productline had **highest sales** in all three years.
- **Trains** Productline had **least sales** in all years.
- Likewise with the filter option we can check a particular country's sales

v. Visualization :



4. Annual Revenue :

i. Introduction :

This graph shows the total sales in all three years.

ii. General Description :

- I've dropped Year Id in columns and sales in rows.
- I've dropped Year Id in label in marks card.

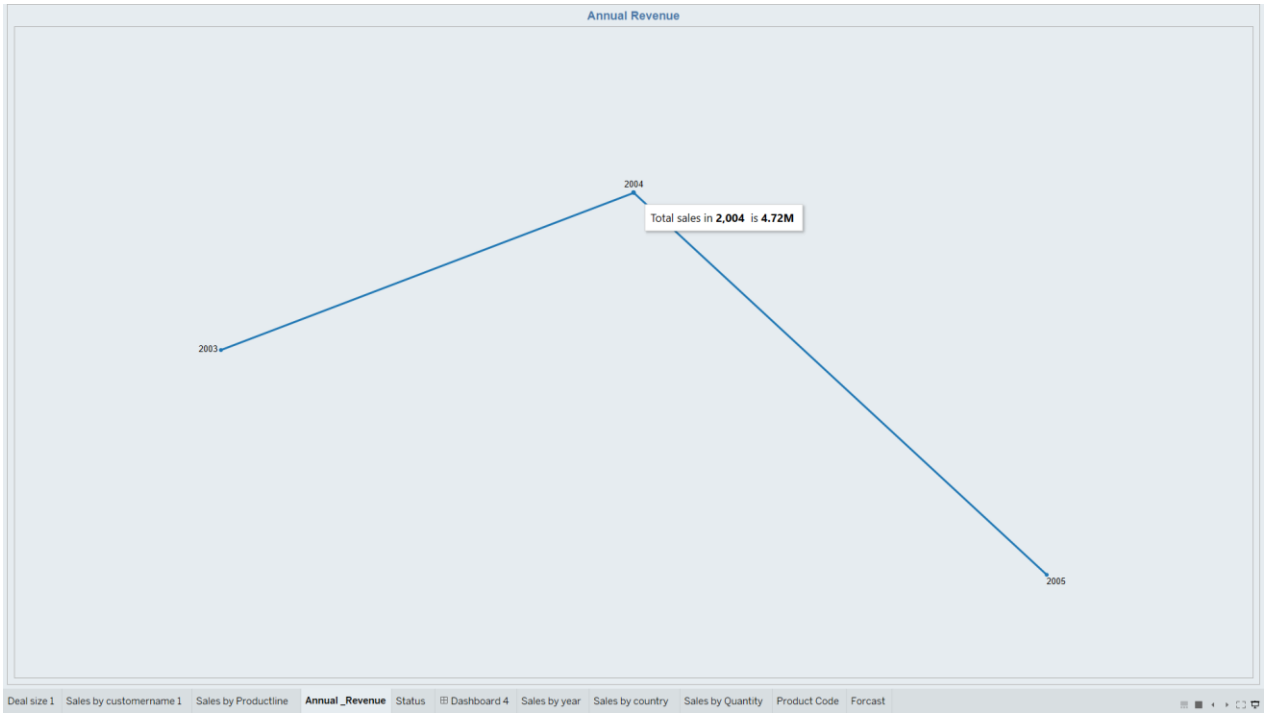
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- To show the sales in Million I have formatted sales in Number custom to show in million and up to 2 decimal points.

iv. Analysis results :

- Total sales in **2003** is **3.52 Million**.
- Total sales in **2004** is **4.72 Million** with **highest** sales in all three years.
- Total sales in **2005** is **1.79 Million** with **least** sales in all three years.

v. Visualization :



5. Status of order :

i. Introduction :

This chart shows us about the status of order , like if it got shipped, is on hold, is resolved, is in process, is disputed or cancelled.

ii. General Description :

- I've dropped status column in color in marks card.
- I've dropped the field "zero" that I made twice in rows and chose pie chart and sales in rows. And I've made a **Doughnut chart** to depict the result.

