Table of Contents

Chapter 4: ERPSim – SAC

Chapter 5: Data Wrangling and Analysis using AQUASTAT and FAOSTAT

Chapter 6: SAP Analytics Cloud for Unsupervised Learning

Chapter 7: Tableau Using Text Analysis

Chapter 1 Microsoft Excel Pivot Tables

1.1 Introduction to Excel Pivot Tables

Microsoft Excel is a spreadsheet program which was first introduced by Microsoft in 1985. Over the years, Excel has evolved into one of the most widely used spreadsheet applications globally, becoming a standard tool for businesses, academics, and individuals for data analysis, financial calculations, and more. Pivot Tables were introduced in Excel 5.0, which was released in 1993. This version marked a significant step forward in Excel's capabilities, introducing several new features, with the Pivot Table being one of the most notable. Pivot Tables provided users with a powerful tool for analyzing and summarizing large datasets in a dynamic and user-friendly way by transforming rows into columns and columns into rows. It enables users to transform raw data into meaningful insights by creating dynamic tables.

1.2 Research Questions and Data Set Used

Using the examples and lecture of Chapter 1 as a reference, I have used Excel pivot tables to analyze the sales data given as a .xlsx file. The data is from the years 2007-2011 for a fictitious company called GBI. The file includes 15 columns of variables including customer, product, date, revenue, sales quantity etc. Since the dataset was complete without any missing data or duplicates, no cleaning was required, and the file was ready to be used.

Using pivot tables, the below questions based on the transformations will be addressed:

1. What was the overall revenue? Which year had the highest revenue and what was the revenue during that year?
2. What is the Revenue for E-Bike Trailwind in 2009? Is there any revenue?
3. In the year that had the highest Net Sales, what division had the highest Net Sales?
4. What customer provided the highest revenue for Accessories (Division AS) in 2009?
5. Is there seasonality in revenue during the year? If so, what month has the highest revenue? Is the seasonality similar from year to year?

1.3 Analysis using Pivot Tables and Outcomes

1.3.1 What was the overall revenue? Which year had the highest revenue and what was the revenue during that year?

A pivot table for the given data was inserted as a new sheet. Using the Pivot Table Builder, Revenue was added to Values and Calendar Year was added to Rows. Revenue was set to Dollars and the Row Label was sorted in descending order by Sum of Revenue.

Row Labels	Sum of Revenue
2011	$5,66,25,742.94
2010	$5,58,54,415.97
2009	$5,26,10,815.06
2008	$5,94,44,067.11
2007	$6,07,15,831.69
Grand Total	**$28,52,50,872.77**

Figure 1.1: Overall Revenue

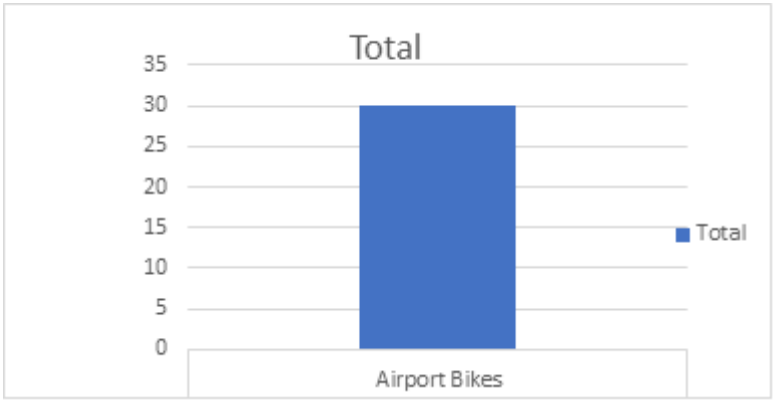


Figure 1.2: Bar Graph of the Overall Revenue

According to Figure 1 & 2, the overall revenue for GBI is $ 28,52,50,872.77. The highest revenue earned was in the year 2007, which was $ 6,07,15,831.69.

1.3.2 a) What is the Revenue for E-Bike Trailwind in 2010?

Using the Pivot Table Builder, add Material Desc to the Rows. Using the dropdown arrow next to Row Label, select field as Material Desc and filter the material to E-Bike Trailwind.

