BA Toolkit Portfolio

Zhang Yijie

Dalhousie University

INFO6513: Business Analytics & Data Vis

Instructor: Dr. Kyung Young Lee

April 8, 2025

Table of Contents

Chapter 1 Microsoft Excel Pivot Table

1.1 About Microsoft Excel Pivot Table

Microsoft Excel's Pivot Table is a powerful data analysis tool that helps users quickly organize, summarize, and analyze large amounts of data. With simple drag-and-drop operations, users can easily adjust the arrangement of data, classify, calculate, and visualize from different dimensions, to more intuitively discover trends and patterns in the data. Its core functions are data aggregation, filtering, grouping, and interactive analysis, such as automatic calculation of sums, counts, averages, etc., while filtering data according to specific conditions to improve analysis efficiency. In addition, Pivot Table can also group data, such as by date, category, etc., to help users view information at different levels more clearly. Its powerful interactive features allow users to adjust rows, columns, values, and filters at any time, and dynamically update data views without manually modifying the original data.

1.2 Dataset Source and Research Questions

I continue and expand on what I learned in Chapter 1 by using Microsoft Excel Pivot Table to analyze Global Bike Inc.'s sales data from 2007 to 2016 and answer the following questions:

1) How does the company's annual revenue change?
2) How does the company's revenue change in different countries?
3) Which product contributes the most revenue? How has its share of total revenue changed?
4) Is seasonal in the company's revenue? How does the company's revenue vary from month to month? What are the best-selling products in the highest-revenue months?
5) In the year 2014, which was the product with the highest sales in terms of quantities

The dataset is GBI_E5_2.xlsx, which is provided in class. The dataset does not need to be cleaned and can be used directly.

1.3 Analysis tool application and results

1.3.1 How does the company's annual revenue change?

Row Labels	Sum of Revenue USD
2014	$142,498,014.18
2012	$138,086,539.24
2011	$136,951,337.89
2013	$136,186,868.71
2016	$135,151,873.09
2015	$131,866,086.75
2008	$118,763,939.67
2007	$115,993,581.83
2010	$109,622,920.05
2009	$101,326,903.66
Grand Total	**$1,266,448,065.07**

Figure 1

Inserted a PivotTable containing all the data as a new worksheet. In the PivotTable Filed, YEAR was added to Rows, Revenue USD was added to values, which is because in the original data, since the income comes from two regions, DE and USA, the exchange rate difference makes it difficult to compare. It is important to convert the exchange rate to the same currency. In this example, Revenue USD is used.

Changed the number type of Revenue USD in values to currency ($) through Value Field Setting. Now I get a crosstab, which the YEAR was sorted in descending order of revenue (Figure 1). To compare the revenue variation more easily, inserted a clustered column chart with different years on the x-axis and income on the y-axis (Figure 2)

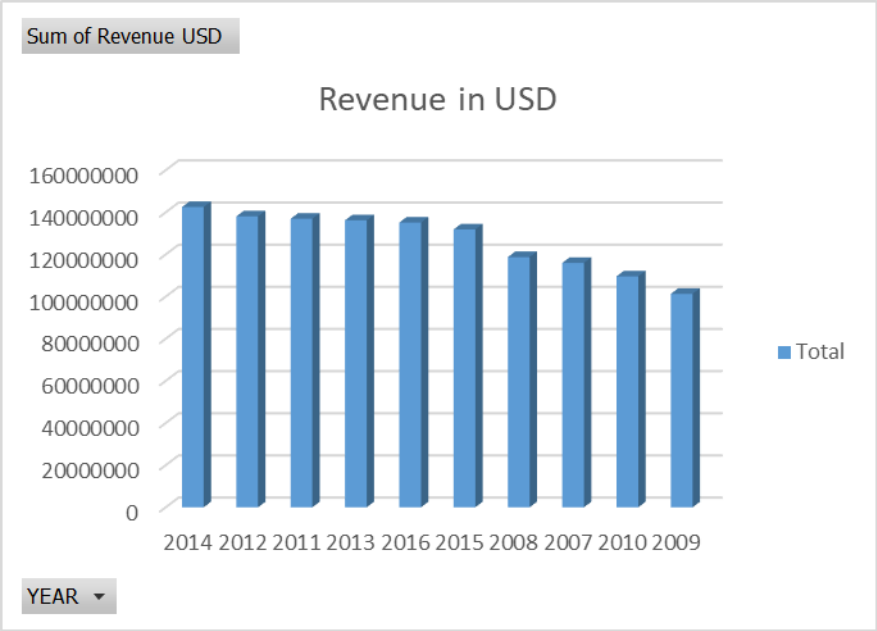


Figure 2

It can be clearly seen from the picture that the company achieved the highest revenue in 2014 and the lowest in 2009. Although this chart can easily compare the differences in revenue between different years, since the years are sorted in descending order based on revenue, it becomes difficult to analyze the trend of the company's revenue changes between 2007 and 2016. By removing the descending order, a new chart was inserted (Figure 3).

Figure 3

Now we can clearly see that from 2007 to 2016, the company's revenue generally showed an upward trend, with a decline from 2007 to 2009, followed by a rapid increase, and remained basically stable from 2011 to 2016.

1.3.2 How does the company's revenue change in different countries?

In the PivotTable Filed, YEAR was added to Rows, Country was added to Columns, Revenue USD was added to values and got a new crosstab (Figure 4).

Sum of Revenue USD	Column Labels		
Row Labels	DE	US	Grand Total
2007	$59,838,733.83	$56,154,848.00	$115,993,581.83
2008	$62,563,714.02	$56,200,225.65	$118,763,939.67
2009	$61,715,877.42	$39,611,026.24	$101,326,903.66
2010	$63,449,297.26	$46,173,622.79	$109,622,920.05
2011	$76,013,814.81	$60,937,523.08	$136,951,337.89
2012	$71,066,139.42	$67,020,399.82	$138,086,539.24
2013	$73,177,517.38	$63,009,351.33	$136,186,868.71
2014	$79,135,301.12	$63,362,713.06	$142,498,014.18
2015	$74,698,783.11	$57,167,303.64	$131,866,086.75
2016	$79,118,833.50	$56,033,039.59	$135,151,873.09
Grand Total	$700,778,011.87	$565,670,053.20	$1,266,448,065.07

Figure 4

Then inserted the Stacked Line with Markers chart. Added Data labels on the chart.

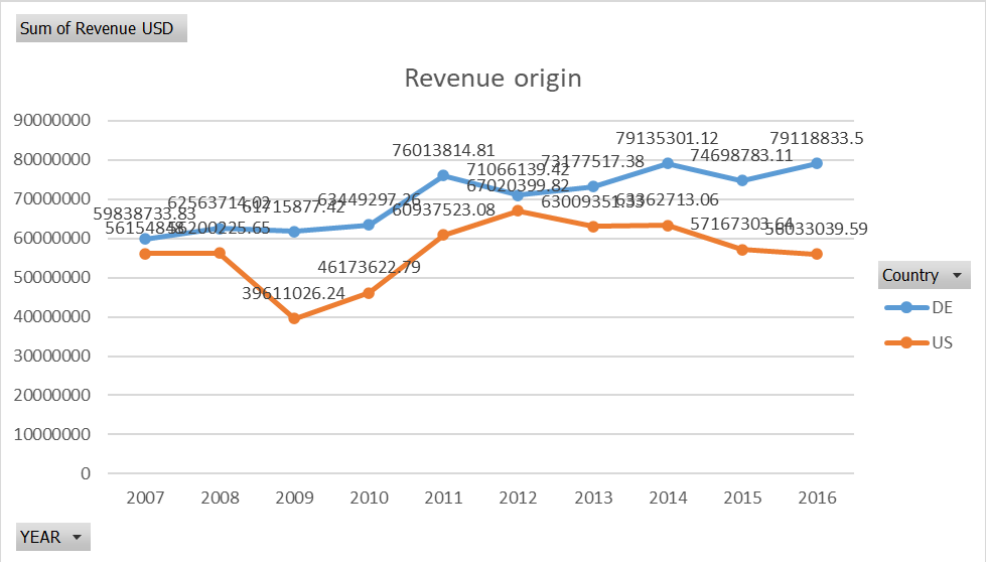


Figure 5

This chart (Figure 5) shows the changes in the company's revenue in the USA and DE from 2007 to 2016. Comparing the revenue from DE and the USA, the revenue from DE is always higher than that from the USA; and comparing the trends of revenue changes, the revenue from DE had increased significantly, while the revenue from the USA had not increased.

1.3.3 Which product contributes the most revenue? How has its share of total revenue changed?

Added Product to Rows. Added Revenue USD to values. Clicked drop row and used Field Settings. Choose ‘Show value as % of Grand Total’. Then a new crosstab was

inserted (Figure 6). Figure 6 shows the proportion of total revenue generated by each product.

Row Labels	Sum of Revenue USD
BOTL1000	0.03%
CAGE1000	0.08%
CITY1000	0.41%
DXRD1000	7.64%
DXRD2000	2.46%
DXTR1000	6.12%
DXTR2000	11.50%
DXTR3000	1.83%
ELBK1000	3.17%
EPAD1000	0.05%
FAID1000	0.12%
FXGR1000	0.19%
HVBD1000	0.07%
KPAD1000	0.06%
OHMT1000	0.08%
ORHT1000	4.85%
ORHT2000	7.62%
ORMN1000	9.42%
ORWN1000	4.84%
PRRD1000	11.97%
PRRD2000	3.36%
PRRD3000	3.38%
PRTR1000	6.13%
PRTR2000	12.29%
PRTR3000	1.97%
PUMP1000	0.21%
RHMT1000	0.07%
RKIT1000	0.05%
SHRT1000	0.04%
Grand Total	**100.00%**

Figure 6

Row Labels	Sum of Revenue USD
PRTR2000	12.29%
PRRD1000	11.97%
DXTR2000	11.50%
ORMN1000	9.42%
DXRD1000	7.64%
ORHT2000	7.62%
PRTR1000	6.13%
DXTR1000	6.12%
ORHT1000	4.85%
ORWN1000	4.84%
PRRD3000	3.38%
PRRD2000	3.36%
ELBK1000	3.17%
DXRD2000	2.46%
PRTR3000	1.97%
DXTR3000	1.83%
CITY1000	0.41%
PUMP1000	0.21%
FXGR1000	0.19%
FAID1000	0.12%
CAGE1000	0.08%
OHMT1000	0.08%
HVBD1000	0.07%
RHMT1000	0.07%
KPAD1000	0.06%
EPAD1000	0.05%
RKIT1000	0.05%
SHRT1000	0.04%
BOTL1000	0.03%
Grand Total	**100.00%**

Figure 7

Sorted Product according to the descending order of revenue (Figure 7). It is easy to see from Figure 7 that most of the company's revenue is contributed by a few major products.

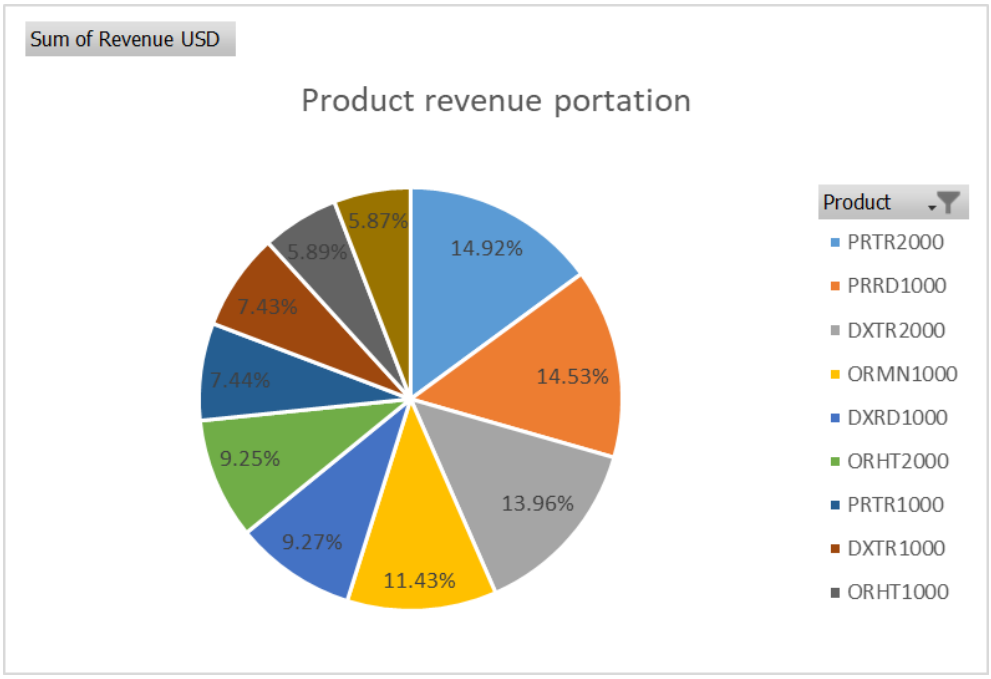


Figure 8

Inserted the pie charts with Data Labels. Right clicked the Product Rows. Used Top 10 Filter to show top 10 items by revenue.
As can be seen from Figure 8, the four products PRTR 2000, PRRD 1000, DXTR 2000 and OMRN 1000 contribute more than half of the company's revenue, while none of the remaining six products accounted for more than 10% of total revenue.

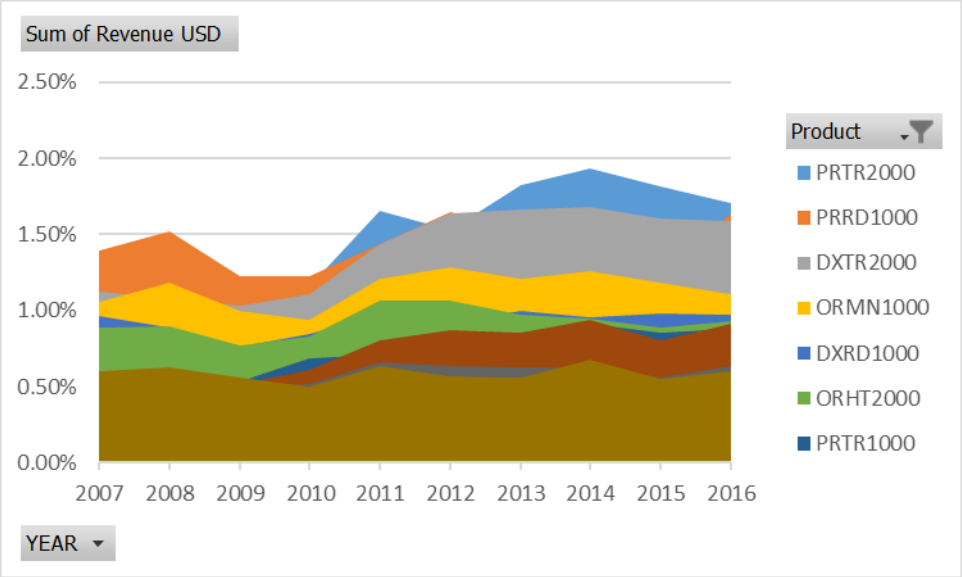


Figure 9

Add YEAR to Rows. Add Products to columns. Add Revenue USD to values. Inserted Area chart. Then Figure 9 was obtained.
As shown in Figure 9, PRRD 1000 was the product that contributed the most to the company's revenue from 2007 to 2009 but was surpassed by PRTR 2000 in 2010. In 2012, the revenue contributions of PRTR 2000, PRRD 1000, and DXTR 2000 were very close. By 2016, PRTR 2000 had been the product that contributed the most to the revenue for four consecutive years, followed by PRRD 1000 and DXTR 2000.

1.3.4 Is seasonal in the company's revenue? How does the company's revenue vary from month to month? What are the best-selling products in the highest-revenue months?
Add MONTH to Rows. Add Products to columns. Add Revenue USD to values. Used Value filter to show top 10 items by revenue. Inserted Stacked column chart.

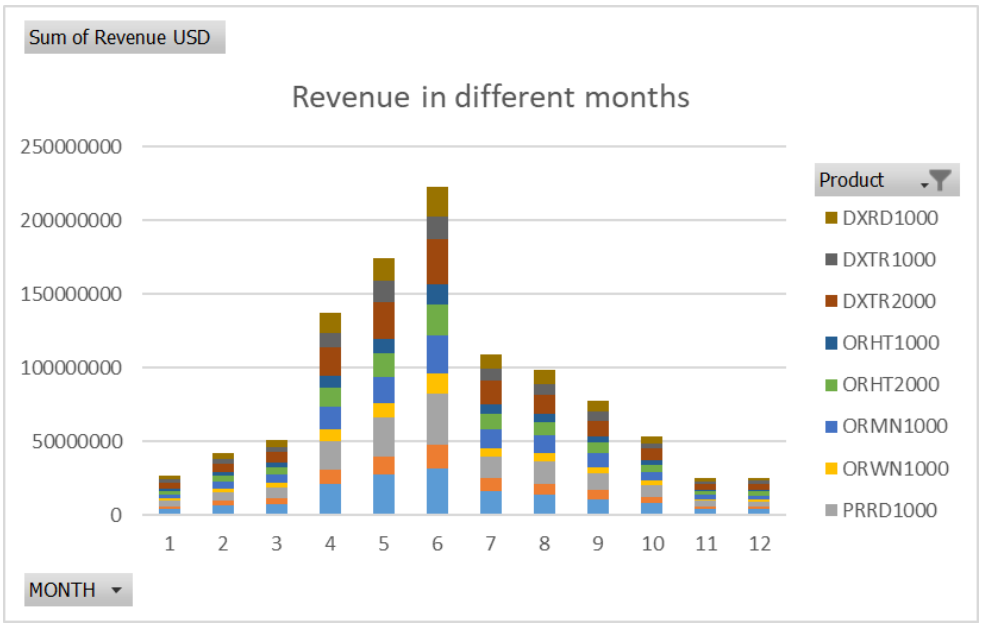


Figure 10

As shown in Figure 10, the company's revenue shows strong seasonal characteristics. Revenue is mainly concentrated in the middle of the year, rising rapidly in April, reaching a peak in June, and then falling rapidly, shrinking month by month in July, August, September, and October, and remaining stable at a low level in November, December, and January. Among them, PRDR 1000 was the product that contributed the most to revenue in June.

1.3.5 Who is the company's largest customer and who is its smallest customer?
Add YEAR to Rows. Add CustDescr to columns. Add Revenue USD to values. Inserted a line chart. Sorted top 10 companies by revenue.

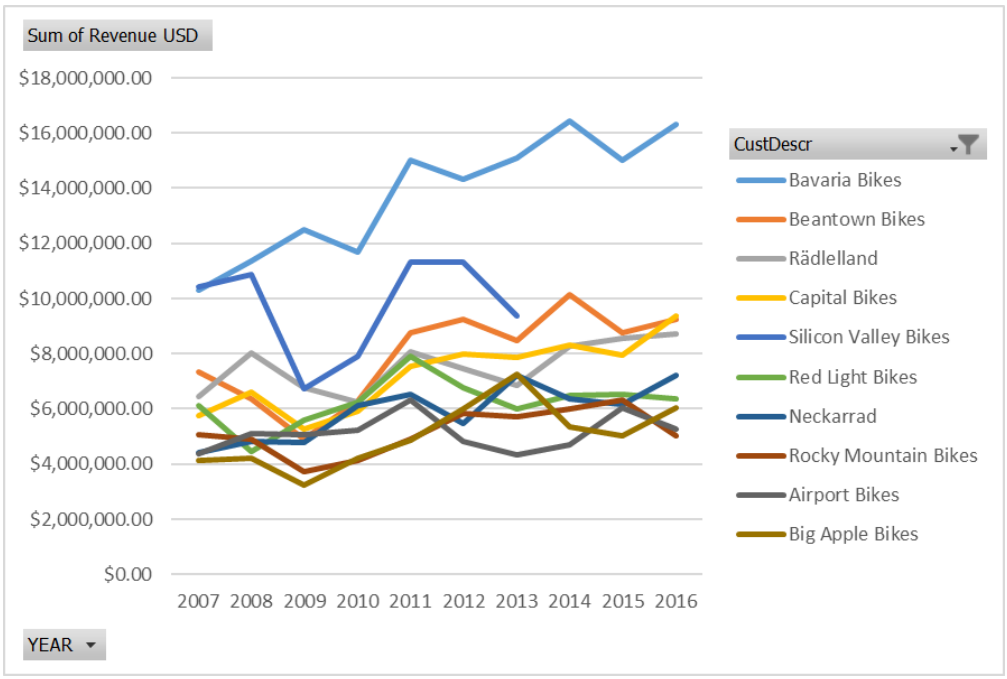


Figure 11

As shown in Figure 11, Bavaria Bikes has been the company's main customer since 2007, and the revenue obtained from the company has generally increased. Neckarrad was the company's second largest customer before 2014 but has not purchased the company's products since 2014.
Sorted bottom 10 companies by revenue.