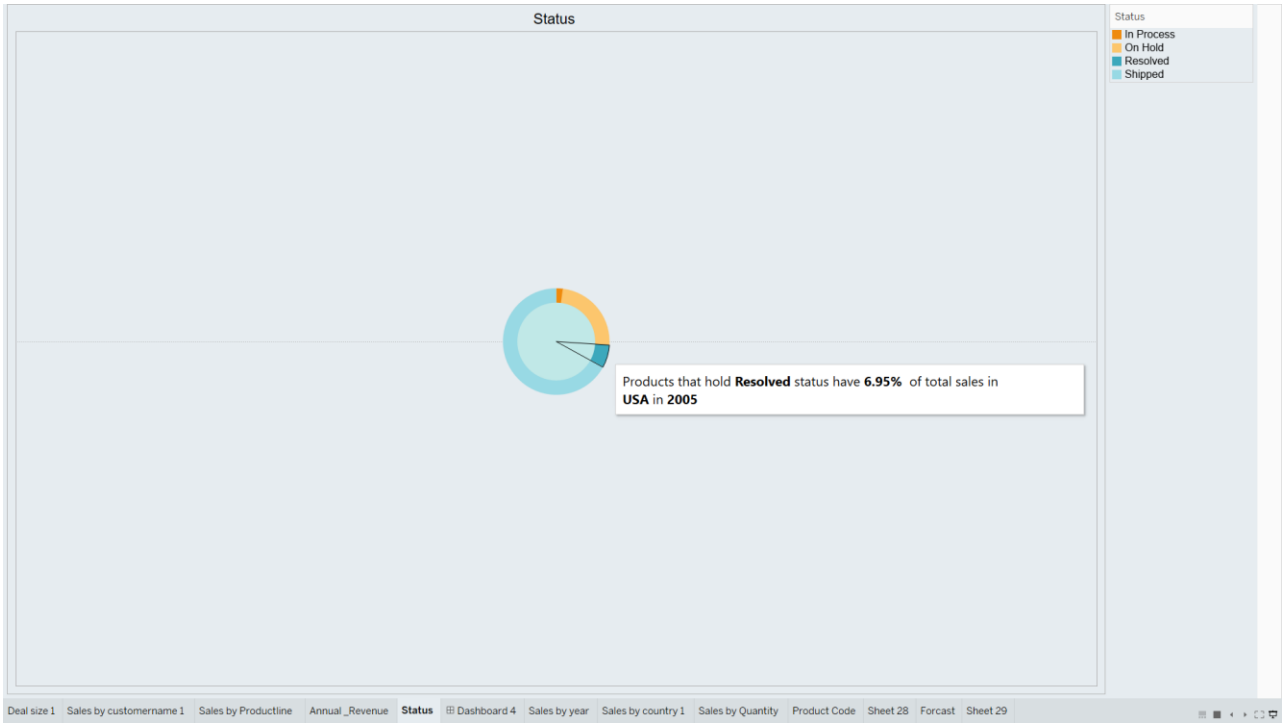
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular country in a particular year.
- I made a **Calculated field** called **Zero** which has value 0 to make Doughnut chart.

iv. Analysis results :

- Most of the orders got shipped only few are on hold, in process and resolved and few more.

v. Visualization :



6. Sales by year :

i. Introduction :

This plot shows month wise sales of a particular country in a particular year.

ii. General Description :

- I've dropped orderdate field in columns and sales and country in rows.
- I chose **Box – Whisker plot** to depict the result.