Row Labels	Sum of Revenue
2010	$20,24,435.69
E-Bike Tailwind	$20,24,435.69
2011	$19,98,759.52
E-Bike Tailwind	$19,98,759.52
Grand Total	$40,23,195.21

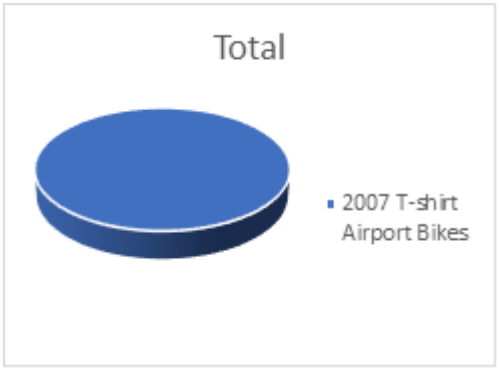


Figure 1.3: Total revenue of Trailwind Figure 1.4: Pie Chart of Trailwind Revenue

According to Figures 3 & 4, the revenue for E-Bike Trailwind in 2010 is $ 20,24,435.69.

b) What is the Revenue for E-Bike Trailwind in 2009? Is there any revenue?

According to Figures 3 & 4, the revenue for E-Bike Trailwind in 2009 was $ 0.

1.3.3 What is the Sales Quantity of T-shirts for Airport Bikes in 2007?

Row Labels	Sum of Sales Quantity
2007	30
T-shirt	30
Airport Bikes	30
Grand Total	30

Figure 1.5: Sales Quantity for T-Shirts for Airport Bikes in 2007

Using the Pivot table builder, Remove Material Desc and Customer Desc. Sort the Calender year in descending order of sales quantity to get the total T-Shirt sales for Airport Bikes in 2007 which is 30 units.

1.3.4 What year had the highest Sales Quantity?

Row Labels	Sum of Sales Quantity
2007	37,537
2008	36,088
2010	32,237
2011	31,880
2009	31,112
Grand Total	1,68,854

Figure 1.6: Yearly Sales Quantity

Using the Pivot Table Builder, add Revenue and sort the Calendar Year in ascending order of Revenue to get the highest Sales Quantity. The year with the highest sales quantity is 2008.

1.3.5 What year had the lowest Revenue?

Row Labels	Sum of Revenue
⊞ 2009	52610815.06
⊞ 2010	55854415.97
⊞ 2011	56625742.94
⊞ 2008	59444067.11
⊞ 2007	60715831.69
Grand Total	285250872.8

Figure 1.7: Each Years Revenue

Using the Pivot Table Builder, add Material Desc to the rows section and sort the Calender Year in ascending order of Revenue followed by sorting the Material Desc in ascending order of Revenue. The year with the lowest revenue is 2009.

1.4 Advantages and Disadvantages of the Tool

Though I have been using Excel all my life, I have never used pivot tables for data analysis before. Since pivot tables are widely used for analyzing data, this skill would be a great addition to my set of skills as a former data engineer.

Some advantages of Pivot tables are that they are easy to learn and use as it is straightforward. The user can drag and drop fields into 4 categories to organize and get insights from the dataset without having any technical skills. Another advantage is its flexibility, where the users can rearrange the data fields and gain different perspectives of the data thereby making business decision-making time efficient and easy to understand.

Some disadvantages of Excel pivot tables are that it is limited only to structured data specifically tabular data. If the data is not well-structured and contains complex relationships that go beyond rows and columns or if the user has an older version of Excel, then pivot tables cannot be used. Though learning basic pivoting techniques is easy, mastering complex techniques requires proper training and experience.

1.5 Conclusion

To sum it up, Pivot Tables in Microsoft Excel emerge as a robust and adaptable ally for dissecting data, granting users the prowess to condense extensive datasets swiftly and dynamically. Their merits span flexibility, interactive capabilities, time-saving efficiency, user-friendliness, and a wealth of customization choices. Yet, it's crucial to be aware of potential hurdles like the learning curve, reliance on precise source data, and limitations when dealing with non-tabular data or intricate relationships. Nevertheless, mastering Pivot Tables bestows users with the authority to unveil valuable insights, facilitate judicious decision-making, and showcase data in a lucid and orderly manner. This solidifies their status as an indispensable asset for adept data analysis within the Excel realm.