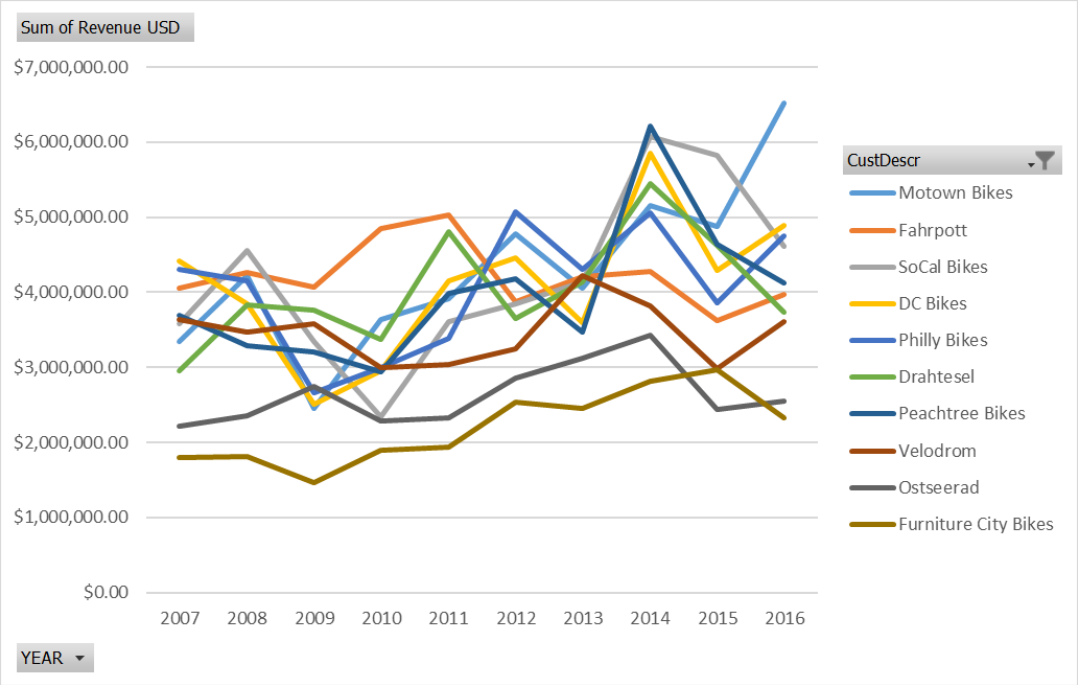


Figure 12

As can be seen from Table 12, Furniture City Bikes was the company's least profitable customer from 2007 to 2016, except for 2015, when Ostseerad was the least profitable customer.

1.4 Personal Reflection and Conclusion

After analyzing the sales data for Global Bike Inc., I really saw how powerful Excel PivotTables can be for business decision making. When looking at the annual revenue trends, I noticed that while the company hit a low point in 2009, overall, the numbers have been rising, especially since 2011. This suggests that businesses need solid strategies to deal with market fluctuations, such as optimizing costs or exploring new markets. When I compared revenues in different countries, it became clear that the German market not only has a strong revenue base, but is also growing rapidly, while the US market seems to be rather stagnant. This means that the company should consider investing more in the German market while also identifying factors that are hindering growth in the US, perhaps more intense competition or changing customer preferences.

I was also impressed by the flexibility and ease of use of PivotTables. With just a few drag-and-drop operations, I can quickly create different analytical views, such as looking at how product revenue shares change over time. For example, PRTR 2000 is becoming a key product, while PRRD 1000's share is constantly declining, indicating

that a product refresh or better marketing may be needed. Seasonality analysis also showed that June was a peak sales month, with PRDR 1000 being the best-selling product – which is great for inventory management and promotions. That being said, I did run into some downsides. The tool would sometimes freeze when processing large data sets, and for more advanced analysis, I still had to rely on formulas or manual work, which slowed things down a bit. If Excel could increase the speed of data processing and add more automation, it would be more useful.

From a business perspective, PivotTables make data analysis more accessible, even for people without a technical background, because they can quickly generate insights that really drive decision making. But there are limitations – the tool works best with structured data, struggles with unorganized or real-time data, and isn't suited for complex analysis, which often requires other tools. Overall, Excel PivotTables are great for small and medium-sized businesses that want to make data-driven decisions, but for more complex situations, a specialized platform or custom solution may be needed. This experience really made me realize that the value of a tool isn't just in its functionality, but how well you use it to align with business goals and turn insights into action.

Chapter 2 FAOSTAT

2.1 About FAOSTAT

FAOSTAT is a global authoritative statistical database developed by the Food and Agriculture Organization of the United Nations (FAO), focusing on agriculture, food security, nutrition, fisheries, forestry and related fields. It covers statistical data from 245 countries and regions in the world since 1961, including key indicators such as crop yields, trade, land use, and greenhouse gas emissions. Its data integrates official statistics, international reports and FAO model estimates from various countries to support academic research, policy making (such as food security strategies, monitoring of sustainable development goals), market analysis and international cooperation (such as disaster response). Users can access it for free through the official website, filter data by country, year or indicator, download it in Excel, CSV and other formats, or integrate it into applications using API interfaces. FAOSTAT provides a reliable basis for global agricultural and food system analysis for governments, businesses and research institutions. It is often used in conjunction with other FAO reports (such as The State of Food Security and Nutrition in the World) and is a core tool for understanding the dynamics of related fields.

2.2 Dataset Source and Research Questions

In this chapter, I will use FAOSTAT W2025 to carry out the SDG Data Challenge. I will mainly clean the obtained dataset and then use Tableau to make meaningful visualizations to answer the following questions:

1) How is the economic development of Asian countries during the epidemic?
2) How do CO2 emissions from food packaging differ in different countries?
3) How will the amount of CO2 produced by industrial wastewater discharge differ in different countries in 2022?
4) Is there a correlation between water use efficiency and water stress levels in different countries?

In addition, all the data sets used in this chapter are from the FAOSTAT database (https://www.fao.org/faostat/en/#data), and the data sets need to be cleaned.

2.3 Analysis tool application and results

2.3.1 How is the economic development of Asian countries during the epidemic?

Domain	Domain	Area Co	Area	Element	Element	Item Coc	Item	Year Co	Year	Unit	Value	Flag	Flag Des	Note
FS	Suite of Fo	050	Banglades	6121	Value	21010	Average di	20182020	2018-2020	%	110.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6121	Value	21010	Average di	20192021	2019-2021	%	110.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6121	Value	21010	Average di	20202022	2020-2022	%	111.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6121	Value	21010	Average di	20212023	2021-2023	%	112.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	220001	Dietary ene	2019	2019	kcal/cap/d	2514.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	220001	Dietary ene	2020	2020	kcal/cap/d	2571.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	220001	Dietary ene	2021	2021	kcal/cap/d	2568.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	220001	Dietary ene	2022	2022	kcal/cap/d	2583.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	220001	Dietary ene	2023	2023	kcal/cap/d	2591.0	E		Estimated Imputed val
FS	Suite of Fo	050	Banglades	6128	Value	22000	Dietary ene	20182020	2018-2020	kcal/cap/d	2550.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	22000	Dietary ene	20192021	2019-2021	kcal/cap/d	2551.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	22000	Dietary ene	20202022	2020-2022	kcal/cap/d	2574.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6128	Value	22000	Dietary ene	20212023	2021-2023	kcal/cap/d	2581.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6121	Value	21012	Share of di	20182020	2018-2020	%	77.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6121	Value	21012	Share of di	20192021	2019-2021	%	75.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6121	Value	21012	Share of di	20202022	2020-2022	%	74.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6123	Value	21013	Average pr	20182020	2018-2020	g/cap/d	63.5	E		Estimated value
FS	Suite of Fo	050	Banglades	6123	Value	21013	Average pr	20192021	2019-2021	g/cap/d	64.7	E		Estimated value
FS	Suite of Fo	050	Banglades	6123	Value	21013	Average pr	20202022	2020-2022	g/cap/d	66.0	E		Estimated value
FS	Suite of Fo	050	Banglades	6123	Value	21014	Average su	20182020	2018-2020	g/cap/d	13.1	E		Estimated value

Figure 13

Selected 10 countries in Asia, downloaded their Suite of Food Security Indicators dataset from FAOSTAT in XLS format. Opened XLS file downloaded, copied original data sets to create a new sheet and named as Sheet2. Applied filters to labels (Figure 13).

Cleaned the data and deleted unnecessary labels, such as: Domain Code, Domain, Area Code (M49), Element Code, Element, Item Code, Unit, Flag, Flag Description, Note. Applied left function and right function in Year columns to get new data (Figure 14).

D2 : =(LEFT(C2,4)+RIGHT(C2,4))/2

	A	B	C	D	E
1	Area	Item	Year Code	Year	Value
2	Bangladesh	Average dietary ene	20182020	2019	110.0
3	Bangladesh	Average dietary ene	20192021	2020	110.0
4	Bangladesh	Average dietary ene	20202022	2021	111.0
5	Bangladesh	Average dietary ene	20212023	2022	112.0
6	Bangladesh	Dietary energy supp	2019	2019	2514.0
7	Bangladesh	Dietary energy supp	2020	2020	2571.0
8	Bangladesh	Dietary energy supp	2021	2021	2568.0

Figure 14

Copied Value column and pasted it in a new column named New_Value. Filtered New_Value column, delete all rows with empty data. Sorted New_Value column from smallest to largest. Use the right function to change all abnormal data (Figure 15).

F1539 : =RIGHT(E1539,3)/2

	A	B	D	E	F
1531	China	Gross domestic product per capita, PPP,	2021	20669.9	20669.9
1532	Thailand	Gross domestic product per capita, PPP,	2022	20751.7	20751.7
1533	China	Gross domestic product per capita, PPP,	2022	21262.3	21262.3
1534	Thailand	Gross domestic product per capita, PPP,	2019	21331.9	21331.9
1535	Republic of Kc	Gross domestic product per capita, PPP,	2020	46506.9	46506.9
1536	Republic of Kc	Gross domestic product per capita, PPP,	2019	46903.8	46903.8
1537	Republic of Kc	Gross domestic product per capita, PPP,	2021	48594.6	48594.6
1538	Republic of Kc	Gross domestic product per capita, PPP,	2022	49977.0	49977.0
1539	Lao People's I	Number of children under 5 years of age v	2019	<0.1	0.05
1540	Lao People's I	Number of children under 5 years of age v	2020	<0.1	0.05
1541	Lao People's I	Number of children under 5 years of age v	2021	<0.1	0.05
1542	Lao People's I	Number of children under 5 years of age v	2022	<0.1	0.05
1543	Lao People's I	Number of newborns with low birthweight	2019	<0.1	0.05
1544	Lao People's I	Number of newborns with low birthweight	2020	<0.1	0.05
1545	Mongolia	Number of moderately or severely food in:	2019	<0.1	0.05
1546	Mongolia	Number of moderately or severely food in:	2020	<0.1	0.05

Figure 15

	Item	Average di	Average di	Average fa	Average pr	Average su	Cereal imp	Coefficient
Banglades	2019	2305	110	34.9	63.5	13.1	12.1	0.26
Banglades	2020	2309	110	37.3	64.7	14.4	13.7	0.26
Banglades	2021	2312	111	39.8	66	15.9	15.7	0.26
Banglades	2022	2314	112					0.26
Banglades	2023	2316						0.26
China	2019	2436	136	90.5	120.9	47.7	4.7	
China	2020	2440	137	92.3	122.9	48.9	6.9	
China	2021	2447	138	93.7	125.2	50.2	8.8	
China	2022	2453	139					
China	2023	2457						
Indonesia	2019	2308	126	69.4	74.5	27.3	14.6	0.28
Indonesia	2020	2311	126	75.4	76.4	28.2	15.3	0.28
Indonesia	2021	2314	124	78.1	77.5	28.5	14.1	0.28
Indonesia	2022	2316	123					0.28
Indonesia	2023	2319						0.28
Lao People	2019	2351	116	41.7	78.9	22.6	-2	0.23
Lao People	2020	2354	116	42.9	79.3	23	-1	0.23
Lao People	2021	2358	115	42.7	79.8	23.5	-0.9	0.22

Figure 16

Inserted a Pivot Table in a new sheet. Added Area to columns. Added Item to Rows. Add New_Value to values. Transposed the whole Table as values in a new sheet. Deleted Sum of New_Value column and all 'Total' rows (Figure 16).

Highlighted all empty value through Conditional Formatting (Figure 17) and cleaned empty value (Figure 18).

	Item	Average di	Average di	Average fa	Average pr	Average su	Cereal imp	Coefficient
Banglades	2019	2305	110	34.9	63.5	13.1	12.1	0.26
Banglades	2020	2309	110	37.3	64.7	14.4	13.7	0.26
Banglades	2021	2312	111	39.8	66	15.9	15.7	0.26
Banglades	2022	2314	112					0.26
Banglades	2023	2316						0.26
China	2019	2436	136	90.5	120.9	47.7	4.7	
China	2020	2440	137	92.3	122.9	48.9	6.9	
China	2021	2447	138	93.7	125.2	50.2	8.8	
China	2022	2453	139					
China	2023	2457						
Indonesia	2019	2308	126	69.4	74.5	27.3	14.6	0.28
Indonesia	2020	2311	126	75.4	76.4	28.2	15.3	0.28
Indonesia	2021	2314	124	78.1	77.5	28.5	14.1	0.28
Indonesia	2022	2316	123					0.28
Indonesia	2023	2319						0.28
Lao People	2019	2351	116	41.7	78.9	22.6	-2	0.23
Lao People	2020	2354	116	42.9	79.3	23	-1	0.23

Figure 17

Area	Item	////[[(!!!!!!!!!!!!!!!!!!!!Prevalence (
Banglades	2019	######04###5#2#######4####
Banglades	2020	######04###5#2#######5####
Banglades	2021	######04###6#2#######5####
Banglades	2022	######04###6#2#######5####
Indonesia	2019	######3766########66690106
Indonesia	2020	######3755########555#0006
Indonesia	2021	######3755########555#0007
Indonesia	2022	######2755########555#0007
Lao People	2019	######001120#3###1###7#797
Lao People	2020	######001120#4###1###7#786
Lao People	2021	######001130#4###1###88776
Lao People	2022	######001130#4###1###86665
Mongolia	2019	######000001##7##1777#0005
Mongolia	2020	######000001##7##1666#0004

Figure 18

Saved XLS file and opened it in Tableau. Right clicked the Area label and set it to Country. Set Item to Date type. Dragged Year to the Columns Shelf. Dragged 'Gross domestic product per capita, PPP, (constant 2017 international $)' to the Rows Shelf. Dragged Area to Color. In the Marks Card, selected Line as the chart type. Clicked Label and checked Show Mark Labels (Figure 19).

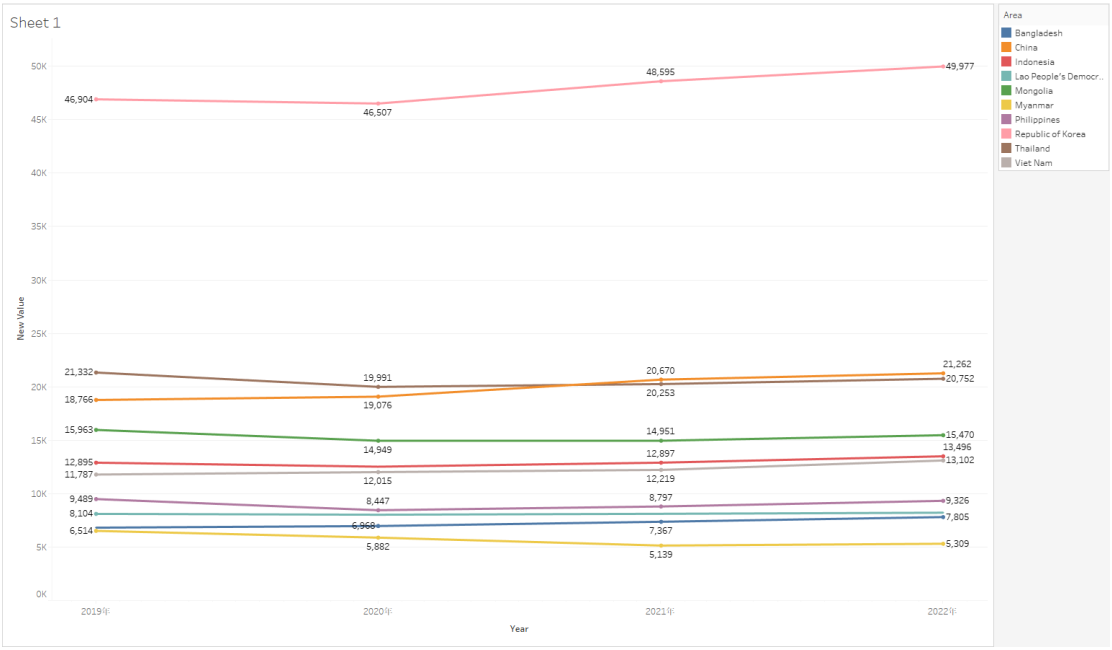
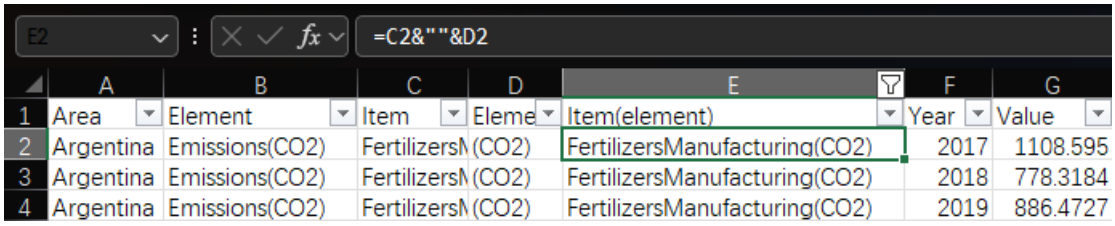


Figure 19

From the Figure 19, South Korea (assuming the pink line) is far ahead in per capita PPP, increasing from 46,904$ in 2019 to 49,977$ in 2022, showing an overall upward trend. Although it fell slightly in 2020, it quickly rebounded, showing strong economic stability and growth potential. China (orange line) also saw a steady increase in per capita PPP, from 18,766$ in 2019 to 21,262$ in 2022, indicating continued economic growth. Thailand (brown line) experienced a certain decline in 2020, from 21,332$ in

2019 to 19,991$, but then recovered to 20,752$, which may have been affected by the epidemic. Other Southeast Asian countries, such as Indonesia, Vietnam, the Philippines, and Bangladesh, have a per capita PPP of mostly between 5,000$ and 15,000$. Some of these countries saw a decline in 2020 but rebounded in 2021-2022. Overall, the economies of East Asia and Southeast Asia are still growing, but the growth rates of different countries vary, greatly affected by industrial development and the epidemic.

2.3.2 How do CO2 emissions from food packaging differ in different countries?
Selected 10 countries, downloaded their Emissions from pre and post agricultural production dataset from FAOSTAT in XLS format. Opened the file, deleted some unnecessary columns. Integrated Item and Element column to a new column named Item(element) (Figure 20).



E2 =C2&""&D2

	A	B	C	D	E	F	G
1	Area	Element	Item	Eleme	Item(element)	Year	Value
2	Argentina	Emissions(CO2)	FertilizersN	(CO2)	FertilizersManufacturing(CO2)	2017	1108.595
3	Argentina	Emissions(CO2)	FertilizersN	(CO2)	FertilizersManufacturing(CO2)	2018	778.3184
4	Argentina	Emissions(CO2)	FertilizersN	(CO2)	FertilizersManufacturing(CO2)	2019	886.4727

Figure 20

Inserted a Pivot Table in a new sheet. Added Area, Year to columns. Added Item(element) to Rows. Add Value to values. Transposed the whole Table as values in a new sheet. Deleted some 'Total' rows. Highlighted all empty value through Conditional Formatting (Figure 21) and cleaned empty value (Figure 22).

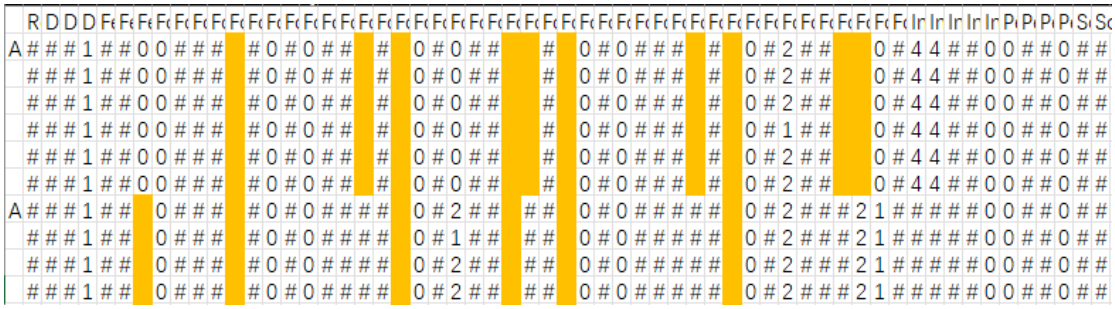


Figure 21

Country	Date	D	F	F	F	F	Fc	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	In	In	P	P	S	S
Argentina	2017	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Argentina	2018	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Argentina	2019	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Argentina	2020	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Argentina	2021	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Argentina	2022	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Australia	2017	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Australia	2018	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Australia	2019	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#

Figure 22

Inserted a Pivot Table. Add Date to columns. Add Country to rows. Add Food Packaging(CO2) to values. Inserted a cluster column chart (Figure 23).