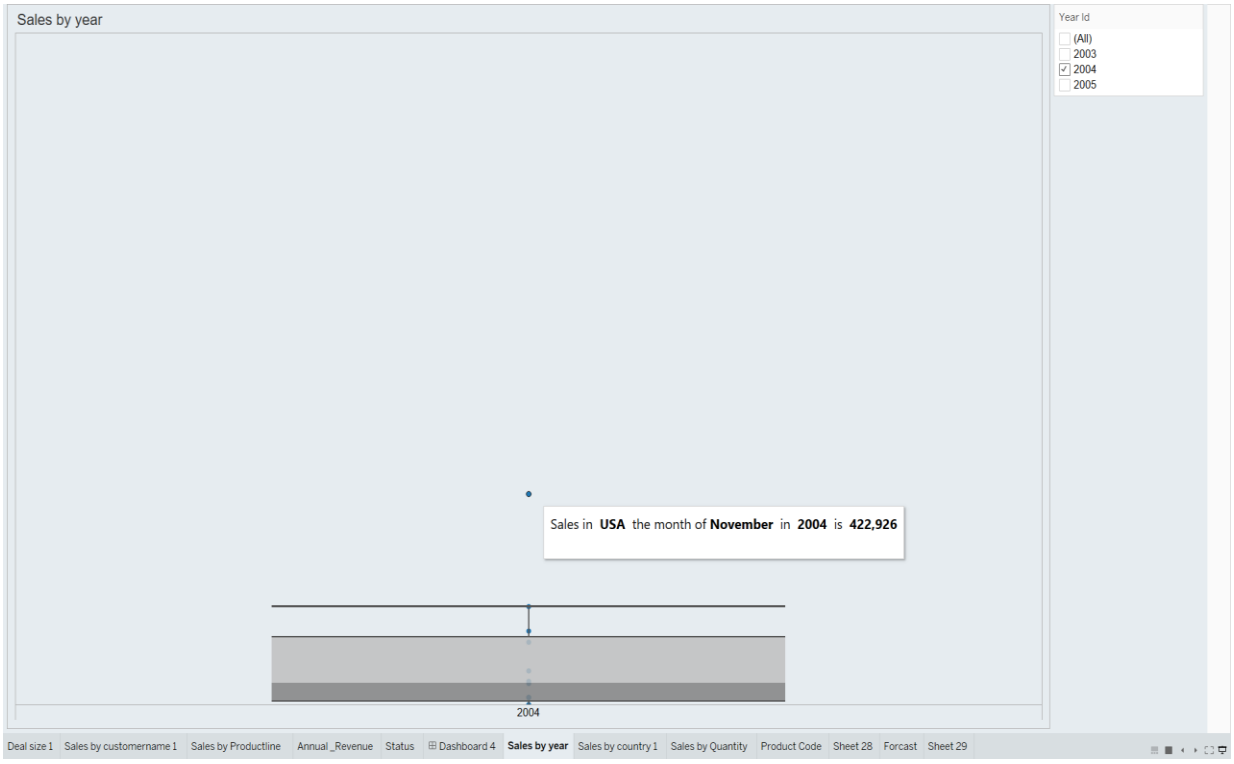
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular country in a particular year.

iv. Analysis results :

- **November** month of all three years had most sales.
- And mostly January has least sales.

v. Visualization :



7. Sales by country :

i. Introduction :

This visualization shows the sales of a country in a particular year.

ii. General Description :

- I've dropped Country field in columns and sales field in rows.
- I chose **Tree map** to show the result.