Chapter 2 SAP Analytics Cloud

2.1 Introduction to SAP Analytics Cloud

SAP Analytics Cloud (SAC) is an integrated cloud platform offered by SAP that combines business intelligence, planning, and predictive analytics functionalities. Launched in 2015, it serves as a comprehensive solution for organizations seeking to analyze, visualize, plan, and make data-driven decisions within a unified environment.

SAC enables users to connect to various data sources both on-premises and cloud databases. It enables its users to create insightful and interactive visualizations that support real-time analytics for business decision-making.

SAC goes beyond traditional BI by providing robust planning and budgeting capabilities. Users can collaborate on planning scenarios, forecast data, and align business objectives seamlessly.

The platform incorporates predictive analytics tools which allow organizations to identify trends, patterns, and potential future outcomes based on historical data.

2.2 Dataset Used and Main Features of SAC Modeler View

The dataset used is the sales data from a fictitious company called Global Bikes. The file (GB_AnalyticsData.xlsx) is used to examine past sales performance and come up with strategies for the future. It is a complete dataset with no cleaning required and hence can be used directly. A few changes were made to the dataset post-uploading, concatenating year, month, day columns into a single column called Year_Month_Day. The data type was then changed to date and the column was renamed to “Date”.

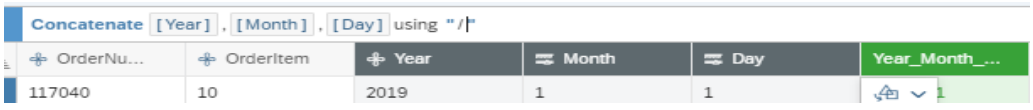


Figure 2.1: Concatenating Year, Month, Day

The Customer and Product columns lacked characteristics and hence to provide more information attributes related to them were added.

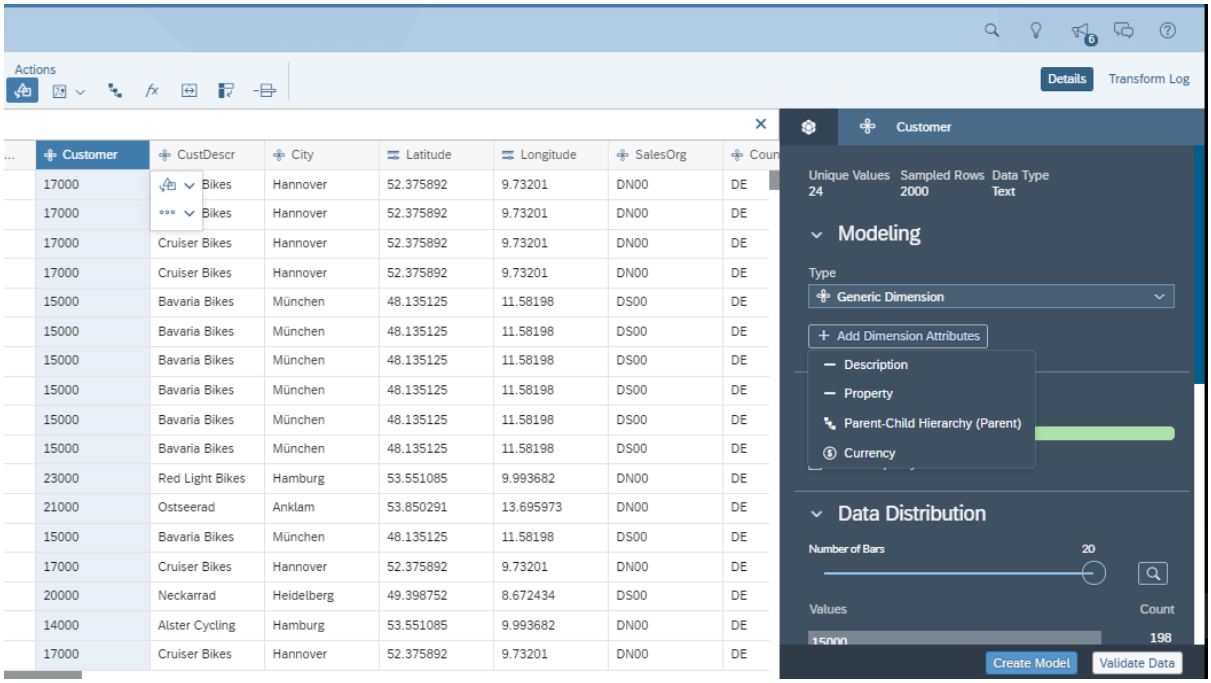


Figure 2.2: Adding Dimension Attributes

Another step to transform the data was creating hierarchies for both the Product column and the Customer Column. Hierarchies are the attributes that are related to each other in a parent-child relationship. This is done by using the action ribbon and choosing the level-based hierarchy option, which opens a dialog box where the selected fields can be added to different hierarchical levels.