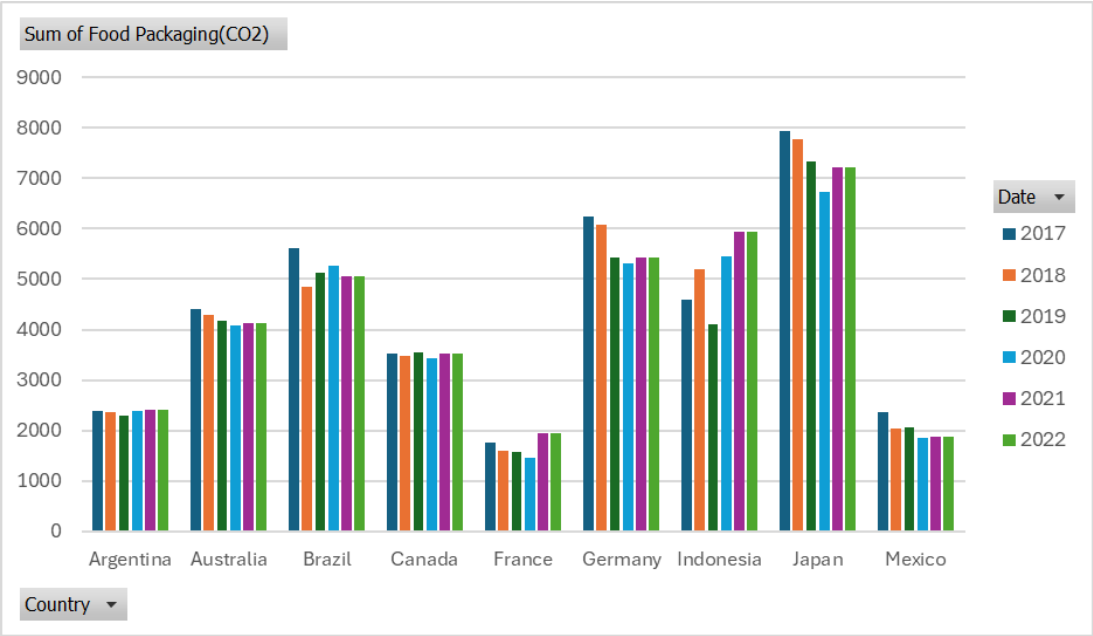


Figure 23

As can be seen from Figure 23, the carbon dioxide emissions of food packaging in different countries are significantly different, and there is no single upward or downward trend. Argentina's emissions are stable and low; Australia fluctuates around 4000-4500; Brazil's overall trend is downward; Canada is very stable; France's emissions are the lowest among the 10 countries, but compared with the previous three years, they have increased in the past two years; Germany has a significant decline, but it is still high; Indonesia has a rapid upward trend; Japan's emissions are still the highest, although they have declined; Mexico's emissions are relatively low, with little change.

2.3.3 How will the amount of CO2 produced by industrial wastewater discharge differ in different countries in 2022?

Saved XLS file and opened it in Tableau. Right click the Country label and set it to Country. Set Item to Date type. Dragged Longitude generated automatically to Columns. Dragged Latitude generated automatically to Rows. Dragged Country to Detail. Dragged Industrial Wastewater(CO2eq) (AR5) to Color. Added Date to Filters and only selected 2022. Chose Map chart (Figure 24).

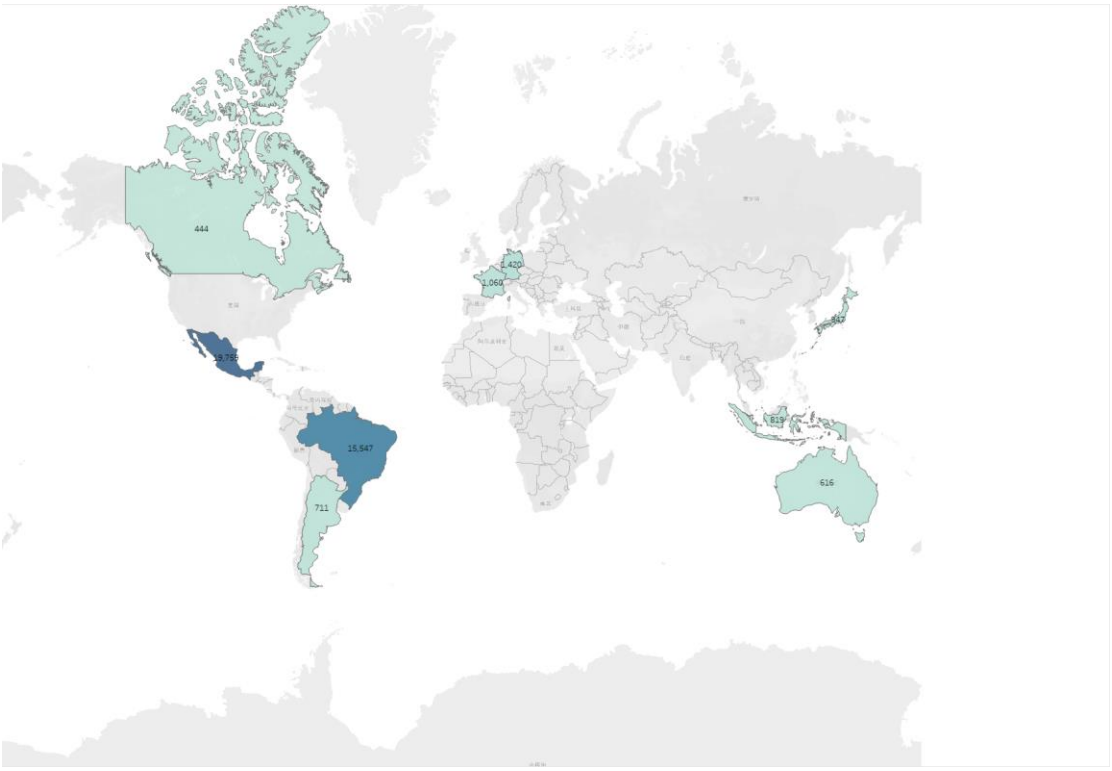


Figure 24

As can be seen from Figure 24, CO2 emissions from industrial wastewater are not uniform across countries, with a few countries producing most emissions. Of these 10 countries, Mexico and Brazil produced most emissions, accounting for 85% of the total of the 10 countries, with Mexico's emissions reaching a staggering 19,759. And Canada had the best emissions performance, only 444.

2.3.4 Is there a correlation between agriculture water use efficiency and agriculture water stress levels in different countries?

Selected 10 countries, downloaded their SDG Indicators dataset from FAOSTAT in XLS format. Deleted some unnecessary columns. Filtered Value column, delete all rows with empty data. Created a new column to remove the number before the text in Item column (Figure 25).

C2 =RIGHT(B2, LEN(B2) - FIND(" ", B2))

	A	B	C	D	E
1	Area	Item(old)	Item	Year	Value
2	Russian Fe	2.c.1 Indicator of Food Price Anomal	Indicator of Food Price Anomalies (IFPA), by type of produc	2023	-3.8
3	Belgium	2.c.1 Indicator of Food Price Anomal	Indicator of Food Price Anomalies (IFPA), by Food CPI	2021	-2.3
4	Italy	2.c.1 Indicator of Food Price Anomal	Indicator of Food Price Anomalies (IFPA), by type of produc	2023	-1.9
5	China, mai	2.c.1 Indicator of Food Price Anomal	Indicator of Food Price Anomalies (IFPA), by Food CPI	2021	-1.9
6	Austria	2.c.1 Indicator of Food Price Anomal	Indicator of Food Price Anomalies (IFPA), by Food CPI	2021	-1.8

Figure 25

Sorted Value column from smallest to largest. Use the right function to change all abnormal data. Inserted a Pivot Table in a new sheet. Added variables to Pivot Table Filed. Transposed the whole Table as values in a new sheet. Highlighted all empty value through Conditional Formatting and cleaned empty value (Figure 26).

Country	Y	[F	[F	A	A	A	F	In	L	L	L	L	N	N	N	N	N	N	N	N	P	P	P	P	P	P	P	W	W	W	W
Australia	2(	3	3	0	0	2	#	2	2	1	3	1	#	#	#	#	#	#	#	#	#	#	#	#	4	3	4	0	0	#	#	#
Australia	2(	4	4	0	1	3	#	#	3	1	5	1	#	#	#	#	#	#	#	#	#	#	#	#	3	3	4	0	1	#	#	#
Austria	2(	5	4	1	1	1	#	1	0	6	9	2	#	#	0	#	#	#	#	#	#	#	3	4	3	1	1	1	0	2	#	#	#
Austria	2(	5	4	1	1	1	#	#	0	6	9	2	#	#	0	#	#	#	#	#	#	#	4	4	5	2	1	2	0	2	#	#	#
Belgium	2(	4	3	0	0	1	#	1	1	#	#	9	#	#	#	#	#	#	#	#	#	#	5	4	5	1	1	1	0	1	#	#	#
Belgium	2(	4	3	0	0	1	#	#	1	#	#	9	#	#	#	#	#	#	#	#	#	#	6	6	6	2	2	2	0	1	#	#	#
Brazil	2(	4	5	0	1	6	#	2	1	0	1	0	#	#	#	#	#	#	#	#	#	#	#	#	#	6	8	5	3	1	#	#	#
Brazil	2(	4	5	0	1	7	#	0	1	0	1	0	#	#	#	#	#	#	#	#	#	#	#	#	#	9	#	7	4	1	#	#	#
Canada	2(	4	5	0	1	2	#	0	0	3	4	1	#	#	1	#	#	#	#	#	#	#	7	7	6	1	1	1	0	0	#	#	#
Canada	2(	4	4	1	2	2	#	#	0	3	4	1	#	#	1	#	#	#	#	#	#	#	8	9	6	1	1	1	0	0	#	#	#
France	2(	5	4	0	0	2	#	1	3	#	#	5	#	#	#	#	#	#	#	#	#	#	6	6	6	1	1	1	0	2	#	#	#
France	2(	5	4	0	0	2	#	#	3	#	#	5	#	#	#	#	#	#	#	#	#	#	7	7	6	2	2	2	0	2	#	#	#

Figure 26

Saved XLS file and opened it in SAP analytics cloud. Added Year, Country to dimensions. Added Level of water stress to Line Axis. Added Water Use Efficiency to column Axis. Selected Combination Column&Line Chart (Figure 27).

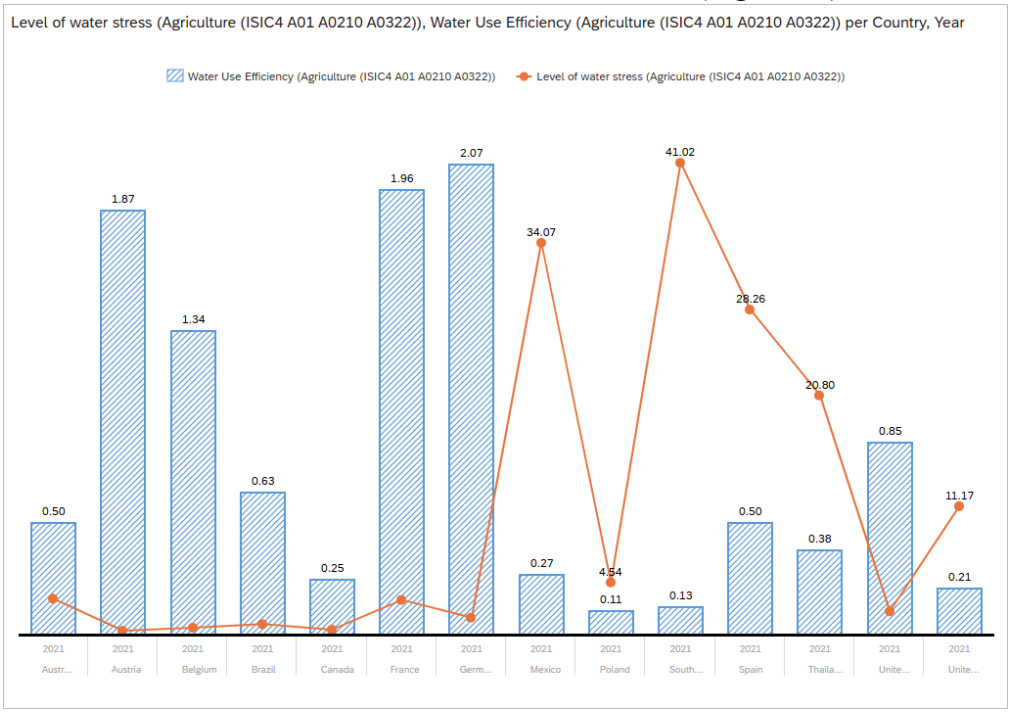


Figure 27

It is obvious from Figure 27 that there are large differences in agriculture water use efficiency and water stress levels in different countries. France, Germany, Belgium, Australia and other countries have high water resource use efficiency and low water resource stress levels. Mexico, South Africa, Spain and Thailand have water resource stress and low water use efficiency. Due to the small sample size (10 countries) and the presence of many factors that may affect water use efficiency and water resource stress levels, such as natural endowment, scientific and technological level, economic conditions, humanistic concepts, etc., it is difficult to draw a conclusion that there is a correlation between agriculture water use efficiency and level of agriculture water stress in different countries.

2.4 Personal Reflection and Conclusion

While my analysis of the FAOSTAT dataset in Chapter 2 provided me with valuable insights into global agricultural and environmental trends, the process also highlighted the strengths and limitations of using such datasets. FAOSTAT is a comprehensive and authoritative database that provides extensive coverage of key indicators across countries and years, making it indispensable for cross-national and longitudinal studies. However, while the completeness of the dataset was generally high, I encountered missing values and inconsistencies, particularly in fields such as the value column and metadata labels. These gaps required thorough data cleaning to ensure accuracy, such as filtering out empty rows, standardizing formats (e.g., using LEFT/RIGHT functions to refine the year field), and removing redundant columns.

Using Excel for data cleaning proved advantageous because of its accessibility and user-friendly interface. Tasks such as filtering, sorting, and applying basic functions (e.g., conditional formatting, pivot tables) were simple and allowed for quick adjustments to the dataset. For example, removing unnecessary columns or transposing data for visualization in Tableau was highly efficient. However, when working with larger or more complex datasets, Excel's limitations became apparent. Its performance lags when processing large numbers of tables, while advanced transformations (e.g., automatic error detection, merging datasets with irregular structures) require tedious manual work. Additionally, Excel lacks native support for version control and scripting, which makes repetitive tasks error-prone and time-consuming.

Overall, this experience highlights the importance of balancing tool selection with project needs. Excel remains a strong entry-level tool for basic data preparation, but its limitations highlight the need to upskill in more advanced platforms in complex scenarios. Despite the need for careful cleaning, FAOSTAT's rich data provides tremendous value for global insights—as long as analysts combine it with the right methods and tools to transform raw data into actionable knowledge.

Chapter 3 SAP Analytics Cloud

3.1 About SAP Analytics Cloud

SAP Analytics Cloud (SAC) is a cloud-based analytics platform developed by SAP that integrates multiple functions such as data visualization, business intelligence, financial planning, and predictive analysis to help companies manage and analyze business data more effectively. Through an intuitive and interactive interface, users can easily create dashboards, conduct data exploration, and use built-in predictive models for trend analysis and decision support, thereby achieving data-driven efficient collaboration and intelligent decision-making, and improving the overall management efficiency and competitiveness of the company.

3.2 Dataset Source and Research Questions

I continue and expand on what I learned in this week by using SAP Analytics Cloud to analyze Global Bike Inc.'s sales data and answer the following questions:

6) What is the geographic distribution of the company's U.S. revenue sources?
7) Does the company's revenue have seasonal characteristics? What changes will it show in the future?
8) In 2023, how will different customers contribute to the company's revenue and what will be the gross profit margin?
9) In 2023, how will different customers contribute to the company's revenue and what will the gross profit margin be?

The dataset is GB_AnalyticsData3.xlsx, which is provided in class. This dataset contains detailed sales-related information, including order details, dates, customer information (such as customer names and geographical locations), product categories, divisions, sales quantities, revenues, discounts, and costs. Financial figures such as revenue, discounts, and costs are converted into USD, facilitating standardized analysis. Additionally, geographic coordinates (latitude and longitude) are provided, making this dataset ideal for analyzing sales performance, regional market distribution, and customer purchasing behaviors through visualization and analytical techniques.

3.3 Analysis tool application and results

Before starting visualizing, modeling the data is necessary. Import the Excel file, GB_AnalyticsData3.xlsx. Convert some variables to dimensions and link these related variables. Creating hierarchies which allows me to look at data at different levels of granularity or detail. Change the data type of Longitude and Latitude to the geo-dimension. Combine three dimensions, Year, Month, and Day. The final completed model as a graph is shown in the following figure.

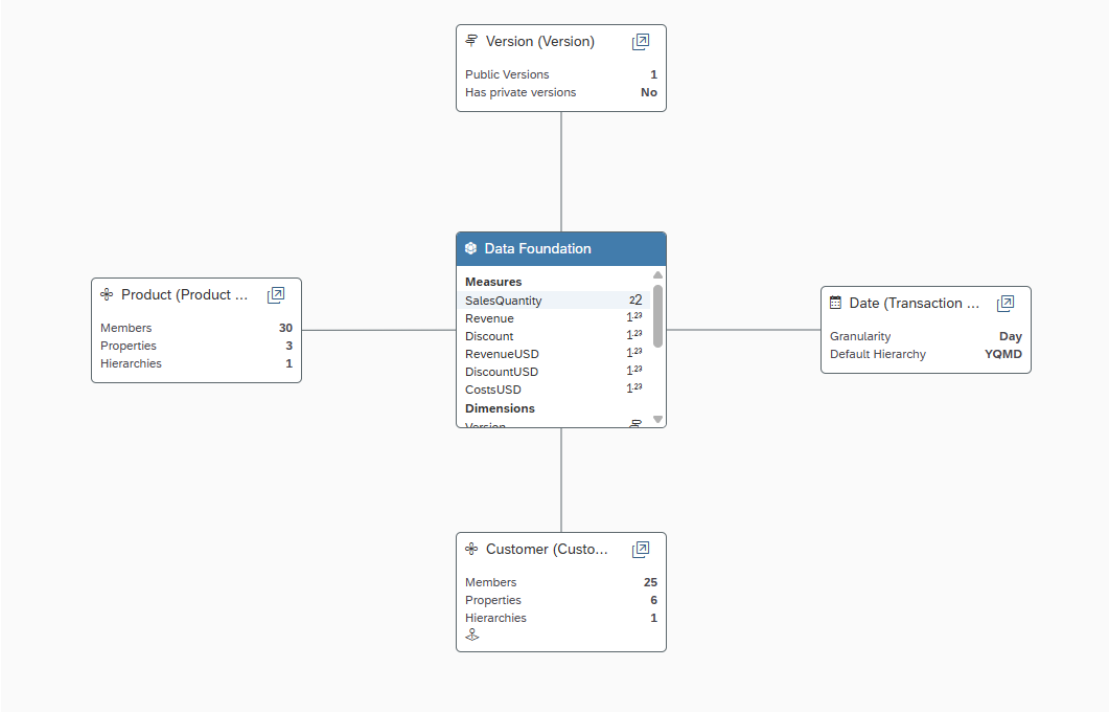


Figure 28

3.3.1 What is the geographic distribution of the company's U.S. revenue sources?

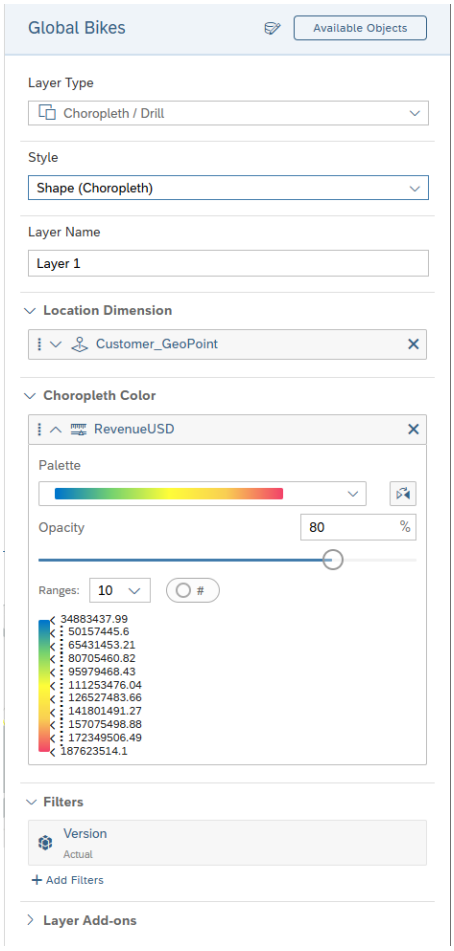


Figure 69

Dra the Geo Map from the widgets to story page. Choose Light Grey as base map and rename the map as “Global Bike Sales”. Then choose Choropleth/Drill as Layer Type. Choose Customer_GeoPoint as Location Dimension and RevenueUSD as Choropleth Color. Expand the Range of Choropleth Color to 10 and toggle to # instead of %. (Figure 29)

Select the hierarchy icon on Layer 1 and select only Region on the Edit Layer Hierarchy screen. The map looks like this. (Figure 30)