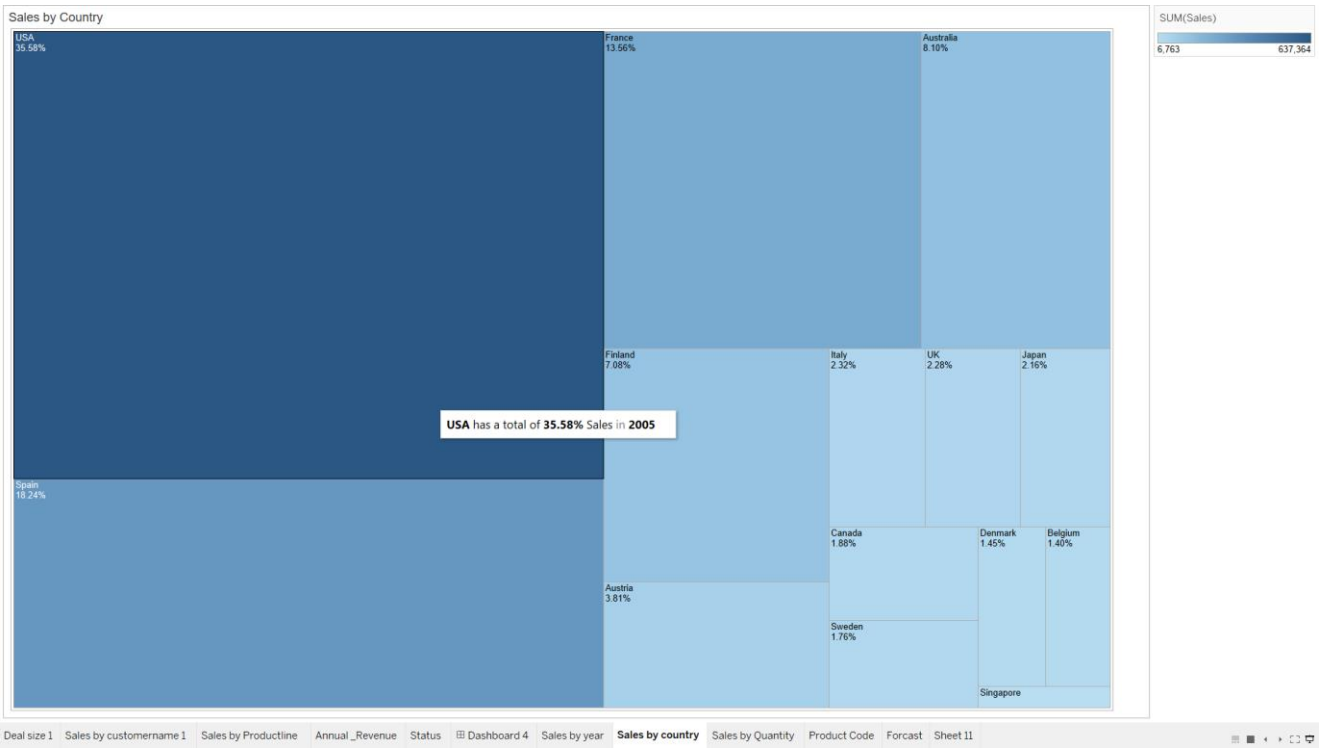
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **Percent of total** option in **Quick table calculation** to show sales in Percentage.
- I used **filters** to show results in a particular year.

iv. Analysis results :

- **USA** had most sales in all three years.
- **Spain** and **France** were fluctuating in second and third positions in all three years.

v. Visualization :



8. Sales by Quantity :