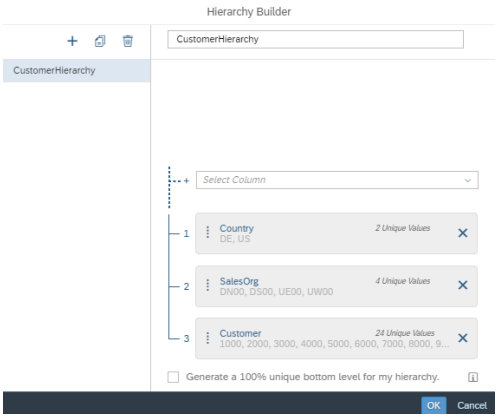


Figure 2.3: Customer Hierarchy

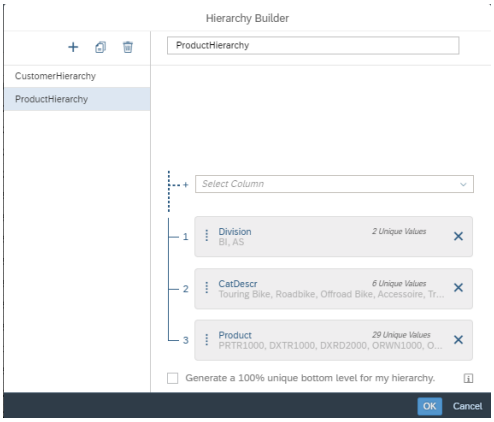


Figure 2.4: Product Hierarchy

The final transformation done was to create a geo-enrichment based on the latitude and longitude of the Customer data given in the dataset using the option - Geo by Coordinates. The purpose of adding the Geographic data is so that the data can be visualized region-wise to get a better understanding of sales in various locations or outlets across the globe.

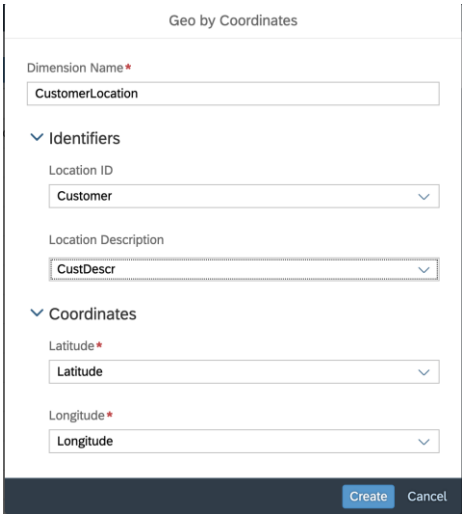


Figure 2.5: Adding a Geo-enrichment to the Customer Dimension

Once all these changes are done, the dataset is now ready to be changed into a data model which can be used for further analysis and derive insights through various visualizations.

2.3 Analysis using SAP Analytics Cloud and Outcomes

After switching the workspace view to the Model Structure (figure 6), it shows the star schema of the data where the fact data is in the center and is surrounded by the different dimension data which is used to further describe the fact data.