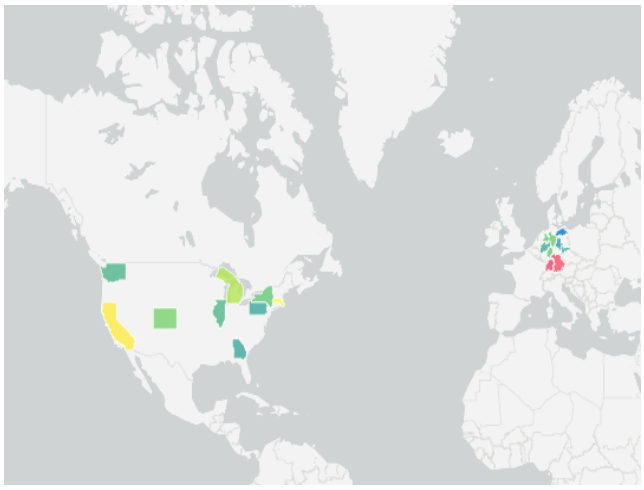


Figure 70

Choose the polygon filter and click the rectangle shape. Draw a rectangle around the U.S. and filter to the continental United States. Select light blue as Background Color. Add Layer 2 and choose Bubble as Layer Type, Customer_GeoPoint as Location Dimension and SalesQuantity as Bubble Color. Expand the Range of Choropleth Color to 10 and toggle to # instead of %. (Figure 31)

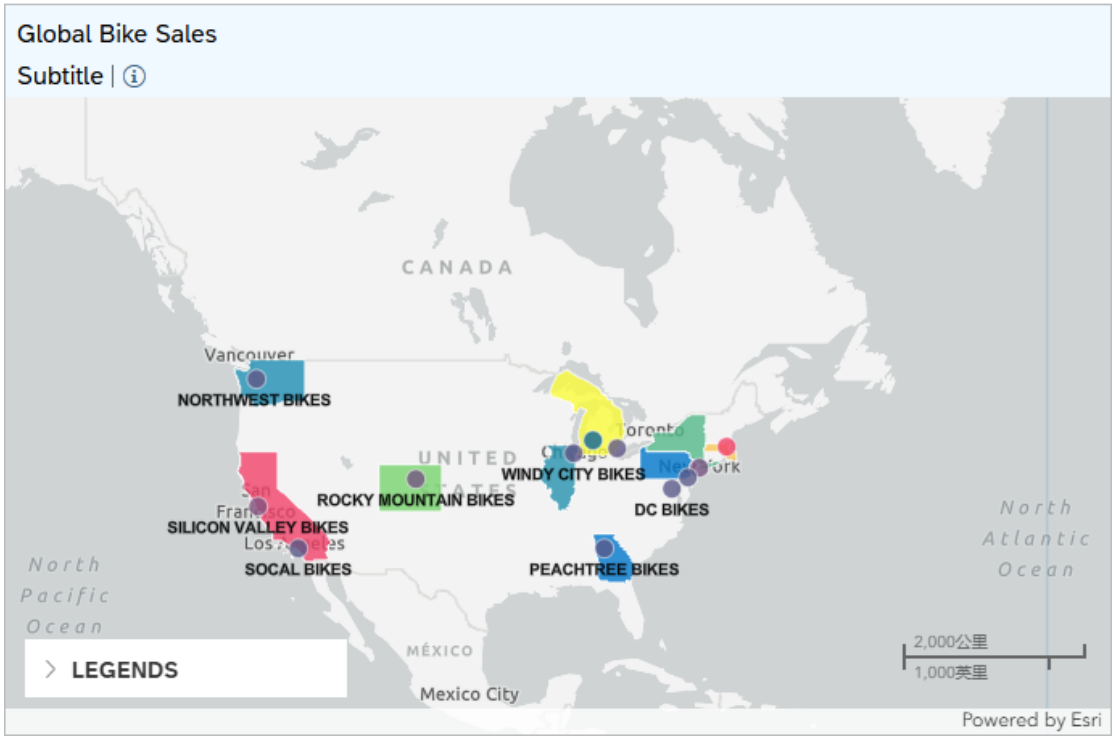


Figure 31

As can be seen from the figure, the company's product sales in the United States are concentrated in several specific provinces. California is the highest, followed by Massachusetts, and Michigan is third. In addition, Beantown was the city with the highest sales.

3.3.2 Does the company's revenue have seasonal characteristics? What changes will it show in the future?

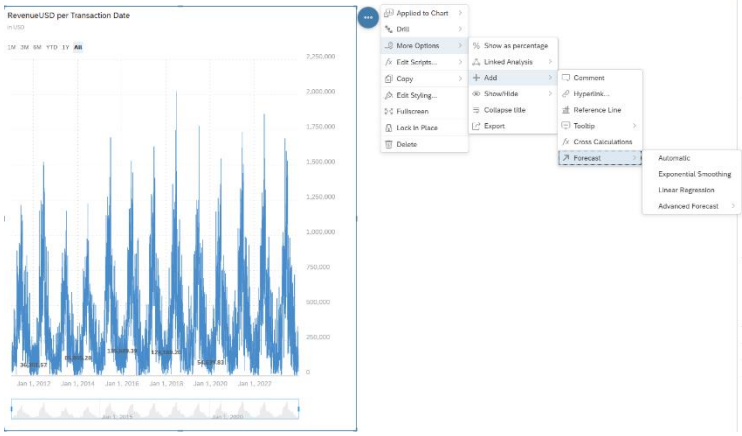


Figure 32

Create a new Responsive Page in my Story. Add a Chart and Transaction date as the dimension and Revenue USD as the Account. Change the chart to a time series. Add Forecast on the Actions menu and choose Linear Regression. (Figure 32) Adjust the time frame to 1 Year.

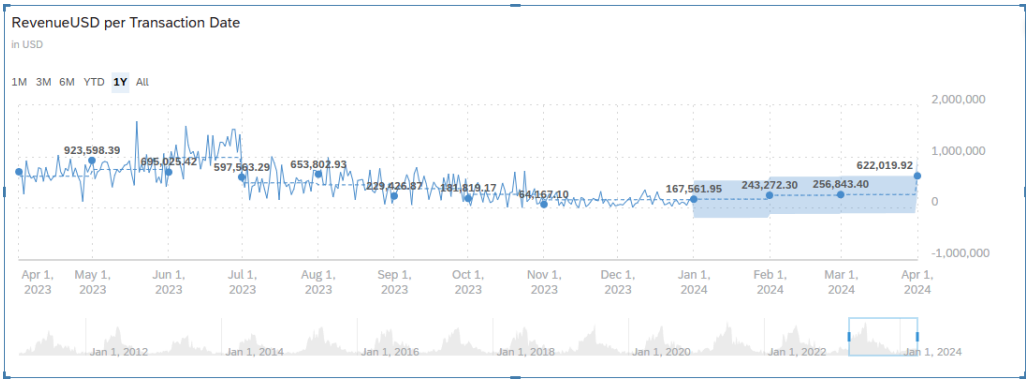


Figure 33

As shown in Figure 33, the company's revenue shows a significant seasonal characteristic, mainly concentrated in the summer, reaching its peak in June, and then gradually decreasing, with less revenue in winter.
According to forecasts, the company's revenue will gradually increase from January to March and will increase rapidly in April.

3.3.3 In 2023, how will different customers contribute to the company's revenue and what will be the gross profit margin?
Add a Responsive Page to the Story. Drag a Chart widget to the Lane canvas. Use "(["Global Bikes":RevenueUSD] - ["Global Bikes":CostsUSD] - ["Global Bikes":DiscountUSD]) /["Global Bikes":RevenueUSD] * 100" to create a calculated measure for Gross Margin Ratio . (Figure 34)

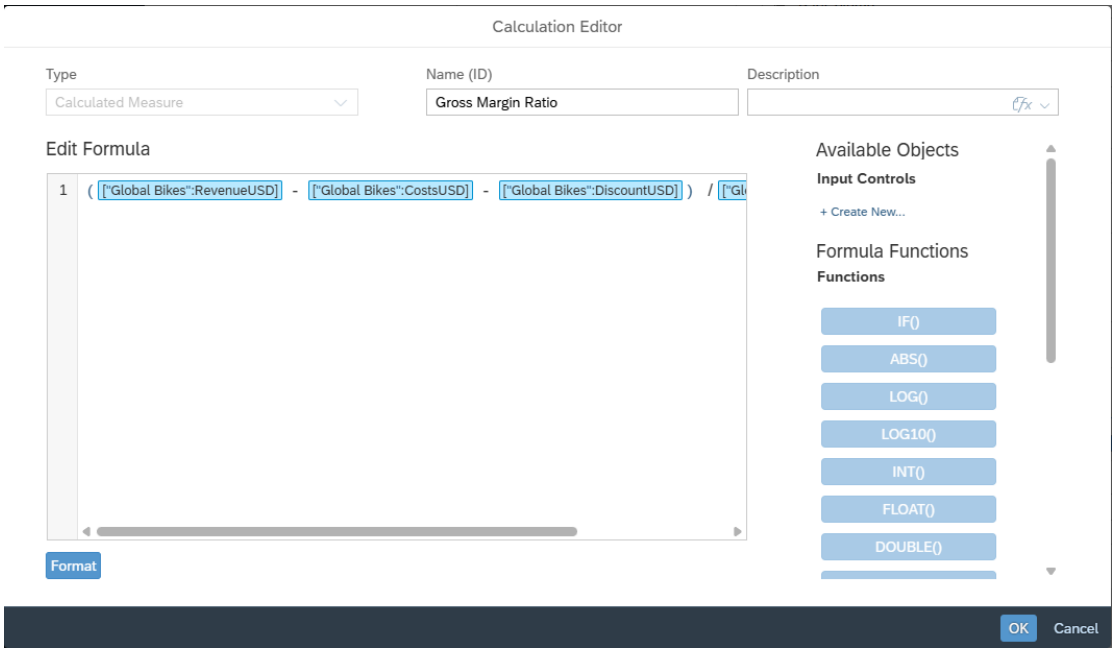


Figure 34

Select a bubble chart. Put Revenue USD on the x-axis, SalesQuantity on the y-axis, and Gross Margin Ratio as bubble size. Add Customer Number to both the dimension and the color. Drill to Level 4 of the hierarchy. Filter to the year 2019. Add the data labels by Gross Margin Ratio to the bubble chart. (Figure 35)

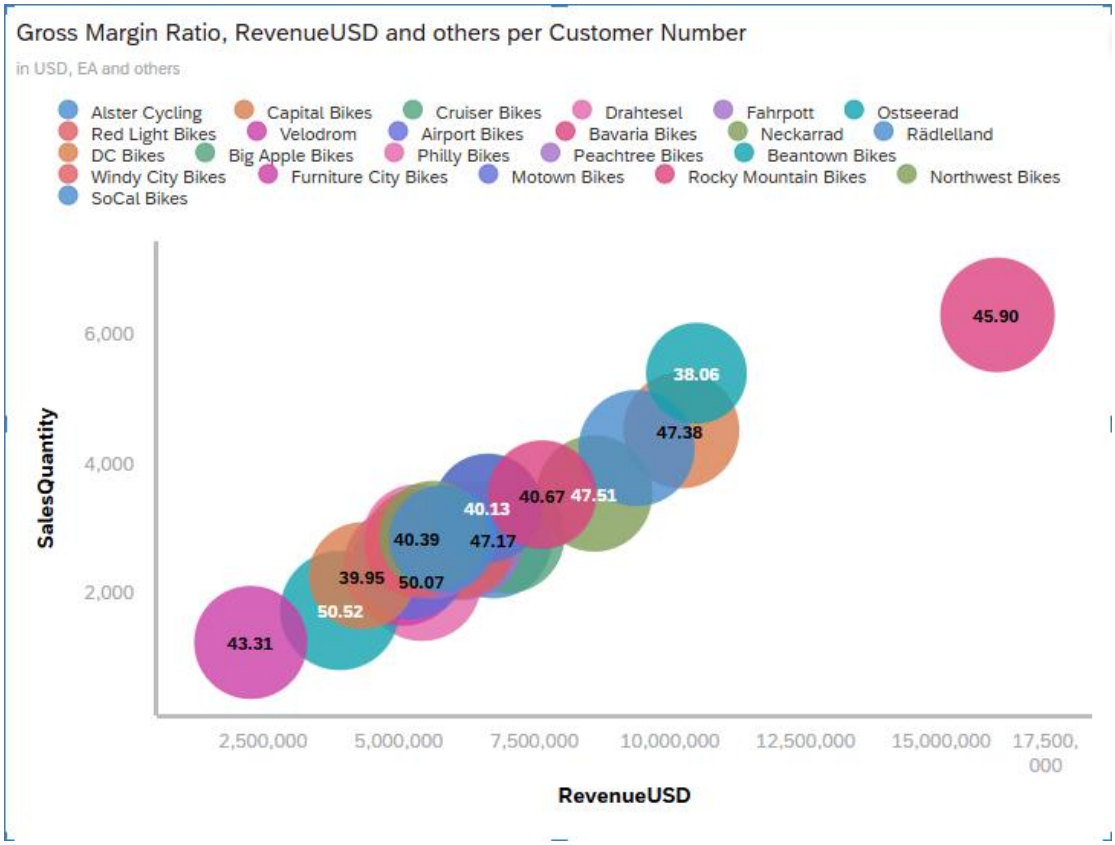


Figure 35

It can be clearly seen from the figure that in 2023, the gross profit margins of most customers were concentrated between 40% and 50%. Bavaria Bikes contributed the highest sales volume and sales revenue, far exceeding other customers, and its gross profit margin reached 45.9%.

3.3.4 In 2023, how will different customers contribute to the company's revenue and what will the gross profit margin be?

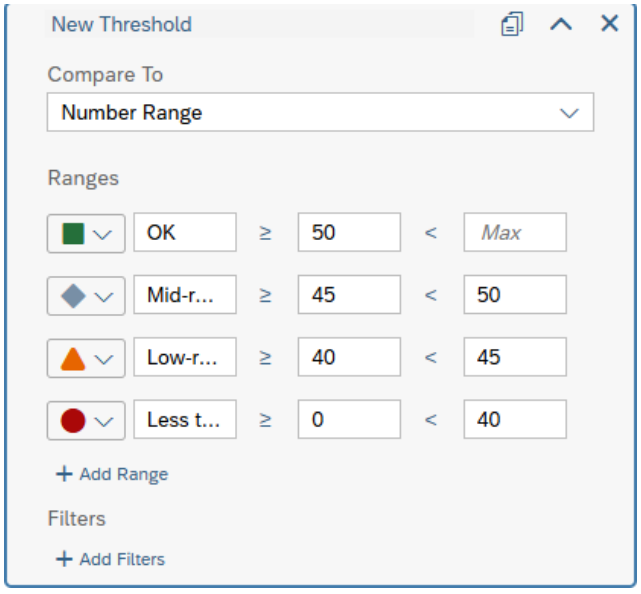


Figure 36

Add a New Responsive Page and drag a Chart to the Lane canvas. Select Bullet Chart, add Gross Margin Ratio to Measures and Customer City to Dimensions. Create a new threshold. Set the Ranges = 50 to Max (OK), 45 to 50 (Mid-range), 40 to 45 (Low-range), and 0 to 40 (Less than Threshold). (Figure 36).

Sort Gross Margin Ratio from High to Low. (Figure 37)

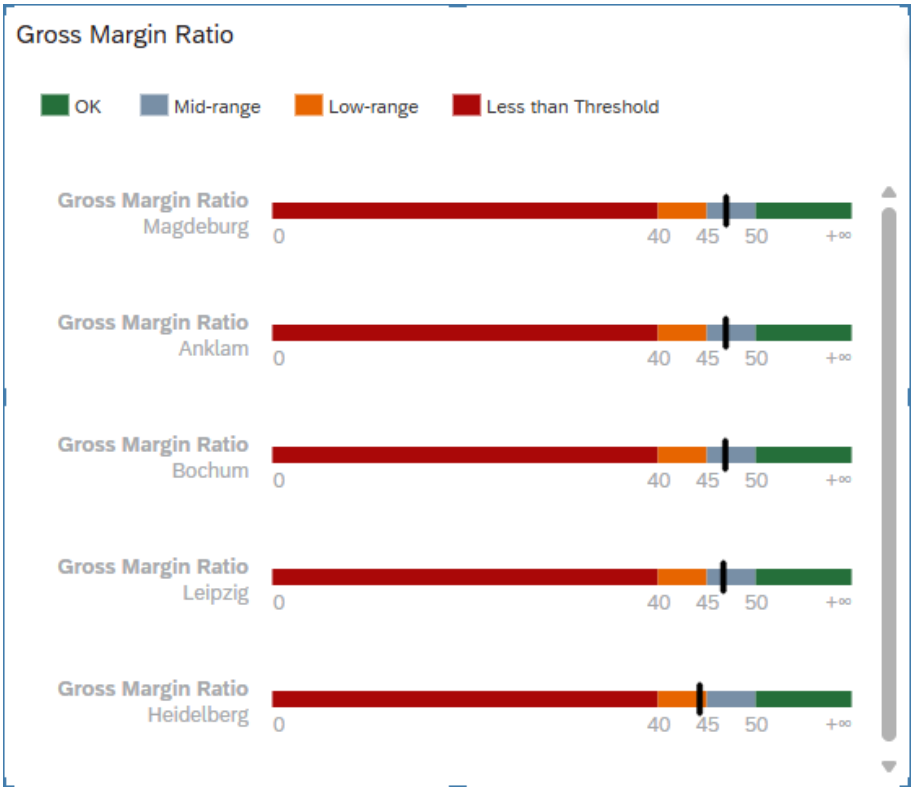


Figure 37

We can clearly see that no city has a gross profit margin of more than 50, and only four cities have a gross margin ratio in the mid-range (above 45), namely Magdeburg, Anklam, Bochum, and Leipzig.

Then sort Gross Margin Ratio from Low to High. (Figure 38)

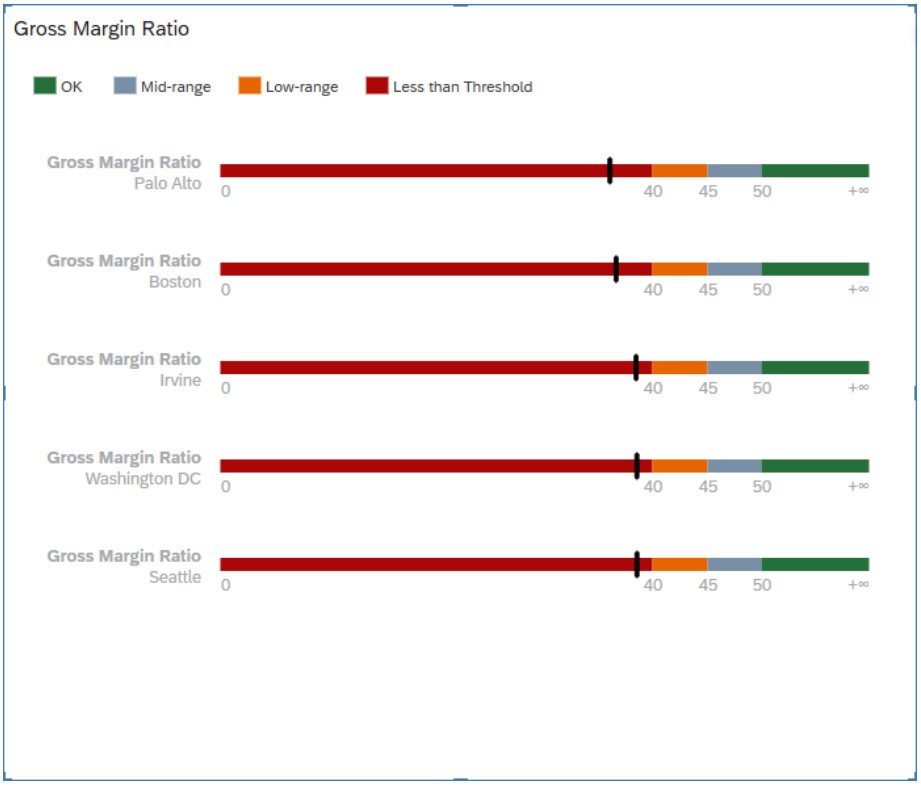


Figure 38

The gross profit margins of several cities with the lowest gross profit margins did not reach 40%, with Pala Alto having the lowest gross profit margins.

3.4 Personal Reflection and Conclusion

Throughout this analysis, I compared and practiced using SAP Analytics Cloud (SAC) and traditional Excel-based data processing methods. From this hands-on experience, I discovered several advantages and disadvantages of each tool.

Firstly, SAP Analytics Cloud, as a cloud-based data analytics platform, has notable advantages. It integrates multiple functionalities, including data visualization, business intelligence, financial planning, and predictive analytics. Its intuitive and interactive interface significantly enhances user experience and effectively improves analytical efficiency. SAC enables users to visualize complex datasets, perform sophisticated predictive analytics, and make data-driven decisions more effectively than Excel's basic regression analysis tools.

However, SAP Analytics Cloud also presents some challenges. The preliminary data modeling and preparation processes can be intricate, involving detailed definition of multiple dimensions and geographical dimensions. This complexity can become a barrier for users unfamiliar with advanced data analytics.

On the other hand, Excel excels in simplicity and accessibility, especially in tasks involving basic data cleaning and visualization through pivot tables. Most users find Excel's interface intuitive and straightforward, making it highly efficient for straightforward tasks and lowering the learning curve significantly. Nevertheless, Excel can be limited when handling large datasets and advanced predictive analytics.

In summary, SAP Analytics Cloud is better suited for processing complex, large-scale data sets that require detailed visualization and predictive analysis, while Excel excels at handling simpler, everyday data cleansing and visualization tasks. In future professional scenarios, choosing the right tool based on complexity and specific analytical requirements will help maximize the advantages of each platform, thereby improving analytical efficiency and decision-making accuracy.