i. Introduction :

This chart shows us the count of Quantity Ordered of a particular Productline.

ii. General Description :

- I've dropped Productline column in columns and QuantityOrdered and country field in rows.
- I chose **Packed Bubbles chart** to depict the result.
- I then changed measure of QuantityOrdered from sum to count.

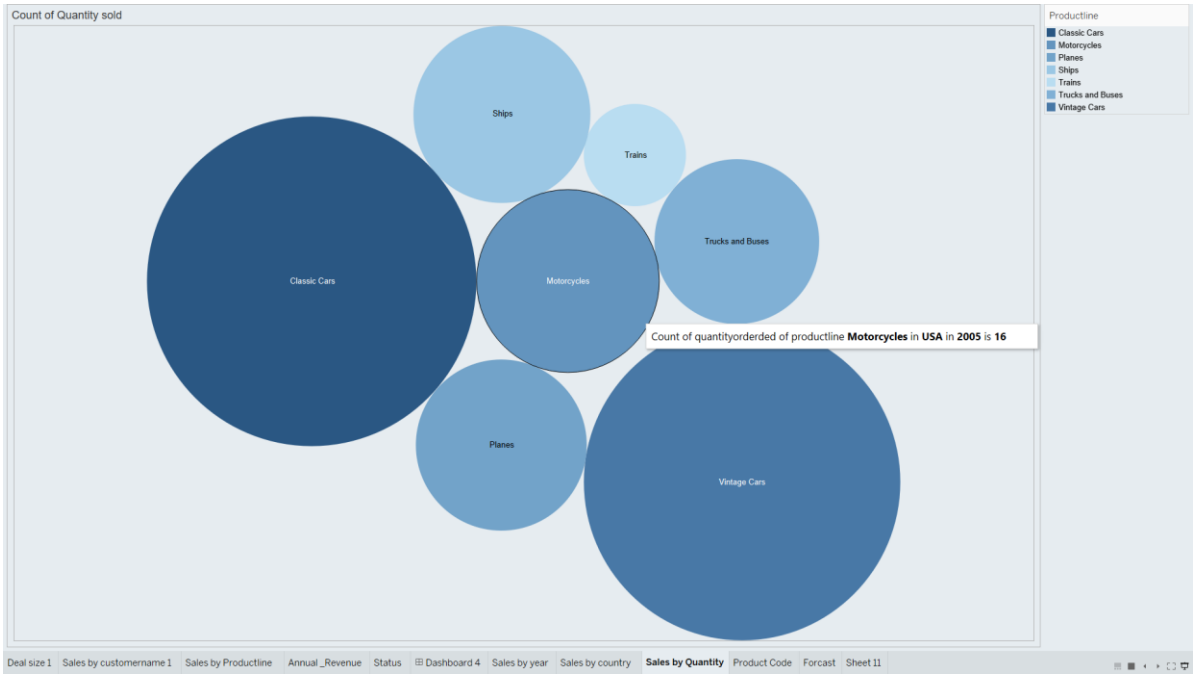
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular country in a particular year.