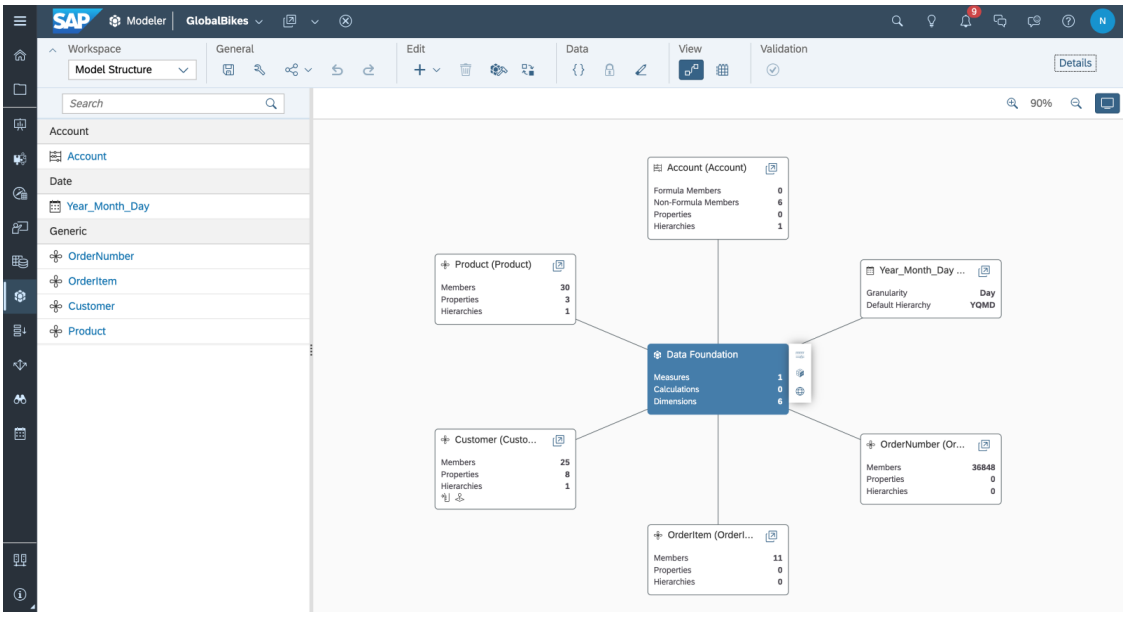
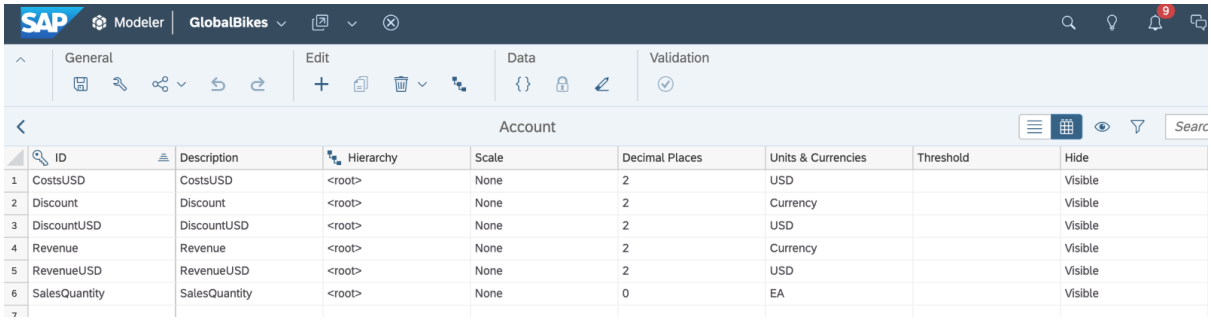


Figure 2.6: the Model Dimension

A fine-grained view is a detailed view in data modelling that refers to a subset of data within the dataset involving specific columns and relationships. The views of this dataset will be explained in detail below:

2.3.1 Dimension 1 – Account Fine Grain View



	ID	Description	Hierarchy	Scale	Decimal Places	Units & Currencies	Threshold	Hide
1	CostsUSD	CostsUSD	<root>	None	2	USD		Visible
2	Discount	Discount	<root>	None	2	Currency		Visible
3	DiscountUSD	DiscountUSD	<root>	None	2	USD		Visible
4	Revenue	Revenue	<root>	None	2	Currency		Visible
5	RevenueUSD	RevenueUSD	<root>	None	2	USD		Visible
6	SalesQuantity	SalesQuantity	<root>	None	0	EA		Visible