Chapter 4 Tableau

4.1 About Tableau

Tableau is a powerful data visualization tool widely used in business intelligence to help users easily connect, visualize, and share data-driven insights. It enables users without extensive technical knowledge to create interactive dashboards, detailed reports, and visually appealing charts. Tableau supports connections to various data sources, such as databases, Excel sheets, and cloud services, allowing businesses to quickly analyze complex datasets. Its intuitive drag-and-drop interface helps users transform raw data into meaningful visual stories, making it ideal for decision-making, trend analysis, and effective communication across teams.

4.2 Dataset Source and Research Questions

I continue and expand on what I learned in this week by using Tableau to analyze Global Bike Inc.'s sales data and global CO2 emission data to answer the following questions:

1) Which year had the highest revenues (in USD) overall and how much were revenues during that year?
2) What was the year with the highest overall gross margin (in USD) and what was the amount?
3) What is the trend of CO2 emissions in countries around the world between 1994 and 2011?
4) What are the differences in per capita CO2 emissions among countries around the world in 2011?

The datasets I used are GBI_E5_2 .xlsx and GB_AnalyticsData3.xlsx, which is provided in class.

The first dataset contains sales transaction data, including details such as order numbers, items, dates, customers, sales quantities, revenue, discounts, and costs. It provides information about products sold (e.g., bikes), customer types, geographical locations, and financial metrics in multiple currencies. This dataset is particularly useful for analyzing sales performance, customer behavior, and financial outcomes.

The second dataset, sourced from the World Bank, includes cleaned CO2 emission data by country and region. It covers total CO2 emissions measured in kilotons (kt) and CO2 emissions per capita in metric tons over various years. The dataset enables the analysis of environmental performance, emission trends, and comparisons across different geographical regions or individual countries.

4.3 Analysis tool application and results

4.3.1 Which year had the highest revenues (in USD) overall and how much were revenues during that year?

Open 'GBI_E5_2 .xlsx' in Tableau and create a new sheet. Convert the Data Type of 'Year' to string. Drag 'Revenue in USD' from Measures into Columns and 'Year' from Dimension into Rows. Change chart type into 'Bar'. Click 'Label' and check 'Show Mark Labels'. Sort Year Descending by 'Revenue in USD'. (Figure 39)

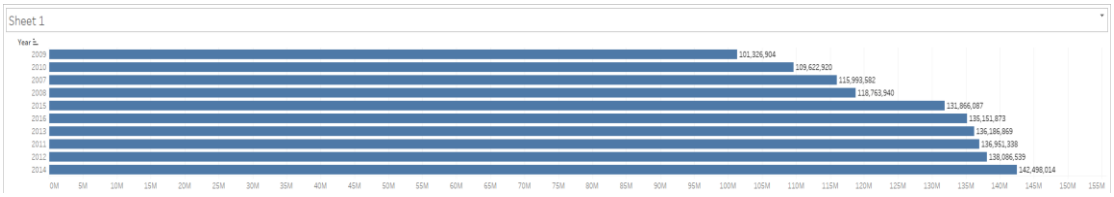


Figure 39

As can be seen from the figure, the company achieved the highest revenue in 2014, exceeding 140M USD, followed by 2012, with revenue reaching 138M. The worst year was 2009, with only 101M USD.

4.3.2 What was the year with the highest overall gross margin (in USD) and what was the amount?

Create a new Calculated Field named ‘Gross Margin in USD’ by entering ‘[Revenue USD] - [Discount USD] - [Costs in USD]’ as formula. (Figure 40)

Figure 40

Drag ‘Revenue in USD’ and ‘Gross Margin in USD’ from Measures into Rows and ‘Year’ from Dimension into Columns. Right-click ‘Year’ within Columns and select all years. Right-click the vertical axis of ‘Gross Margin in USD’ and choose ‘Dual Axis’. Choose ‘Side-by-side’ bar chart and show ‘Labels’. Drag ‘Country’ into Colour. (Figure 41)

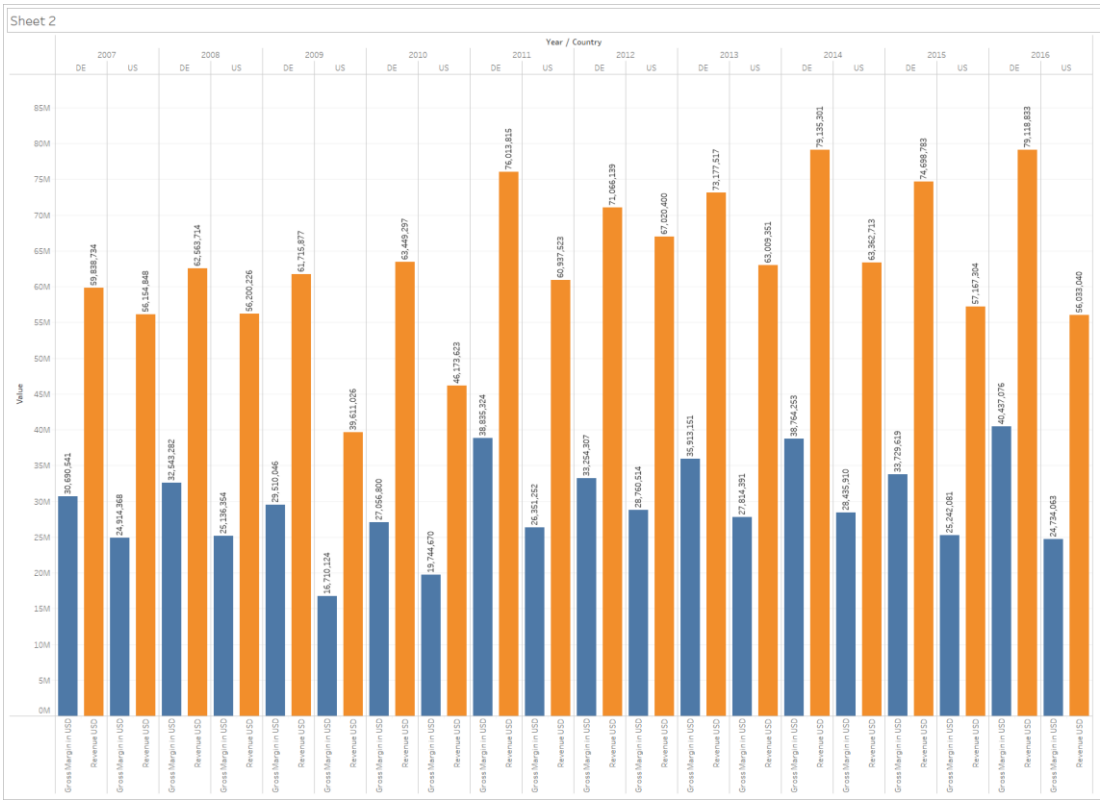


Figure 41

As can be seen from the figure, the company's gross profit in Germany was the highest in 2016, exceeding 40M, followed by 2011, also in Germany, exceeding 38M. Overall, the company's gross profit and annual revenue showed an upward trend from 20007 to 2016, with the German market contributing more than the US market.

4.3.3 What is the trend of CO2 emissions in countries around the world between 1994 and 2011?

Open ‘World_Bank_CO2.xlsx’ in Tableau. Drag ‘CO2 (kt) RAW DATA’ to the ‘Drag sheet here’ box. Use ‘Data Interpreter’ to clean the dataset. Select year columns from 1961 to 2011 and create ‘Pivot Field Names’ and ‘Pivot Field Values’ columns and rename them as ‘Years’ and ‘CO2 Emission’. (Figure 42)

CO2 (kt) RAW DATA Country Name	CO2 (kt) RAW DATA Country Code	Abc CO2 (kt) RAW DATA Indicator Name	Abc CO2 (kt) RAW DATA Indicator Code	Abc Pivot Year	# Pivot CO2 Emission
Aruba	ABW	CO2 emissions (kt)	EN.ATM.CO2E.KT	1960	*null*

Figure 42

Drag and Drop ‘CO2 Emission’ to Rows and ‘Year’ to Columns. Choose Line as chart type. Change screen size to Fit Width. Drag and Drop ‘Country Name’ to Color. Drag and Drop ‘Year’ to Filters and select all years after 1994. Drag and Drop ‘Country Name’ and remove all regions. (Figure 43)

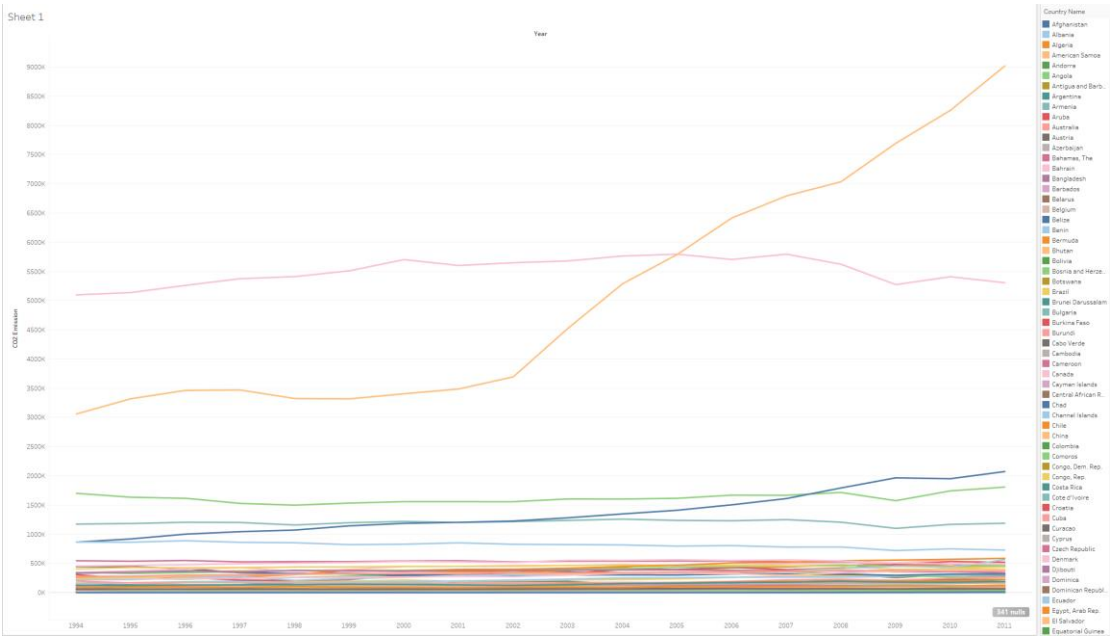


Figure 43

As shown in the figure, China's CO2 emissions have risen sharply, from 3000k in 1994 to 9000k in 2011, and surpassed the United States in 2006 to become the country with the largest CO2 emissions in the world. The United States' emissions during this period showed a trend of first rising and then falling, but the overall change was not large, hovering around 5500k. India is another country with a rapid increase, rising from less than 1000k to more than 2000k, and surpassed Japan in 2008 to become the third in the world.

4.3.4 What are the differences in per capita CO2 emissions among countries around the world in 2011?

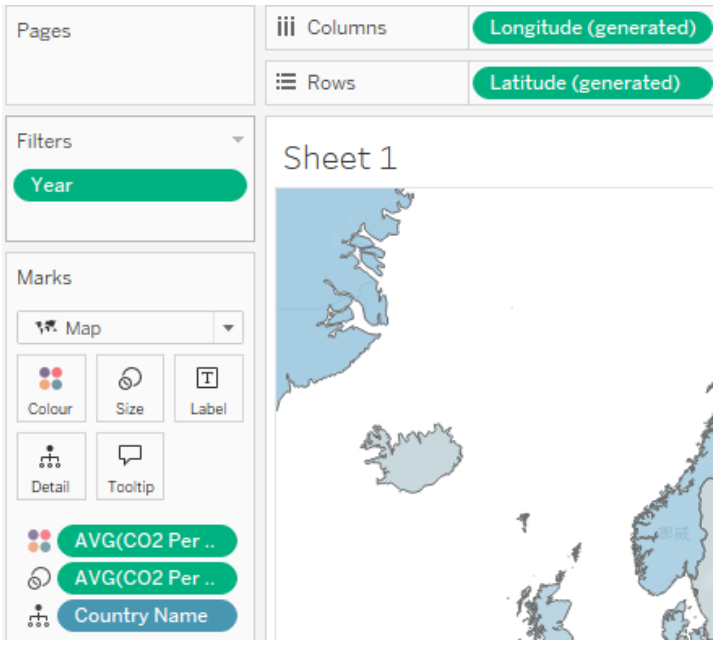


Figure 44

Drag ‘CO2 Data Cleaned’ to the data box. Drag and Drop ‘Longitude’ into columns and ‘Latitude’ into rows. Drag and Drop ‘Country Name’ into Deta and ‘CO2 Per Capita’ into Size. Drag and Drop ‘CO2 Per Capita’ into Color. Change the measure of ‘CO2 Per Capita’ from Sum to Average. Filter ‘Year’ and only select 2011. (Figure 44)

Select ‘Red-Black-White Diverging’ as Palette. Check ‘Reversed’ and ‘Full Color Range’. (Figure 45)

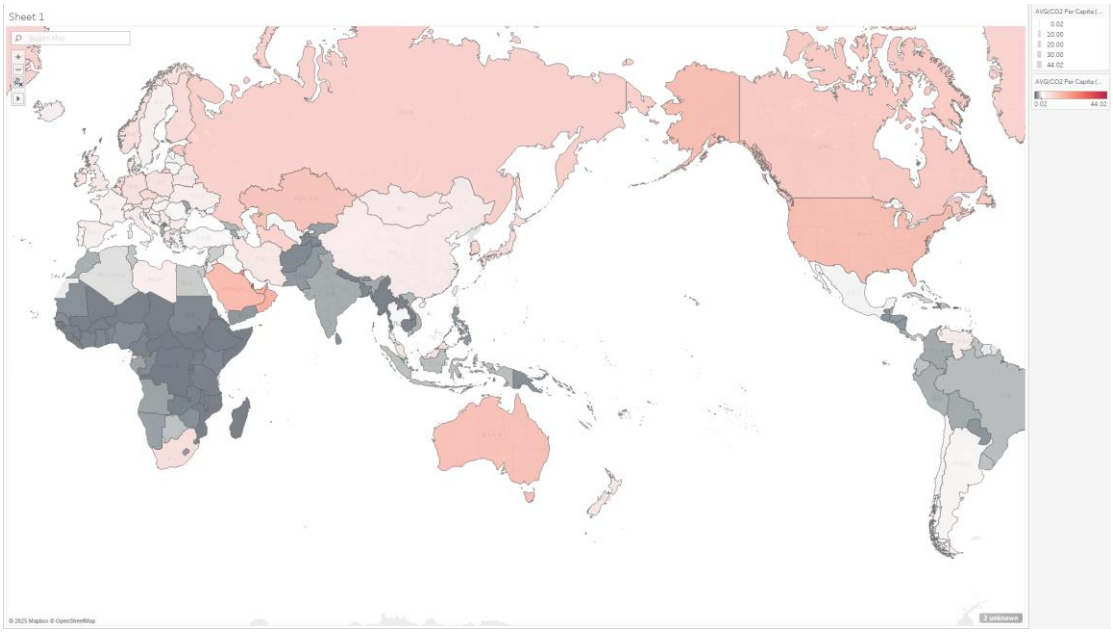


Figure 45

It can be clearly seen from the figure that there are significant differences in per capita CO2 emissions in different countries. Europe, East Asia, North America, and the Gulf countries have relatively high per capita emissions, while India, sub-Saharan Africa, and Latin America have relatively low per capita emissions. The reasons for this phenomenon may be related to economic development, industrial structure and other factors.

4.4 Personal Reflection and Conclusion

Through this analytical experience, I explored and been familiared the useful analytical tools: Tableau, understanding its strengths and weaknesses concerning data cleaning, visualization, and predictive analysis.

Tableau stands out primarily for its powerful and intuitive data visualization capabilities. Its straightforward drag-and-drop interface allows users, even without extensive technical backgrounds, to create visually appealing, interactive dashboards and insightful visual narratives. Tableau is particularly advantageous for rapidly connecting to various data sources and transforming large and complex datasets into understandable visual insights, facilitating clear communication across teams. However, Tableau can become challenging when complex data preparation and advanced predictive analytics are required, as its predictive modeling capabilities are limited compared to specialized analytical software.

SAP Analytics Cloud (SAC) excels in integrating multiple analytical functionalities into one cloud-based platform, including advanced data visualization, business intelligence, financial planning, and sophisticated predictive analytics. The intuitive user interface of SAC supports complex analytical tasks effectively, enabling users to create detailed dashboards and conduct predictive modeling and forecasting efficiently.

Nevertheless, SAC requires significant upfront effort in data modeling and preparation, involving detailed management of data dimensions, hierarchies, and geo-dimensions. Such complexities could pose challenges for users unfamiliar with advanced analytics techniques.

In conclusion, Tableau is ideal for dynamic, sophisticated visual storytelling and quick data insights; SAP Analytics Cloud is better suited for complex analytical tasks requiring integration of multiple analytics capabilities and predictive modeling. Understanding the unique strengths of each tool is crucial for selecting the right platform to optimize analytical effectiveness and decision-making precision in future professional scenarios.

Chapter 5 SAP Analysis For MS Excel

5.1 About SAP Analysis For MS Excel

SAP Analysis for Microsoft Excel is a powerful add-in designed to seamlessly integrate SAP Business Warehouse (BW) and SAP Analytics Cloud data directly into Excel, enabling users to efficiently analyze, visualize, and report on enterprise information. With this tool, users can leverage familiar Excel functionalities combined with advanced SAP analytic features to perform real-time data exploration, create dynamic reports, and execute complex calculations without needing extensive technical knowledge. It simplifies tasks such as filtering, slicing, and drilling down data, empowering business users to make informed decisions by interacting with live SAP data within Excel's intuitive environment.

5.2 Research Questions

I continue and expand on what I learned in this week by using SAP Analysis For MS Excel to analyze Global Bike Inc.'s sales data and answer the following questions:

1) What are the changes in the company's revenue and product sales volume from 2017 to 2019?
2) In 2007, which product contributed the most revenue to the company?
3) How did the company's air pump sales revenue and expenses change from 2007 to 2019?
4) What Customer provided the highest Revenue in 2009?