iv. Analysis results :

- Classic cars were the most ordered Productline.
- Trains were the least ordered Productline.
- Likewise with the filter option we can check a particular country's sales.

v. Visualization :



9. Sales by ProductCode :

i. Introduction :

This graph shows the sales of a particular Product Code of a particular Productline.

ii. General Description :

- I've dropped Productcode field and Productline field in Columns and sales and country in rows.
- I chose to show the results in **Line Graph** .

iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular Productline.

iv. Analysis results :

- The Productcode **$10_4757** of **Classic Cars** had **17,273** sales.
- The Productcode **$12_2823** of **Motorcycles** had **9829** sales.
- Likewise with the filter option we can check a particular country's sales.

v. Visualization :

10. Future sales :

i. Introduction :

This graph shows the hypothetical future sales .

ii. General Description :

- I've dropped Orderdate field in columns and changed it to continuous month.
- I've dropped sales in rows
- Then I dropped forecast and trendline from analytical pane into the sheet .

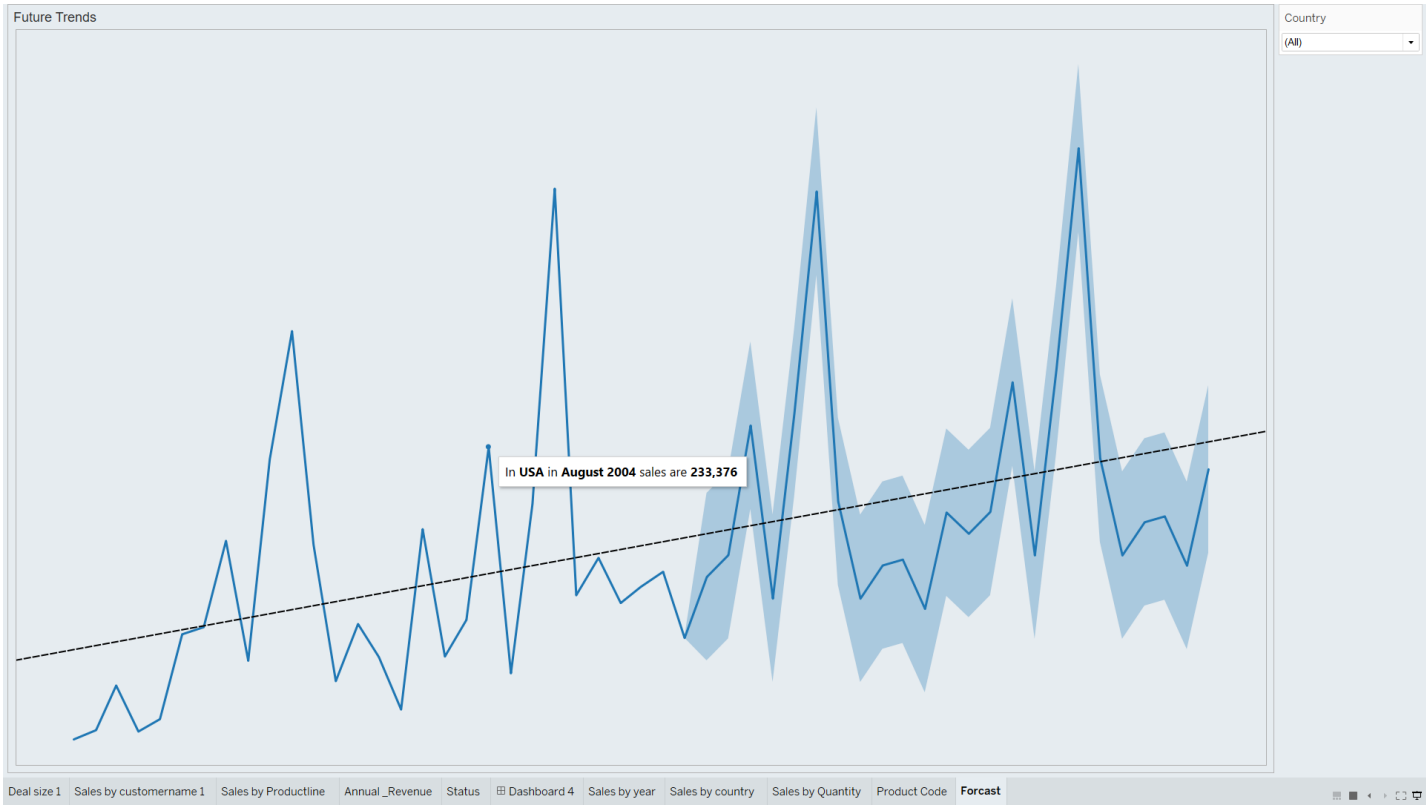
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular Country in a particular year.

iv. Analysis results :

- Since November had highest sales in all three years tableau predicted the trend and shown in the graph the same for future years that November might have highest sales.
- And December had least sales in all three years so tableau predicted the trend and shown in the graph the same for future years that December might have least sales.

v. Visualization :



11. Priceeach :

i. Introduction :

This graph shows us the count of Quantity ordered of a product that priced between ranges like 29 – 31, 31 – 33 up to 97 – 99.

ii. General Description :

- I've dropped Priceeach bin field in rows and QuantityOrdered in columns.
- I chose **Histogram** to depict the result.