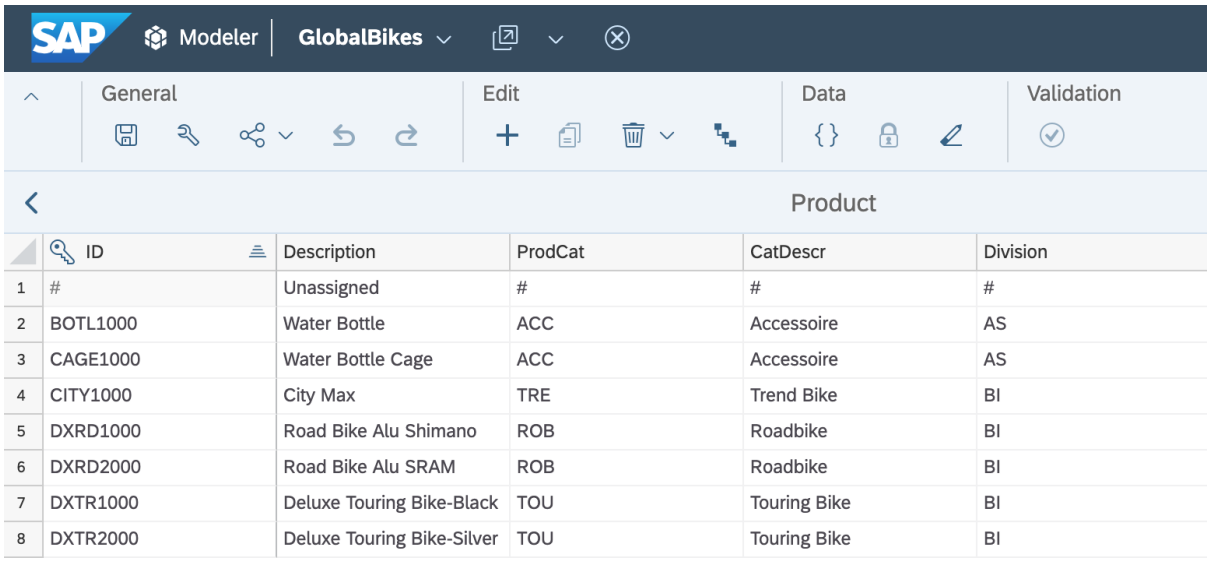
Figure 2.7: Account Dimension – Fine Grain View

The account view consists of a total of six rows and seven columns, each containing data (Figure 7). Within the ID column, all dataset measures necessary for calculations are found. These six measures include CostUSD, Discount, DiscountUSD, Revenue, RevenueUSD, and SalesQuantity.

At the root level, all measures exist without any parent or child attributes; they are at the fundamental level. The initial five measures are configured to display two decimal places, while SalesQuantity is set to zero since it exclusively represents whole numbers. CostUSD, DiscountUSD, and RevenueUSD are denominated in USD, while Discount and Revenue are designated as Currency, allowing for representation in either Euro or US dollars. SalesQuantity is specified as EA, reflecting that the products are sold individually.

This dimension allows for various visualizations, such as illustrating trends over time sing a line graph, a bar chart to represent all six measures, and a scatter plot to represent the relationship between cost and revenue.

2.3.2 Dimension 2 – Product Fine Grain View



	ID	Description	ProdCat	CatDescr	Division
1	#	Unassigned	#	#	#
2	BOTL1000	Water Bottle	ACC	Accessoire	AS
3	CAGE1000	Water Bottle Cage	ACC	Accessoire	AS
4	CITY1000	City Max	TRE	Trend Bike	BI
5	DXRD1000	Road Bike Alu Shimano	ROB	Roadbike	BI
6	DXRD2000	Road Bike Alu SRAM	ROB	Roadbike	BI
7	DXTR1000	Deluxe Touring Bike-Black	TOU	Touring Bike	BI
8	DXTR2000	Deluxe Touring Bike-Silver	TOU	Touring Bike	BI

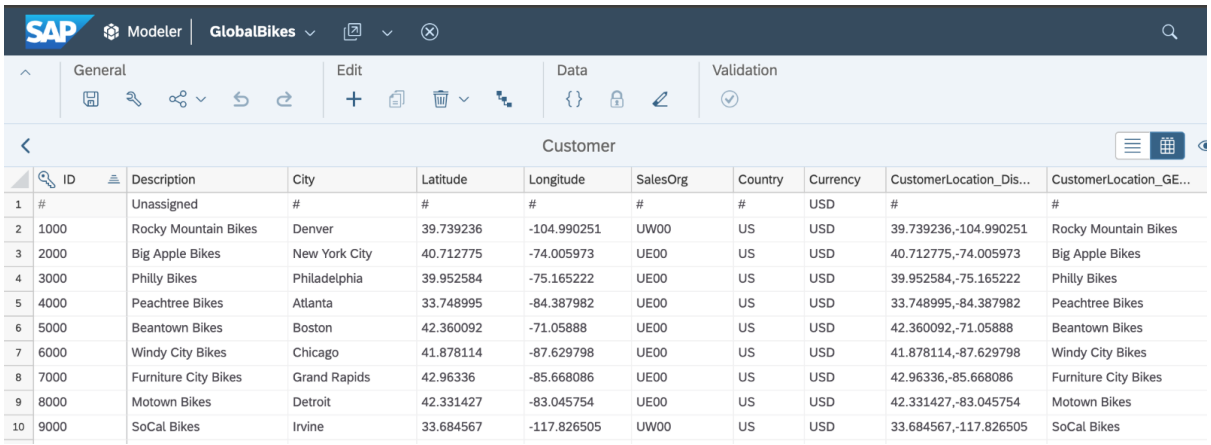
Figure 2.8: Product Dimension – Fine Grain View