5.3 Analysis tool application and results

5.3.1 What are the changes in the company's revenue and product sales volume from 2017 to 2019?

Open a new Excel folder. Click on Analysis tab. Click on Insert Data Source and choose Select Data Source for Analysis. Choose 'Belfest' as SAP BW server and enter my login information. In the Search tab, search for DALCP1KX1095 and select my query. (Figure 46)

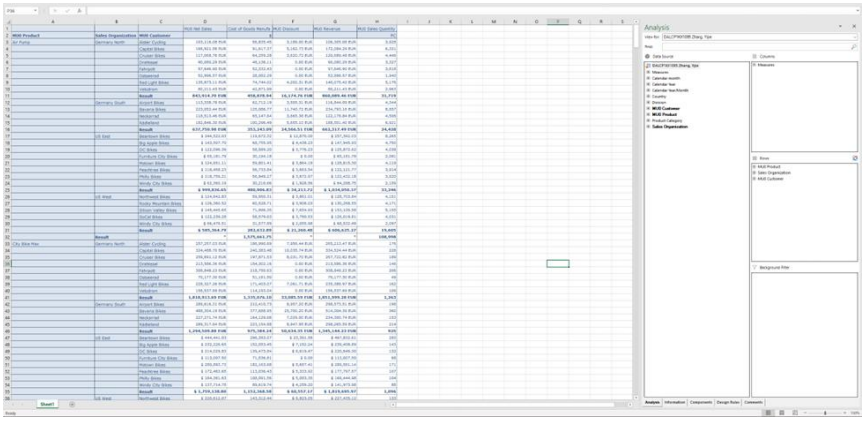


Figure 46

Click on Measures and use Currency Translation in the Analysis ribbon. Convert EURO to USD. (Figure 47)

	A	B	C	D	E	F	G	H
1				MU0 Net Sales	Cost of Goods Manufa	MU0 Discount	MU0 Revenue	MU0 Sales Quantity
2	**MU0 Product**	**Sales Organization**	**MU0 Customer**	$	$	$	$	PC
3	Air Pump	Germany North	Alster Cycling	138,845.80	56,835.45	4,293.99	143,139.79	3,928
4			Capital Bikes	224,759.88	91,617.37	6,951.62	231,711.50	6,321
5			Cruiser Bikes	157,633.09	64,259.28	4,875.30	162,508.38	4,446
6			Drahtesel	121,293.11	48,136.11	0.00	121,293.11	3,327
7			Fahrpott	131,481.55	52,332.43	0.00	131,481.55	3,618
8			Ostseerad	71,359.88	28,082.29	0.00	71,359.88	1,940
9			Red Light Bikes	182,953.14	74,744.02	5,658.41	188,611.55	5,176
10			Velodrom	108,004.69	42,871.99	0.00	108,004.69	2,963
11			**Result**	**1,136,331.14**	**458,878.94**	**21,779.31**	**1,158,110.46**	**31,719**

Figure 47

Drag ‘Sales Quantity’ and ‘Revenue’ to Measures, ‘Calendar Year’ to Rows. (Figure 48)

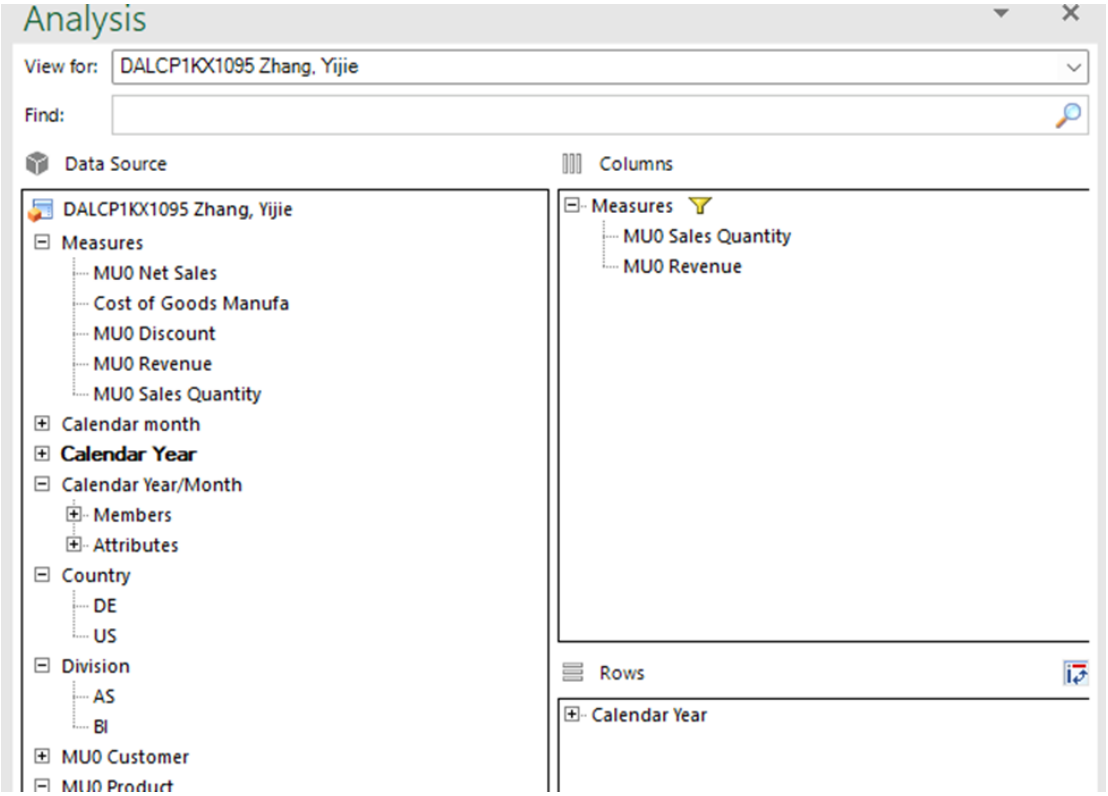


Figure 48

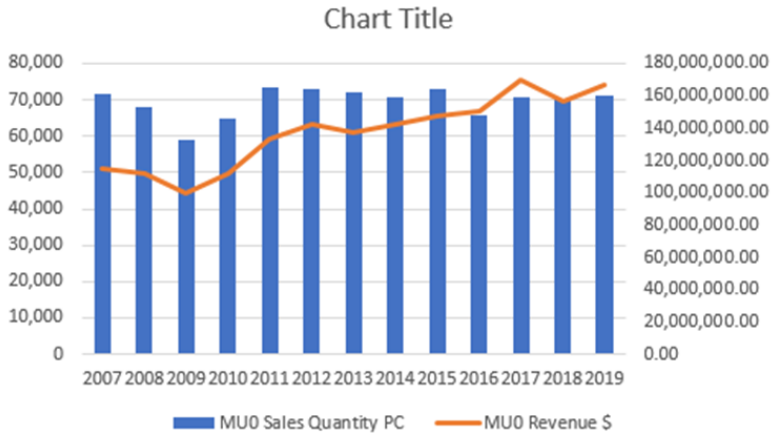


Figure 49

Insert a combo table. (Fugure 49)

As can be seen from the figure, the company's revenue has generally shown an upward trend, from 120 million to 160 million, a significant increase; at the same time, the sales volume of products has not changed much, except for 2009, it has basically remained at the level of 70,000 units per year.

5.3.2 In 2007, which product contributed the most revenue to the company?

Drag ‘Net Sales’ to Measures. Drag 2007 in ‘Calendar Year’ and ‘Product’ to Rows. (Figure 50)

Insert a bar chart and rename it as ‘Product Net Sales in 2007’. (Figure 51)

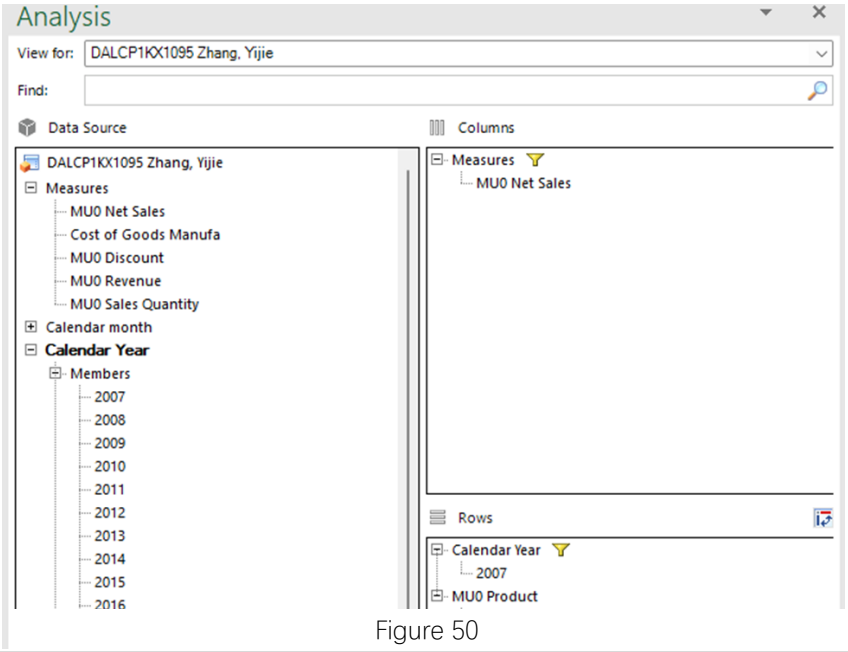


Figure 50

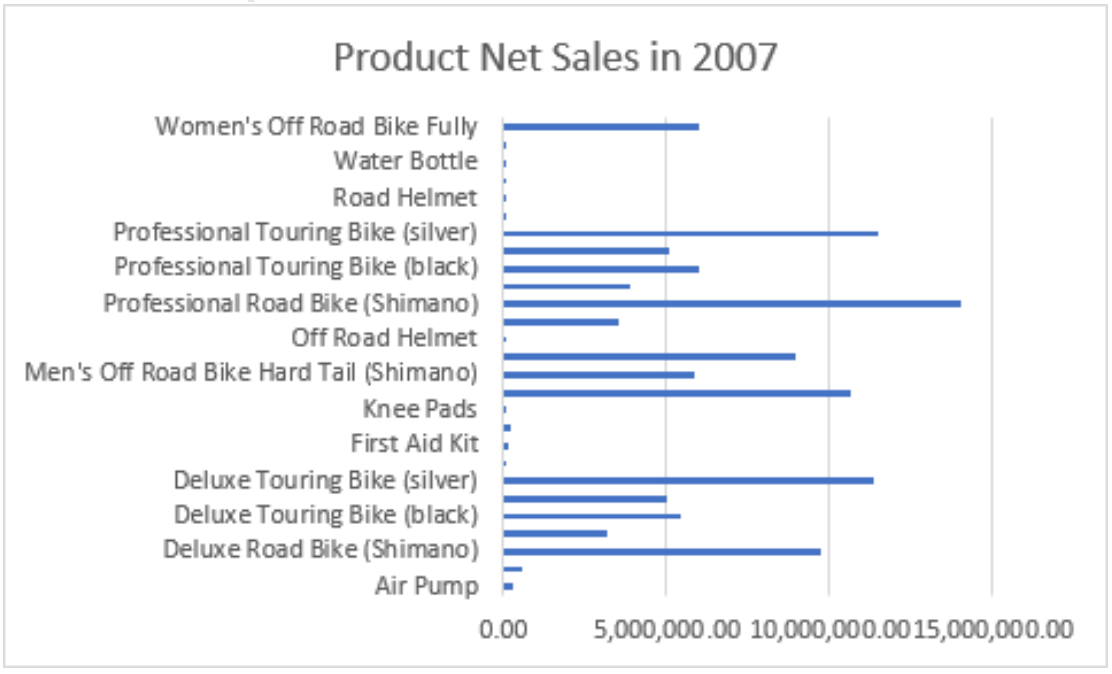


Figure 51

As can be seen from the chart, the company's revenue contribution from different products varies greatly. Professional Road Bike (Shimano) contributes the most, Professional Road Bike (silver) and Deluxe Touring Bike (silver) rank second and third respectively. Various accessories such as Water Bottle, Helmet, etc. account for a relatively small proportion of the company's sales revenue.

5.3.3 How did the company’s air pump sales revenue and expenses change from 2007 to 2019?

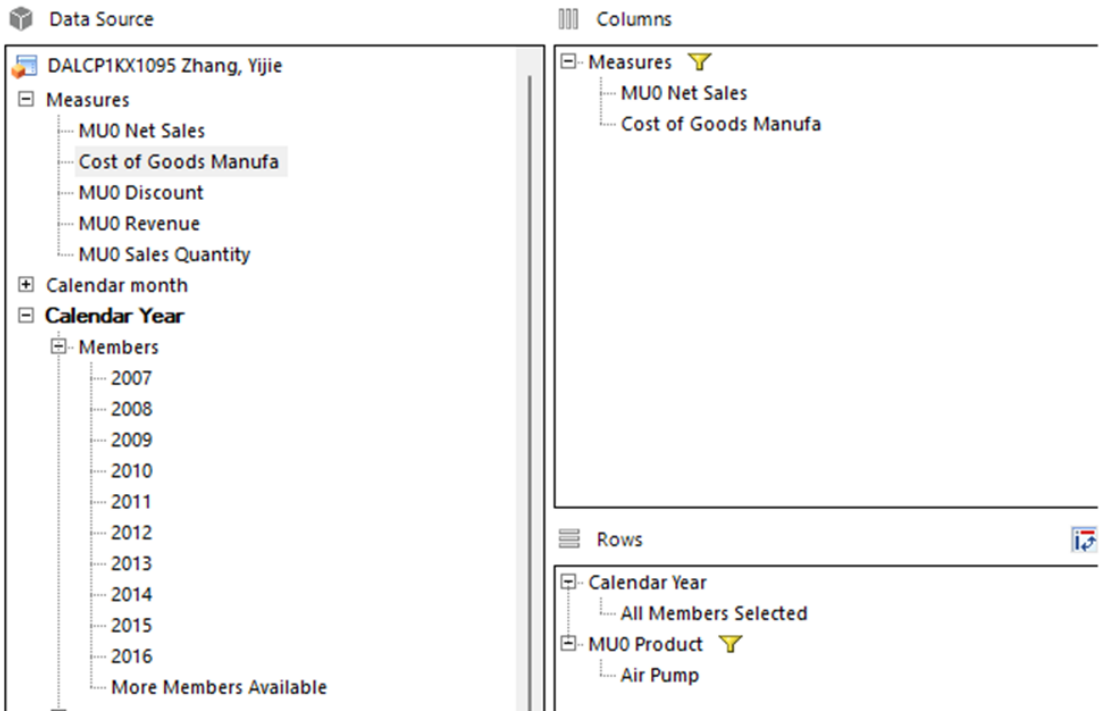


Figure 52

Drag ‘Net Sales’ and ‘CoGM’ to Measures. Drag ‘Calendar Year’ and ‘Air Pump’ in ‘Product’ to Rows. (Figure 52) Insert a Cluster Column table.

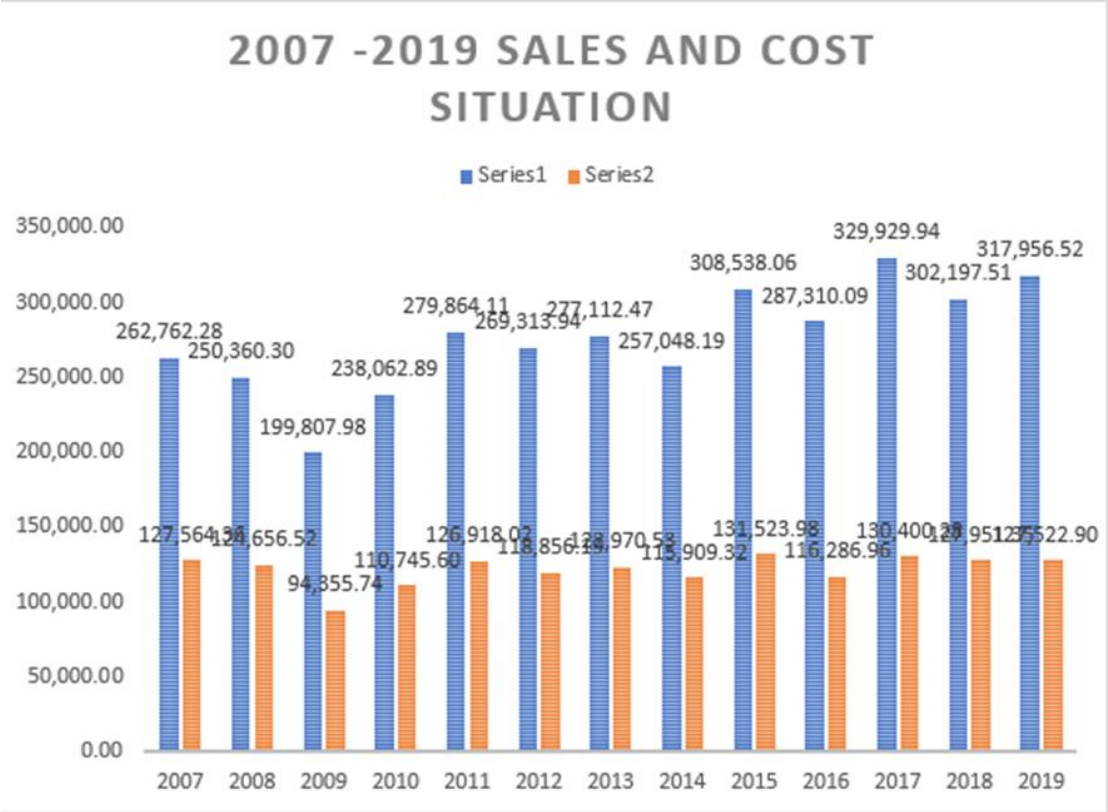


Figure 53

As shown in the Figure 53, from 2007 to 2019, the company's air pump sales revenue fluctuated, but generally showed an upward trend. At the same time, the cost did not

increase and remained around 120,000.

5.3.4 What Customer provided the highest Revenue in 2009?

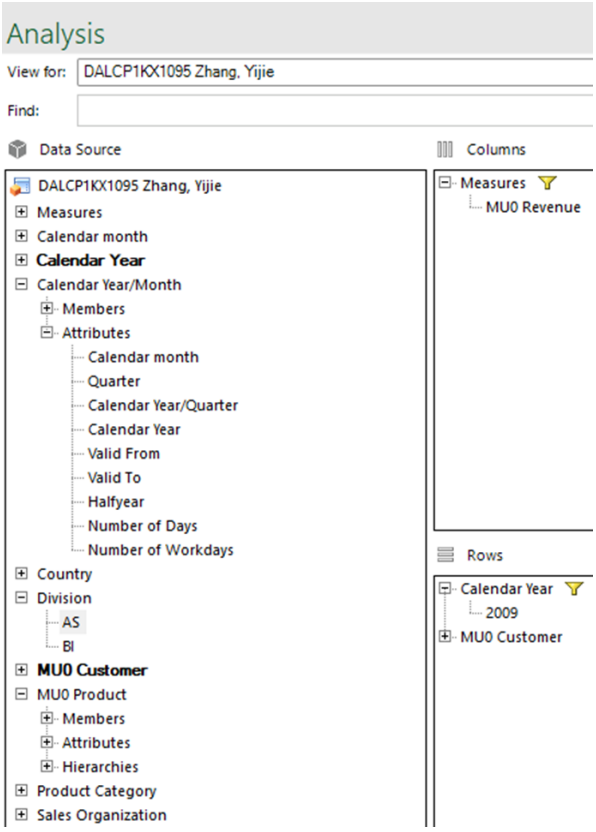


Figure 54

Drag 'Revenue' to Measures. Drag 2009 in 'Calendar Year' and 'Customer' to Rows. (Figure 54) Then insert an Area table. (Figure 55)

As shown in the figure, in 2009, Bavaria Bikes was the company's largest customer, with revenue of $12 million, far exceeding the company's other customers. In addition, there were two customers with revenue contributions of more than $6 million, and three customers with revenue contributions between $4 million and $6 million. The vast majority of customers contribute between $2 million and $4 million in revenue. Furniture City Bikes contributed the least revenue, less than $2 million.

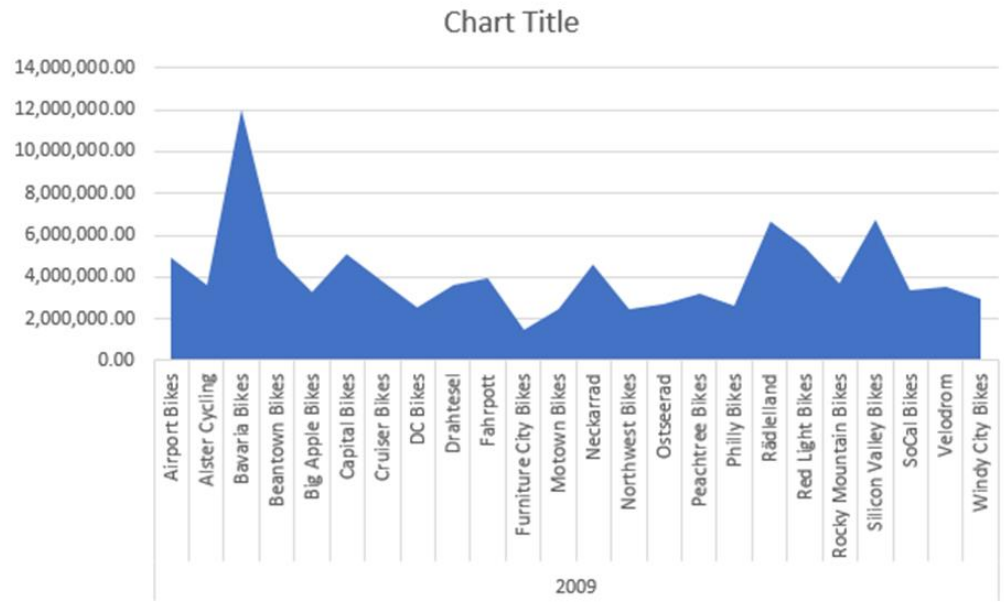


Figure 55

5.4 Personal Reflection and Conclusion

By using SAP Analysis for Microsoft Excel, I gained valuable insights into how this tool effectively integrates the power of SAP data analytics directly into Excel, combining the best of both environments.

SAP Analysis for MS Excel allows users to interact directly with live SAP datasets within the familiar Excel interface, significantly simplifying the process of data exploration and analysis. Its major advantage lies in combining Excel's familiarity and ease of use with SAP’s powerful backend data processing capabilities. Users can easily manipulate data through slicing, filtering, and drilling down without extensive technical training. This integration makes SAP Analysis particularly beneficial for users comfortable with Excel but needing more robust analytical capabilities provided by SAP's real-time data connections.

However, the tool has some limitations. Its dependence on Excel’s interface can limit the range and depth of advanced data visualizations and predictive analytics available. Additionally, although suitable for dynamic reporting and basic data exploration, the tool may not be the best choice for creating highly interactive or visually engaging dashboards compared to dedicated visualization platforms such as SAP Analytics Cloud or Tableau.

In summary, SAP Analysis for Excel is the perfect bridge between basic Excel functionality and advanced SAP analytics, especially for users who rely heavily on Excel for daily tasks. It is best suited for dynamic, real-time data analysis that requires the depth of SAP but the accessibility of Excel. The future choice between SAP Analysis for Excel, SAP Analytics Cloud, or Tableau should depend on specific analytical needs, complexity, user proficiency, and depth of insight required.

Chapter 6 Titanic Association Analysis

6.1 About Analysis

SAP Analytics Cloud (SAC) offers a robust platform for performing Cluster Analysis, a data mining technique that groups similar data points into clusters while ensuring distinct differences between groups. This method is particularly useful for tasks like market segmentation or customer behavior analysis, where multidimensional datasets are common. SAC simplifies the process by providing an intuitive interface and powerful algorithms, enabling users to uncover meaningful patterns even in complex, high-dimensional data that is difficult to visualize directly.

The process begins by importing a dataset into SAC, which can be sourced from Excel, databases, or cloud storage, containing variables such as customer purchase history or product sales metrics. Within SAC's analysis workspace, users can leverage features like "Smart Discovery" or "Predictive Analytics" to initiate clustering. A key step is selecting the number of clusters (K) for the K-means algorithm, which SAC uses to assign data points to clusters and calculate their centroids. The algorithm iteratively optimizes to minimize intra-cluster distances (differences within a cluster) while maximizing inter-cluster distances (differences between clusters), ensuring well-defined groups.