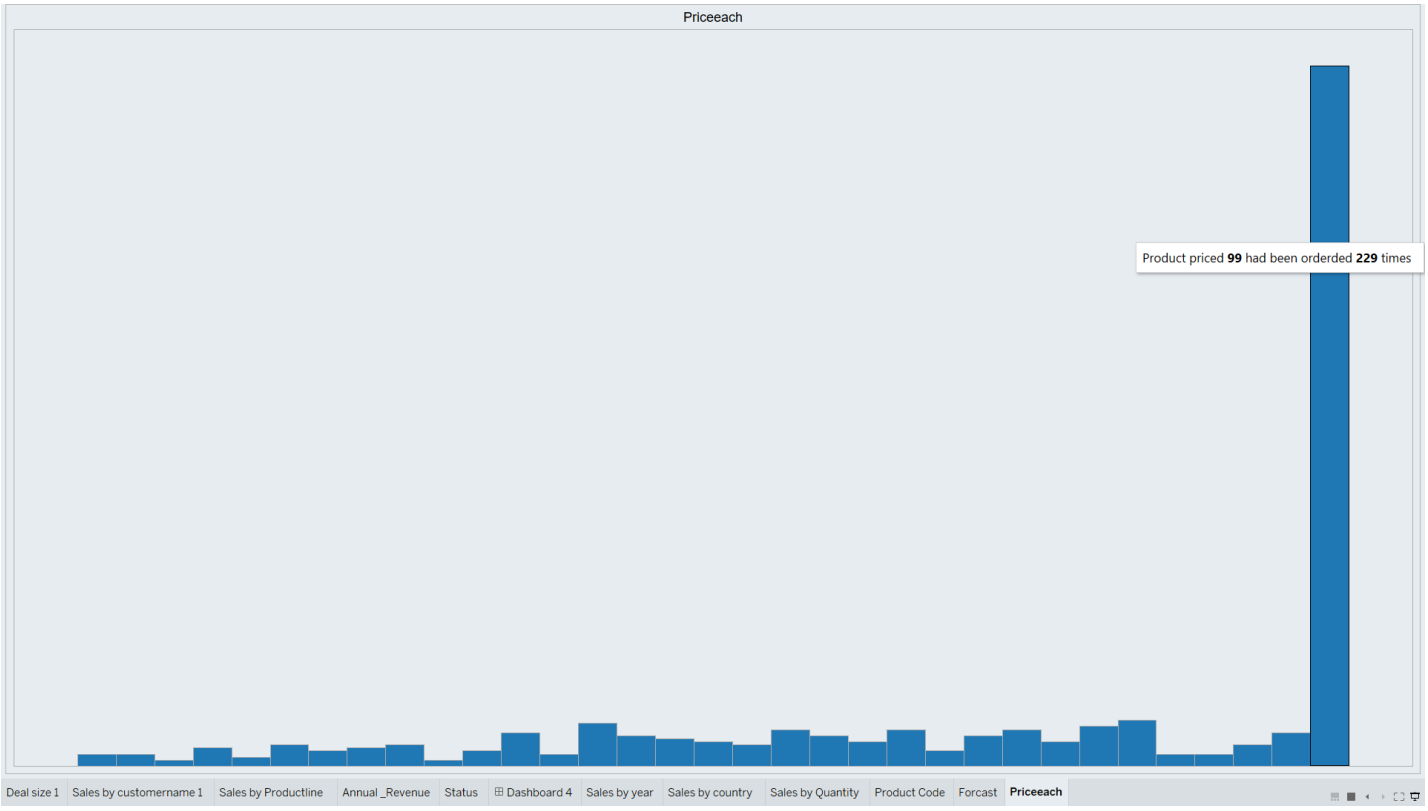
iii. Specific Requirements, functions and formulas and prediction models :

- I customized **Tool tip** to show the result in more readable way.
- I used **filters** to show results of a particular Country and a particular Productline.

iv. Analysis results :