The Product view comprises a total of 30 rows and five columns containing data (Figure 8). Within the ID column, you'll find all the product IDs for the various products offered by the company. The description column provides information on each product, while ProdCat categorizes these 30 products into six distinct categories: ACC, TRE, ROB, TOU, ORB, and EBI.

The CatDescr column offers brief explanations for the abbreviations used in the ProdCat column, where ACC corresponds to Accessoires, TRE to Trend Bike, ROB to Roadbike, TOU to Touring Bike, ORB represents off-road bike, and EBI denotes E-bike. The division column further classifies the entire product list into two categories: Accessories or bikes.

This dimension allows for various visualizations, such as illustrating product distribution, tracking category trends over time, and exploring the relationships between product categories and divisions.

2.3.3 Dimension 3 – Customer Fine Grain View



	ID	Description	City	Latitude	Longitude	SalesOrg	Country	Currency	CustomerLocation_Dis...	CustomerLocation_GE...
1	#	Unassigned	#	#	#	#	#	USD	#	#
2	1000	Rocky Mountain Bikes	Denver	39.739236	-104.990251	UW00	US	USD	39.739236,-104.990251	Rocky Mountain Bikes
3	2000	Big Apple Bikes	New York City	40.712775	-74.005973	UE00	US	USD	40.712775,-74.005973	Big Apple Bikes
4	3000	Philly Bikes	Philadelphia	39.952584	-75.165222	UE00	US	USD	39.952584,-75.165222	Philly Bikes
5	4000	Peachtree Bikes	Atlanta	33.748995	-84.387982	UE00	US	USD	33.748995,-84.387982	Peachtree Bikes
6	5000	Beantown Bikes	Boston	42.360092	-71.05888	UE00	US	USD	42.360092,-71.05888	Beantown Bikes
7	6000	Windy City Bikes	Chicago	41.878114	-87.629798	UE00	US	USD	41.878114,-87.629798	Windy City Bikes
8	7000	Furniture City Bikes	Grand Rapids	42.96336	-85.668086	UE00	US	USD	42.96336,-85.668086	Furniture City Bikes
9	8000	Motown Bikes	Detroit	42.331427	-83.045754	UE00	US	USD	42.331427,-83.045754	Motown Bikes
10	9000	SoCal Bikes	Irvine	33.684567	-117.826505	UW00	US	USD	33.684567,-117.826505	SoCal Bikes

Figure 2.9: Customer Dimension – Fine Grain View

The Customer view comprises a total of 25 rows and 10 columns, each populated with data (Figure 9). Within the table, the ID column contains unique customer IDs, while the Description column provides the corresponding customer names. The city column denotes the city in which the customer is based. The subsequent two columns offer the latitude and longitude coordinates, specifying the precise global location of the store, followed by details about the sales organization to which they are affiliated.

The next two columns, country, disclose the country of operation and the currency utilized for transactions. The CustomerLocation_Distribution column provides a calculated position, facilitating the plotting of the exact location on a map. Lastly, the last column replicates the information found in the Description column.

This dimension facilitates diverse visualizations, including the depiction of the precise geographical coordinates of individual stores, the creation of a customer location density map, and the generation of a heatmap illustrating the correlation between the currency used for transactions and the corresponding countries.

2.4 Potential Future Outcomes and Analyses:

The three described tables - Account View, Product View, and Customer View - offer a rich source of information for diverse analyses and insights.

Geospatial Analysis: Geographical coordinates in both Product and Customer Views enable geospatial analysis, helping to understand store and customer distribution globally.