Once the analysis is complete, SAC presents the results through visualizations like scatter plots, heatmaps, or grouped tables, making it easy to interpret the characteristics of each cluster. For example, users can identify distinct customer segments based on purchasing patterns. SAC also provides explanatory insights, highlighting key variables that define each cluster. These results can be applied to business scenarios such as targeted marketing or customer segmentation, and SAC's integration capabilities allow users to combine clustering outcomes with other analytical tools for deeper insights, empowering data-driven decision-making.

6.2 Dataset Source and Research Questions

I continue and expand on what I learned by using SAP Analytics Cloud and Microsoft Excel to analyze survivability of the Titanic passenger. And answer the following questions:

1) Which rule occurs most frequently in the data set? What does this mean in the associate analysis?
2) Which rule would be considered the most important rule? Why?
3) What does the chart tell me about survivability on the Titanic?
4) What happens to the association analysis if confidence and support are increased or decreased

The data set used is Titanic.csv. This dataset appears to be a historical record of passengers from a maritime disaster, likely the Titanic, given the context of class, survival status, and demographics. It contains 2,201 entries, each representing an

individual passenger or crew member, and is structured with five columns: Passenger, Class, Sex, Age, and Survived. The tabular format makes it well-suited for analyzing survival patterns and demographic trends.

The "Passenger" column is a numerical identifier for each individual, ranging from 1 to 2,201, serving as a unique index. The "Class" column categorizes individuals into four groups: 1st, 2nd, 3rd (indicating passenger classes), and Crew, reflecting their social or occupational status on the ship. The "Sex" column is binary, listing each person as either Male or Female, while the "Age" column simplifies age into two categories: Child or Adult. Finally, the "Survived" column indicates the outcome for each individual, with values of Yes or No, showing whether they survived the disaster.

All columns in this dataset are categorical, making it ideal for statistical analysis, such as calculating survival rates across different classes, sexes, or age groups, or for performing cluster analysis to identify groups with similar characteristics. For example, one could explore whether first-class passengers had a higher survival rate compared to third-class passengers, or how survival differed between adults and children. The dataset's structure supports such multidimensional analysis, providing a foundation for understanding the factors influencing survival in this historical event.

6.3 Analysis tool application and results

Add data from the Tools menu and upload Titanic.csv. Change the aggregation of Passenger to Count. Duplicate the column Survived and rename it HRS. Highlight columns, Class, Sex, and Age to create a transformation LHS. Highlight the columns LHS and RHS to create a transformation Rule (Figure 56).

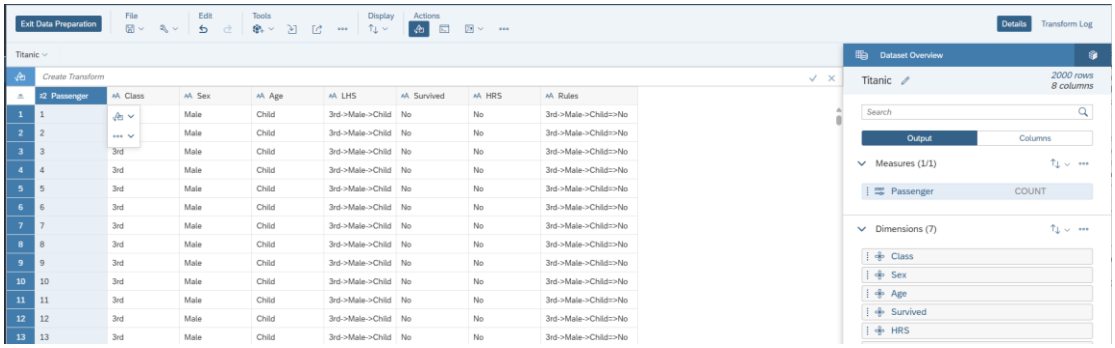


Figure 56

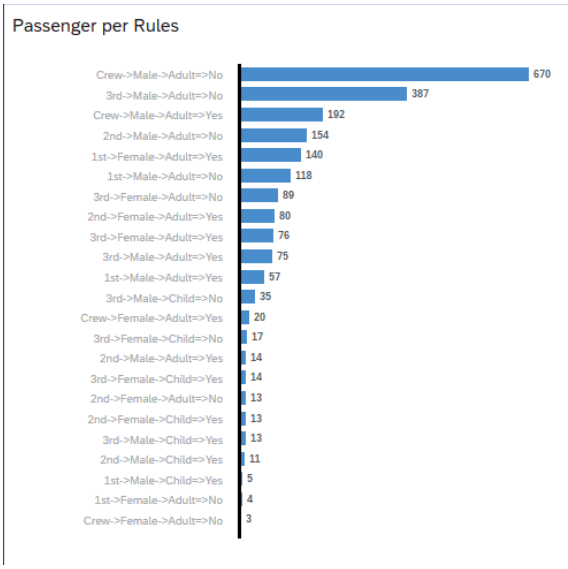


Figure 57

6.3.1 Which rule occurs most frequently in the data set? What does this mean in the associate analysis?

Exit Data Preparation. Add a bar chart using Passenger as the measure and Rules as the Dimension. Sort by Passenger count highest to lowest (Figure 57).

From the chart, the most frequent rule is "Crew->Male->Adult->No," with a count of 387 passengers. This means that the largest group in the dataset consists of male adult crew members who did not survive the Titanic disaster. This finding indicates a strong association between being a male adult crew member and not surviving.

6.3.2 Which rule would be considered the most important rule? Why?

Create three calculated dimensions Passenger by LHS, Passenger by RHS and Passenger by Rule. Create three measures from the dimensions Passenger by LHS Measure, Passenger by RHS Measure and Passenger by Rules Measure. Create the measures support, confidence and lift. Support = Passenger / grand total of Passenger. Confidence = Passengers by Rule Measure / Passengers by LHS Measure. Lift = Confidence / (Passenger by RHS Measure/grand total of Passenger). Format Confidence, Support, and Lift as 6 Decimal Places. Add a bar chart. Select Support, Confidence and Lift in Measures and Rules in Dimensions. Duplicate the bar chart and change it to a bubble chart (Figure 58).

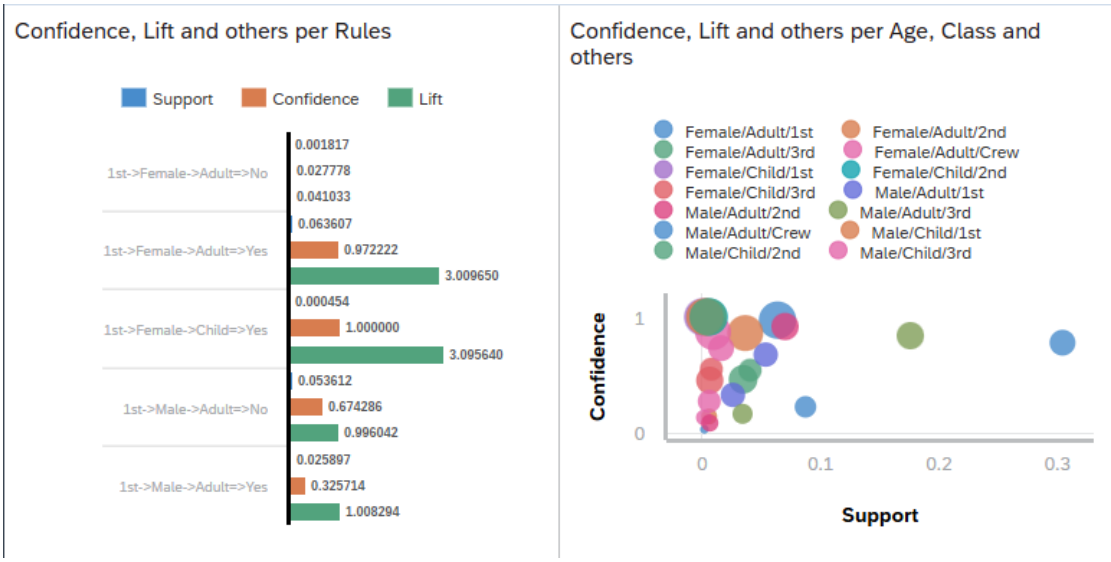


Figure 58

Analyzing these metrics, the most important rules are "1st->Female->Child->Yes" and "2nd->Female->Child->Yes," both with a Confidence of 1.0 and Lift values of 3.056

and 3.094, respectively, indicating a perfect association between being a female child in 1st or 2nd class and surviving, far beyond random chance.

6.3.3 What does the chart tell you about survivability on the Titanic?
Export the data with the calculated measures of support, confidence and lift to a new csv file. Add a Canvas page to Story and upload the new file. Create a bar chart with the same measures and dimension setting as the last question setting. Create a story filter to show Support > 0.01 and Confidence >0.80 (Figure 59).

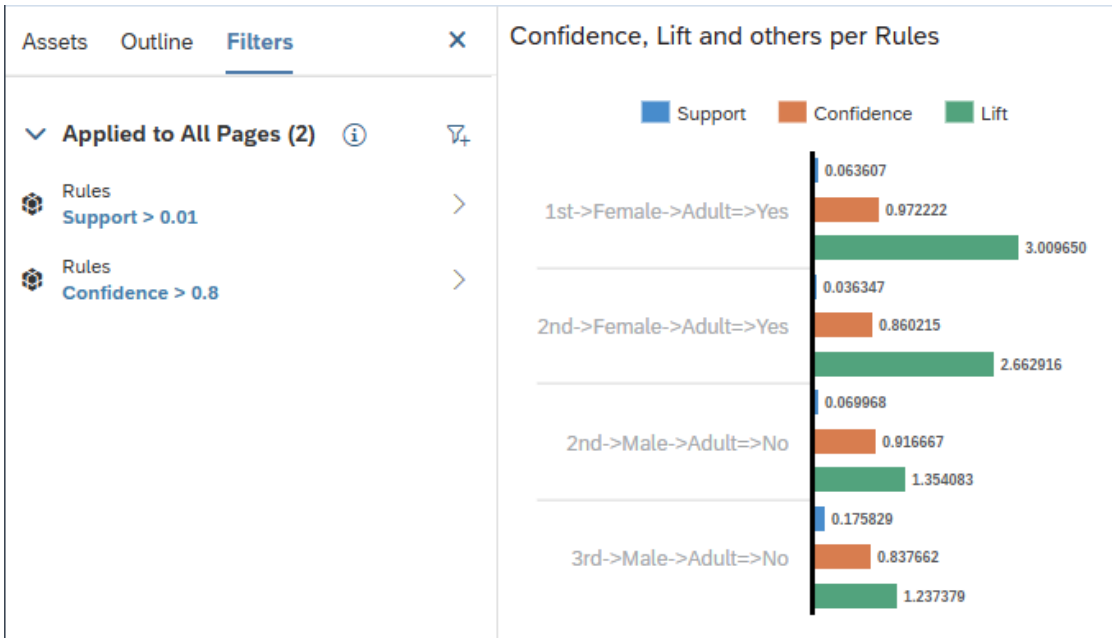
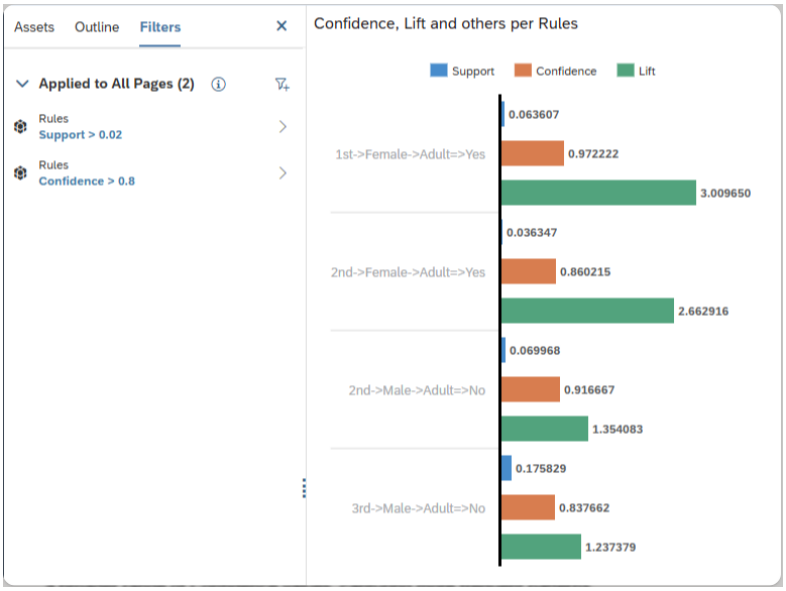


Figure 59

The chart reveals distinct patterns in survivability on the Titanic based on passenger characteristics. First, female adults in the 1st and 2nd classes had a very high likelihood of survival, with Confidence values of 0.972222 and 0.860215, respectively, and high Lift values (3.009650 and 2.662916). This indicates that being a female adult in a higher class (1st or 2nd) was strongly associated with survival, likely due to the "women and children first" evacuation policy and the better access to lifeboats for higher-class passengers.
Conversely, male adults in the 2nd and 3rd classes were highly likely not to survive, with Confidence values of 0.916667 and 0.837662, respectively. The Support for "3rd->Male->Adult->No" (0.175829) is the highest among the filtered rules, indicating that this was the most frequent outcome, reflecting the large number of 3rd-class male adults who perished.

6.3.4 What happens to the association analysis if confidence and support are increased or decreased?
Go to the story filters applied in the previous steps. Increase the Support threshold from 0.01 to 0.02. Increase the Confidence threshold from 0.8 to 0.9 and 0.8 to 0.5 (Figure 60).


Assets
Outline
Filters
Applied to All Pages (2)
Rules
Support > 0.02
Rules
Confidence > 0.8
Confidence, Lift and others per Rules
Support
Confidence
Lift
1st->Female->Adult=>Yes
0.063607
0.972222
3.009650
2nd->Female->Adult=>Yes
0.036347
0.860215
2.662916
2nd->Male->Adult=>No
0.069968
0.916667
1.354083
3rd->Male->Adult=>No
0.175829
0.837662
1.237379


Assets
Outline
Filters
Applied to All Pages (2)
Rules
Support > 0.01
Rules
Confidence > 0.5
Confidence, Lift and others per Rules
Support
Confidence
Lift
1st->Female->Adult=>Yes
0.063607
0.972222
3.009650
1st->Male->Adult=>No
0.053612
0.674286
0.996042
2nd->Female->Adult=>Yes
0.036347
0.860215
2.662916
2nd->Male->Adult=>No
0.069968
0.916667
1.354083
3rd->Female->Adult=>No
0.040436
0.539394
0.796783
3rd->Male->Adult=>No
0.175829
0.837662
1.237379
3rd->Male->Child=>No
0.015902
0.729167
1.077111
Crew->Male->Adult=>No
0.304407
0.777262
1.148157

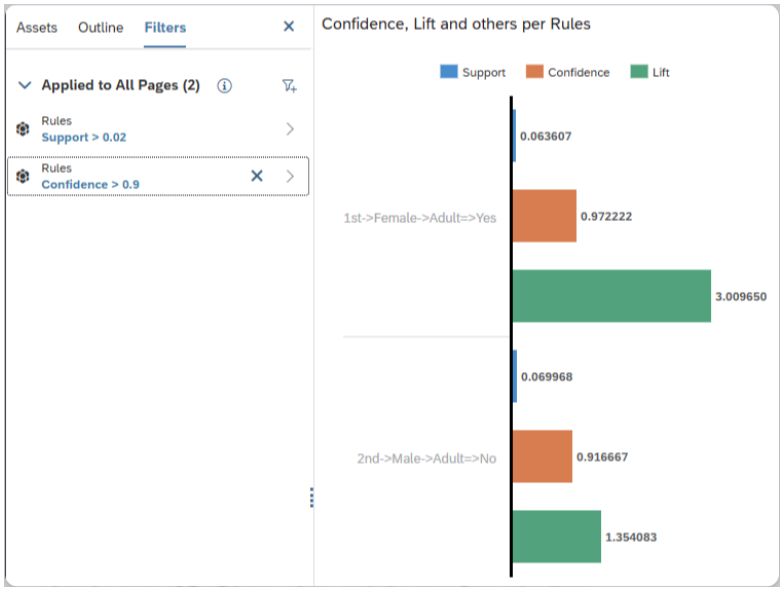

Assets
Outline
Filters
Applied to All Pages (2)
Rules
Support > 0.02
Rules
Confidence > 0.9
Confidence, Lift and others per Rules
Support
Confidence
Lift
1st->Female->Adult=>Yes
0.063607
0.972222
3.009650
2nd->Male->Adult=>No
0.069968
0.916667
1.354083

Figure 60

When both Confidence and Support filters are increased, the association analysis becomes more restrictive, resulting in fewer rules being displayed. In the first chart (Support > 0.01, Confidence > 0.5), eight rules are shown, capturing a wide range of patterns, including less reliable ones like "Crew->Male->Adult->No" with a Confidence of 0.777262. Increasing the Support to > 0.02 (second chart) removes rules with lower frequency, such as "3rd->Male->Child->No" (Support: 0.015902), reducing the number of rules to four. Further increasing the Confidence to > 0.9 (third chart) filters out rules with Confidence below 0.9, such as "2nd->Female->Adult->Yes" (Confidence: 0.860215), leaving only the two most reliable rules: "1st->Female->Adult->Yes" and "2nd->Male->Adult->No." This progressive filtering focuses the analysis on the most frequent and reliable associations, highlighting only the strongest patterns—female adults in 1st class surviving and male adults in 2nd class not surviving—while excluding less frequent or less certain rules, thus sharpening the focus on the most significant insights from the Titanic dataset.

6.4 Personal Reflection and Conclusion

Through the Titanic association analysis using SAP Analytics Cloud (SAC) and Microsoft Excel, I gained a deeper understanding of how SAC facilitates complex data mining tasks like association analysis. SAC proved to be a powerful platform for uncovering patterns in the Titanic dataset, such as the strong association between being a female in higher classes and surviving, or the high likelihood of male adults in lower classes not surviving. Its intuitive interface and robust features, like the ability to calculate Support, Confidence, and Lift, allowed me to efficiently create and filter association rules, while the visualization tools (bar and bubble charts) made it easy to interpret multidimensional data. The integration of predictive analytics and data preparation within SAC streamlined the process, enabling me to focus on deriving meaningful insights rather than wrestling with technical complexities.

However, SAC's data preparation phase, such as creating transformations and calculated measures, was somewhat intricate and required careful attention to detail, similar to the challenges noted in Chapter 3 with SAC's data modeling. Compared to Tableau (Chapter 4), which excels in quick, visually appealing storytelling, SAC offers more depth in analytical capabilities but demands a steeper learning curve for advanced tasks. In contrast to SAP Analysis for MS Excel (Chapter 5), SAC provides superior visualization and predictive features, but it lacks the seamless familiarity of Excel's interface for users accustomed to simpler workflows. In conclusion, SAC is best suited for in-depth, multidimensional analyses like association or cluster analysis, where its advanced features can uncover significant patterns in complex datasets. For future analyses, I would choose SAC for tasks requiring detailed rule-based insights or predictive modeling, while opting for Tableau for rapid visualizations or SAP Analysis for Excel when working within a familiar Excel environment, ensuring the tool aligns with the specific analytical needs and user proficiency.

Chapter 7 Text Analysis with Wine Description Data

7.1 About Analysis

Text Analytics is a critical process in data analysis that involves extracting meaningful insights from unstructured textual data, a skill increasingly essential for data analysts as the volume of such data grows. This process enables the understanding of patterns, sentiments, and trends within text, which can inform strategic decision-making, such as improving customer experiences or developing targeted recommendations. The workflow typically begins with importing a text dataset, such as customer reviews or social media posts, into an analytical platform. Tools like Tableau and Voyant are commonly used for their robust text analysis and visualization capabilities. In Tableau, users can connect to the dataset, preprocess the text by cleaning and tokenizing it, and then apply techniques like sentiment analysis to gauge emotional tone or bag-of-words to identify frequent terms. Voyant complements this by offering detailed text exploration features, such as word clouds, frequency lists, and contextual analysis, allowing users to visualize word distributions and relationships. Together, these tools help transform raw text into actionable insights, enabling analysts to uncover customer behavior patterns and craft data-driven strategies effectively.

7.2 Dataset Source and Research Questions

I continue and expand on what I learned by using Tableau to analyze wine reviews. And answer the following questions:

1) How do average prices and ratings differ by continent and country?
2) Do the price of a wine is significantly related to the review points of a wine?
3) What is the sentiment of the comments?
4) What do the top 10 positive reviews look like?

The dataset utilized in this analysis is Wine_Description_3Continents_W2025_New.csv, which focuses on wine reviews, providing a comprehensive collection of data for text analytics and customer behavior studies. It comprises two subsets: the original dataset, sourced from a data portal, includes approximately 130,000 wine reviews with attributes such as country, province, description, rating points, price, title, variety, and winery. A reduced subset, specifically prepared for this exercise, contains 6,000 sampled data points from the original set, with an additional column for continent to facilitate geographic analysis. This smaller dataset, while not fully representative of the country proportions in the original, retains key dimensions like wine descriptions, which are used for sentiment analysis. Polarity scores for these descriptions were calculated using a natural language processing tool, enabling the evaluation of sentiment in each review. This dataset is well-suited for exploring customer preferences, regional wine characteristics, and sentiment trends through text-based visualizations and analytical techniques.ues.