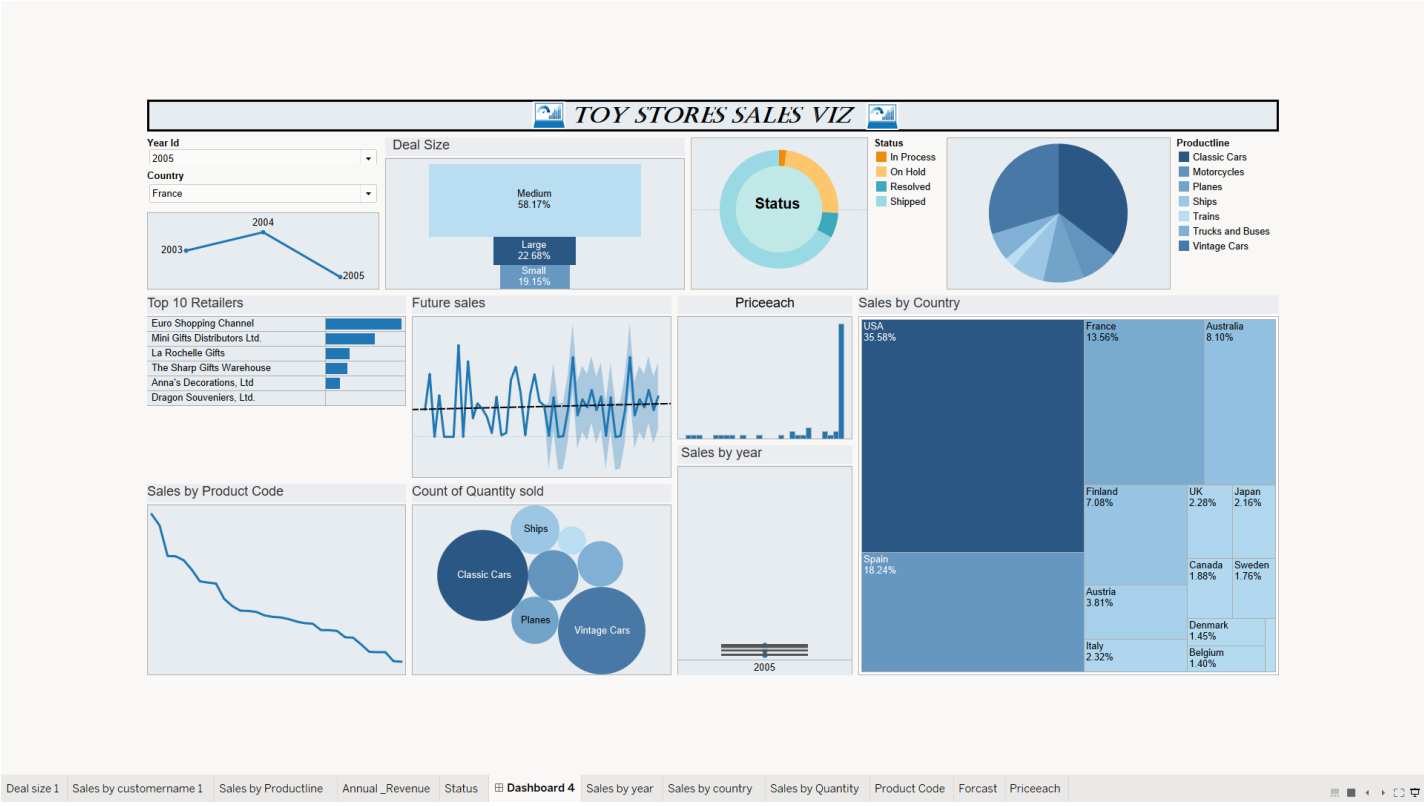
- Price ranged between 97 – 99 had been ordered most of the times.
- Rest all price ranges had been only few times.

v. Visualization :



Dashboard :

This is the final dashboard with two filters Year Id and Country.



LIST OF ANALYSIS WITH RESULTS :

S.NO	ANALYSIS	RESULT
1	Deal size	In all the three years **Medium dealsize** has the **most sales.**
2	Top 10 Retailers	In all the three years **Euro Shopping Channel** owned by **Diego Freyre** and **Mini Gifts Distributors LTD.** owned by **Valarie Nelson** had **First** and **second** highest sales respectively.
3	Sales by Productcode	The Productcode **$10_4757** of **Classic Cars** had **17,273** sales. The Productcode **$12_2823** of **Motorcycles** had **9829** sales
4	Annual Revenue	Total sales in **2003** is **3.52 Million**. Total sales in **2004** is **4.72 Million** with **highest** sales in all three years
5	Status of order	Most of the orders got shipped only few are on hold, in process and resolved and few more
6	Sales by year	**November** month of all three years had most sales. And mostly January has least sales.
7	Sales by country	**USA** had most sales in all three years. **Spain** and **France** were fluctuating in second and third positions in all three years.
8	Sales by Quantity	Classic cars were the most ordered Productline. Trains were the least ordered Productline.
9	Sales by Productline	**Classic Cars** Productline had **highest sales** in all three years. **Trains** Productline had **least sales** in all years.
10	Future sales	Since November had highest sales in all three years tableau predicted the trend and shown in the graph the same for future years that November might have highest sales.
11	Priceeach	Price ranged between 97 – 99 had been ordered most of the times.

REFERNCES :

[1] It was Originally Written by María Carina Roldán, Pentaho Community Member, BI consultant (Assert Solutions), Argentina. This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License. Modified by Gus Segura June 2014.
[2] Collaborated by Gus Segura, Director, Principal Data Science Engineering at Blueskymetrics.com Seattle, Washington, United States

BIBLIOGRPHY :

Referred following websites :

https://help.tableau.com/

https://www.analyticsvidhya.com/

https://www.tutorialspoint.com/