7.3 Analysis tool application and results

7.3.1 How do average prices and ratings differ by continent and country?

Select and open Wine_Description_3Continents_W2025_New.csv. Go to Worksheet, drag the Continent dimension and drop it above Country within the "Country, Province Hierarchy." Rename the hierarchy to Location. Change the aggregation of Points and Price to Average. Drag Continent, Country to the Columns. Drag Price and Points to the Rows. Sort the chart by Price at the Country level. Drag Country to Detail. Drag Price to the Color Marks (Figure 61).

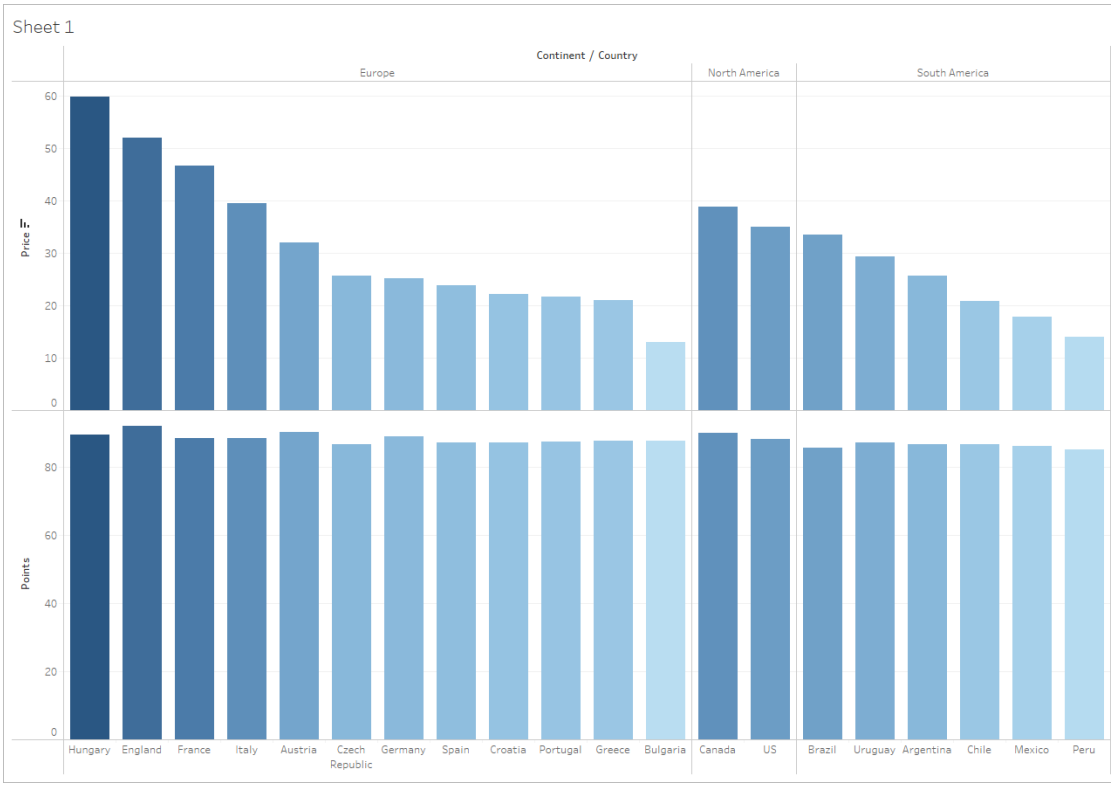


Figure 61

Average prices of wines differ significantly by continent and country, with Europe showing the largest variation (20-55 USD), followed by South America (15-35 USD), and North America displaying the least variation (30-35 USD). In contrast, average ratings are much more consistent, ranging from 85 to 90 points across all regions, with differences within continents being minimal (3-5 points). The lack of a strong correlation between price and rating indicates that factors other than perceived quality—such as brand reputation, production costs, or market dynamics—may drive pricing differences, while wine quality remains relatively uniform across the dataset.

7.3.2 Do the price of a wine is significantly related to the review points of a wine?

Drag Price to the Columns shelf. Drag Points to the Rows shelf. Drag Country to the Color Marks card to differentiate countries by color. Drag Province to the Marks card, placing it just below "Country," to add a more granular level of detail. Adjust the Y-Axis Scale and set the range from 80 to 100 (Figure 62).

Based on the scatter plot, the price of a wine does not appear to be significantly related to its review points. The visualization shows a wide distribution of review points across all price ranges, with no clear trend indicating that higher prices consistently result in higher ratings. For instance, wines priced between 20 and 50 USD often achieve ratings

of 88-92 points, which are comparable to ratings of wines priced above 80 USD. Similarly, some low-priced wines (below 20 USD) score as high as 90-92 points, while some higher-priced wines (50-70 USD) score as low as 85 points. This lack of a distinct upward trend suggests that factors other than price—such as wine quality, variety, or reviewer preferences—play a more significant role in determining review points. Therefore, while there are instances where higher-priced wines receive higher ratings, the overall relationship between price and review points is not significant, as the scatter plot indicates a weak correlation at best.

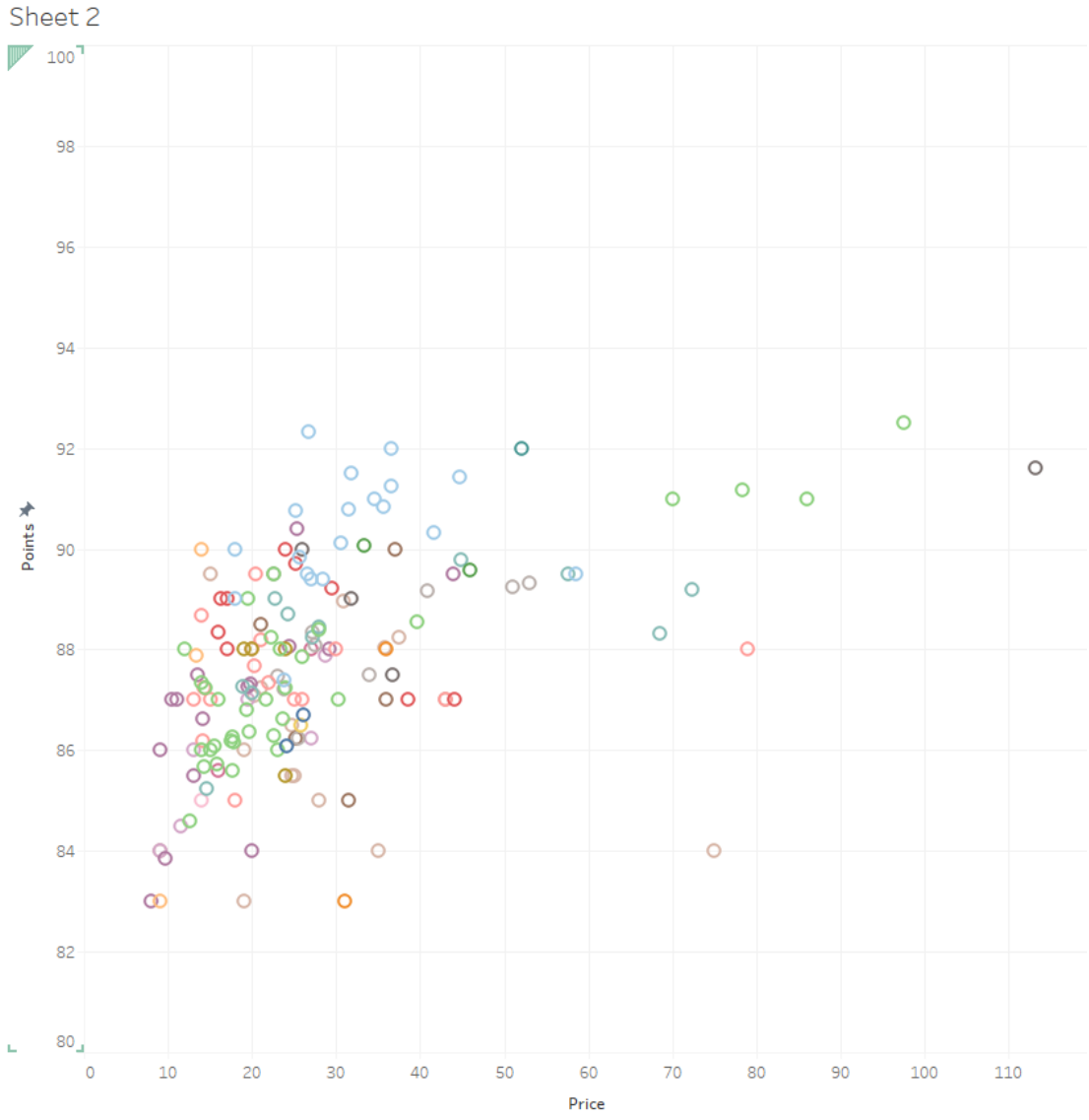


Figure 62

7.3.3 What is the sentiment of the comments?

Go to https://communalytic.org/ and import dataset. Set "Skip Field" under "Reviewer" to "user_id." and "Skip Field" under "description" to "Text." Click Sentiment Analyzer to start Analysis (Figure 63, 64, 65).

Based on the analysis of **6000** out of 6000 posts, the results are as follows. **As per request, all posts were assumed to be in English.**

	# of Posts	Negative Sentiment [-1..-0.05]	Neutral Sentiment (-0.05..0.05)	Positive Sentiment [0.05..1]
VADER (English/EN)	6000	614 (10.23%)	706 (11.77%)	4680 (78.00%)
TextBlob (English/EN)	6000	809 (13.48%)	975 (16.25%)	4216 (70.27%)

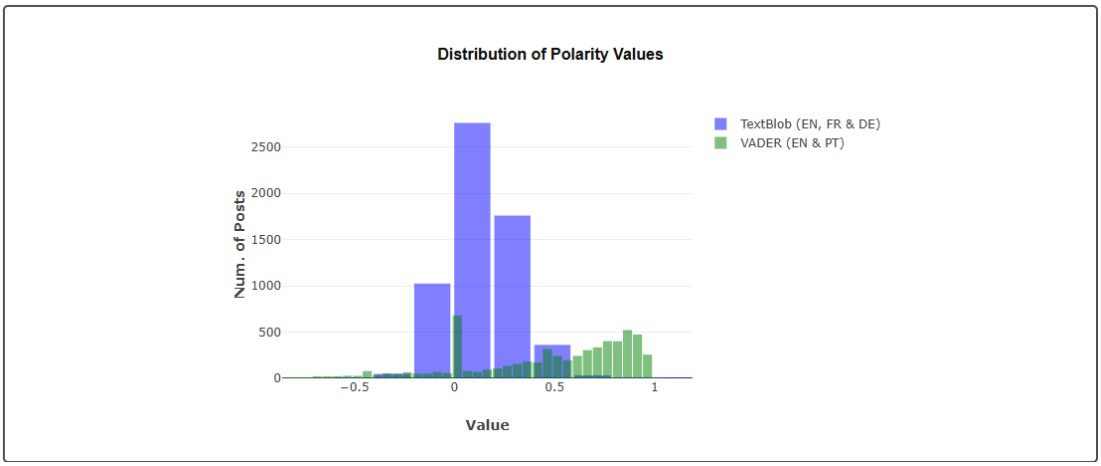


Figure 63

Comparing Results between VADER and TextBlob (for posts in **English**)

Excluding duplicates (such as reposts/retweets), VADER and TextBlob agree on how to categorize **4215 (70.73%)** out of **5959** English language posts. This agreement level is considered to be **fair** (Cohen's kappa statistic = **0.302**;)

Specifically, both libraries agree on:

- **261 (6.19%)** posts with negative sentiments (polarity scores <= -0.05),
- **210 (4.98%)** posts with neutral sentiments (polarity scores between -0.05 and 0.05), and
- **3744 (88.83%)** posts with positive sentiments (polarity scores >= 0.05).
- These cases are shown with the green background in the confusion matrix below.

To determine which of the two sentiment analysis algorithms is better suited/more accurate at detecting polarity of posts in your dataset, we suggest examining all or a sample of the polarity scores produced by both libraries to cross-validate results.

Start by **downloading** the full dataset, and then use either Excel or Google Sheet to manually review (all or a sample of) cases where VADER and TextBlob disagree, especially in cases where posts were assigned opposite polarity scores.
For your information, based on the analysis of your dataset, there are:

- **310** posts when VADER assigned positive polarity scores and TextBlob assigned negative, and
- **187** posts when VADER assigned negative polarity scores and TextBlob assigned positive.
- These cases are shown with the red background in the confusion matrix below.

Note: When you download the dataset with the polarity scores, the 'vader_sentiment_compound' column will contain VADER's polarity score and the 'textblob_polarity' column will contain TextBlob's polarity score.

Figure 64

Confusion Matrix (excluding duplicates)

The following table shows both agreement and disagreement counts across sentiment labels as determined by VADER and TextBlob.

	VADER - Negative [-1..-0.05]	VADER - Neutral (-0.05..0.05)	VADER - Positive [0.05..1]
TextBlob - Negative [-1..-0.05]	261	235	310
TextBlob - Neutral (-0.05..0.05)	162	210	595
TextBlob - Positive [0.05..1]	187	255	3744

Figure 65

The sentiment of the wine review comments is overwhelmingly positive, with 70-78% of the comments expressing favorable opinions, as determined by both VADER and TextBlob. A smaller portion of comments are neutral (11.77-16.25%) or negative

(10.23-13.48%), indicating that while most reviews are positive, there is a notable minority expressing neutral or critical views. The consistency between VADER and TextBlob in identifying a large positive sentiment, despite some disagreements, reinforces the overall positive tone of the comments.

7.3.4 What do the top 10 positive reviews look like?
Import Wine_Description_3Continents_W2025_SentimentAnalyzed. Csv in Tableau. Drag Compound Sentiment to the Rows and Description to the Columns. Drag Description to the Filters shelf and set to "Top 10 by Compound Sentiment Sum." Select "Bubble Chart." (Figure 66)

Figure 66

The top 10 positive reviews, as shown in the bubble chart, are characterized by overwhelmingly favorable descriptions of the wines, focusing on their appealing flavors, balanced structures, and pleasant finishes. These reviews frequently highlight fruity notes such as cherry, plum, berry, strawberry, and grapefruit, often describing the wines as "fresh," "smooth," or "crisp." For example, one review notes a wine with "ripe

berry aromas mixed with touches of minerality and latex" and a "sweet, lightly chocolate finish," while another praises a wine for its "fine textured acidity along with a pink grapefruit flavor." Many reviews also emphasize the wines' balanced structure, using terms like "crisp acidity," "balanced," and "smooth palate," indicating they are well-rounded and enjoyable. The finishes are often described as "sweet," "fruity," or "chocolatey," enhancing the overall positive impression. Some reviews mention versatility, noting the wines' suitability for immediate drinking or food pairing, such as with salmon, while others add notes of complexity with descriptors like "leathery," "earthy," or "toasty oak."

7.4 Personal Reflection and Conclusion

Through the text analysis of wine reviews in Chapter 7 using Tableau, I gained valuable insights into the capabilities of this tool for handling unstructured textual data, while also reflecting on its strengths and limitations in comparison to the tools explored in previous chapters. Tableau proved to be highly effective for visualizing and interpreting the wine review dataset, allowing me to uncover patterns such as the predominantly positive sentiment of the reviews and the lack of a strong correlation between price and review points. Its intuitive drag-and-drop interface made it easy to create visualizations like scatter plots, histograms, and bubble charts, which provided clear insights into customer preferences and regional pricing trends. The ability to integrate sentiment analysis results from Communalytic and visualize them in Tableau was particularly powerful, enabling me to quickly identify the top 10 positive reviews and understand their characteristics through descriptive text snippets.

However, the process also highlighted some challenges. Preparing the data for sentiment analysis required external tools like Communalytic, as Tableau's native capabilities for advanced text processing and sentiment analysis are limited. This dependency on external tools mirrors the data preparation complexities I encountered with SAP Analytics Cloud (SAC) in Chapters 3 and 6, though Tableau's visualization process was more straightforward. Compared to SAC, which excels in predictive analytics and multidimensional analysis (as seen in the Titanic association analysis in Chapter 6), Tableau focuses more on rapid, visually appealing storytelling, aligning with my observations in Chapter 4. In contrast to SAP Analysis for MS Excel (Chapter 5), Tableau offers superior visualization capabilities but lacks the seamless integration with Excel's familiar interface for users who prefer a spreadsheet environment. Additionally, while Tableau handled the visualization of the 6,000-row dataset efficiently, it required careful setup to ensure accurate filtering and aggregation, similar to the data cleaning challenges I faced with FAOSTAT data in Chapter 2 using Excel.

In conclusion, Tableau is an excellent choice for transforming text data into actionable visual insights, particularly for tasks requiring quick, interactive visualizations and customer behavior analysis, as demonstrated in this chapter. However, its reliance on external tools for advanced text analytics and the need for precise data preparation underscore the importance of selecting the right tool based on the task's complexity.

For future projects, I would use Tableau for rapid visualization and storytelling, SAC for in-depth predictive and multidimensional analyses, and Excel or SAP Analysis for MS Excel for simpler, spreadsheet-based tasks, ensuring the tool aligns with the analytical goals and my proficiency level. This experience reinforced the value of combining multiple tools to leverage their unique strengths, ultimately enhancing the quality and impact of data-driven insights.

Chapter 8 Summary

This document chronicles a comprehensive journey through various data analysis tools and techniques, applied across diverse datasets to address a range of business and environmental questions. Each chapter focuses on a specific tool—Microsoft Excel Pivot Tables, FAOSTAT with Tableau, SAP Analytics Cloud (SAC), Tableau, SAP Analysis for MS Excel, and text analysis with Tableau—demonstrating their unique capabilities and limitations in handling structured and unstructured data. The analyses span sales performance, environmental impact, survival patterns, and customer sentiment, providing a holistic view of how these tools can be leveraged for data-driven decision-making.

In Chapter 1, I used Excel Pivot Tables to analyze Global Bike Inc.'s sales data, revealing trends such as an overall revenue increase from 2007 to 2016, with a peak in 2014, and identifying key products like PRTR 2000 as major revenue contributors. The analysis also highlighted seasonal patterns, with June as the peak sales month, driven by products like PRDR 1000. Excel's simplicity and accessibility were evident, but its limitations in handling large datasets and advanced analytics were noted.

Chapter 2 explored FAOSTAT data to assess economic and environmental trends, using Excel for data cleaning and Tableau for visualization. The analysis showed varied economic resilience in Asian countries during the pandemic, significant differences in CO2 emissions from food packaging and industrial wastewater across countries, and no clear correlation between agricultural water use efficiency and stress levels. FAOSTAT's comprehensive data was valuable, but required extensive cleaning, and Excel's limitations in processing large datasets were apparent.

In Chapter 3, SAC was applied to Global Bike Inc.'s sales data, uncovering geographic revenue concentrations in the U.S. (e.g., California), seasonal revenue peaks in June, and customer contributions to revenue in 2023, with Bavaria Bikes leading and gross profit margins mostly between 40-50%. SAC's predictive and visualization capabilities were powerful, though its complex data preparation phase posed challenges.

Chapter 4 utilized Tableau to analyze Global Bike Inc.'s sales and global CO2 emissions, identifying 2014 as the highest revenue year (140M USD), 2016 as the peak gross margin year (over 40M USD in Germany), and trends in CO2 emissions, with China's emissions rising sharply from 1994 to 2011. Per capita CO2 emissions in 2011 varied significantly, with higher emissions in Europe and North America. Tableau's visualization strengths were evident, but its limited predictive capabilities were noted.

Chapter 5 employed SAP Analysis for MS Excel to further analyze Global Bike Inc.'s sales, showing a revenue increase from 120M to 160M USD from 2017 to 2019, Professional Road Bike (Shimano) as the top revenue contributor in 2007, and stable air pump sales costs despite revenue growth from 2007 to 2019. Bavaria Bikes was the

largest customer in 2009. The tool's integration with Excel was seamless, but its visualization capabilities were limited compared to SAC and Tableau.

Chapter 6 used SAC for association analysis on the Titanic dataset, identifying "Crew->Male->Adult->No" as the most frequent rule and "1st->Female->Child->Yes" as the most important due to its perfect confidence. The analysis highlighted survival disparities, with higher-class females more likely to survive, and showed how increasing support and confidence thresholds in association analysis focuses on the most reliable rules. SAC's analytical depth was impressive, though data preparation was complex.

Chapter 7 focused on text analysis of wine reviews using Tableau, revealing significant price variations across continents (Europe: 20-55 USD) but consistent ratings (85-90 points), no strong price-rating correlation, predominantly positive sentiment (70-78%), and top positive reviews emphasizing fruity flavors and balanced structures. Tableau excelled in visualization, but required external tools like Communalytic for sentiment analysis, highlighting its limitations in advanced text processing.

Overall, this document demonstrates the importance of selecting the right tool for specific analytical tasks. Excel Pivot Tables and SAP Analysis for MS Excel are ideal for structured data and spreadsheet-based analysis, offering accessibility but limited scalability. SAC provides depth for predictive and multidimensional analyses, though with a steeper learning curve. Tableau shines in rapid, visually appealing storytelling, particularly for text and structured data, but lacks advanced analytical features. FAOSTAT's rich data supports global insights, requiring careful cleaning. Each tool's strengths and weaknesses underscore the need for a strategic approach to tool selection, balancing complexity, user proficiency, and analytical goals to maximize the impact of data-driven insights. This journey has equipped me with a nuanced understanding of these tools, preparing me to tackle diverse analytical challenges effectively in future scenarios.