Sales Performance: Utilizing data from the Account View, one can conduct thorough sales performance analyses, identify trends, and make informed business decisions.

Product Strategy: The Product View facilitates a deep dive into product categories, aiding in product strategy formulation and inventory management.

Customer Insights: Combining information from the Customer View with other tables could lead to valuable insights into customer demographics, preferences, and regional variations.

Currency and Country Relations: The Currency and Country columns in the Customer View can be leveraged to analyze transaction patterns, exchange rate impacts, and global business trends.

By integrating and analyzing data from these views, businesses can gain a holistic understanding of their financial performance, product dynamics, and customer interactions, paving the way for strategic decision-making and targeted improvements.

2.5 Conclusion

In brief, SAP Analytics Cloud (SAC) is a robust, integrated solution that combines business intelligence, planning, and predictive analytics on a unified platform. It boasts seamless collaboration, real-time analytics, scalability, and an intuitive interface. However, I feel that users may face a learning curve, connectivity dependencies, and considerations regarding costs and integration. Despite these challenges, SAC empowers organizations to make informed, data-driven decisions, enhancing agility and efficiency in a rapidly changing business landscape. As businesses prioritize analytics for strategic insights, SAC proves to be a valuable, comprehensive, cloud-based solution for analytical and planning needs.

Chapter 3 Tableau – Data Manipulation for Analysis

3.1 Introduction to Tableau

Tableau is a useful tool for looking at data and making sense of it. This tool is like a superhero for numbers. It's not just regular software; it's like magic for people who want to use their data in smart ways. We'll take a close look at what Tableau can do, talking about all its cool features and how it turns boring data into pictures and stories that are easy to understand.

As we dig into Tableau, we'll find out how it can play with data in smart ways. It's not just about making charts and graphs; Tableau is like a wizard that helps you make smart decisions with your data. We'll explore its tricks, like those interactive charts and dashboards that make data less boring.

But here's the exciting part – we're not just going to talk about the technical stuff. We'll also look at where Tableau fits into the real world. Whether it's big companies figuring out money stuff or schools dealing with tricky numbers, Tableau seems to be super helpful everywhere.

So, get ready! We're on a mission to discover why Tableau is so awesome, not just in the techy bits but also in the real world where people need to make sense of their data. This introduction is like the starting point for our journey, where we'll find out why Tableau is like a superhero for anyone dealing with data in all sorts of places.

3.2 Explanation of the Dataset Used

Using the examples and lecture of Chapter 5 as a reference, I have used Tableau to analyze the wholesale data given as a .xlsx file. The data is from the years 2007 to 2016 for a fictitious company called GBI. The file includes 23 columns of variables including customer, product, date, revenue, sales quantity etc. and 132,760 rows. Since the dataset was complete without any missing data or duplicates, no cleaning was required, and the file was ready to be used.

Using the Following steps, Import given file (GBI_E5_2) into Tableau:

1. Run Tableau (Public) → Connect Microsoft Excel → Open.
2. Choose GBI_E5_2.xlsx → Click ‘New worksheet’ (Bottom left).

3. Check the formats of measures and dimensions to see if they are in the right format.
4. If not in the appropriate format, change to the desired format by right clicking the '#' button.

3.3 Research Questions to be Solved.

Using Tableau, the below questions based on the transformations will be addressed:

1. Which year had the highest revenues (in USD) overall and how much were the revenues during that year?
2. What was the year with the highest overall gross margin (in USD) and what was the amount?

   a. During this year, what was the gross margin in dollars for Germany and what was it

   for the U.S.?

   b. During this year, what was the gross margin as a percentage of Net sales (the gross margin ratio) for Germany and the U.S.?

3. Which sales organization has the highest revenues for 2016?
4. In the year 2016, which was the product with the highest sales in terms of quantities sold

   and what was that quantity?

5. Which product (represented by 'product description) was the lowest-selling accessory

   (Division AS) in terms of sales quantities sold during 2016?

3.4 Analysis of the Research Questions and Outcomes

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

To get the highest revenue overall and the revenue during that year, drag the Measure which is Revenue in USD into the Rows and the Dimension Year into the column.

Change the chart type to a bar under the Marks section.

Click 'Label' and check 'Show Mark Labels' to view the revenue for each bar.

Sort year in descending by 'Revenue in USD'.

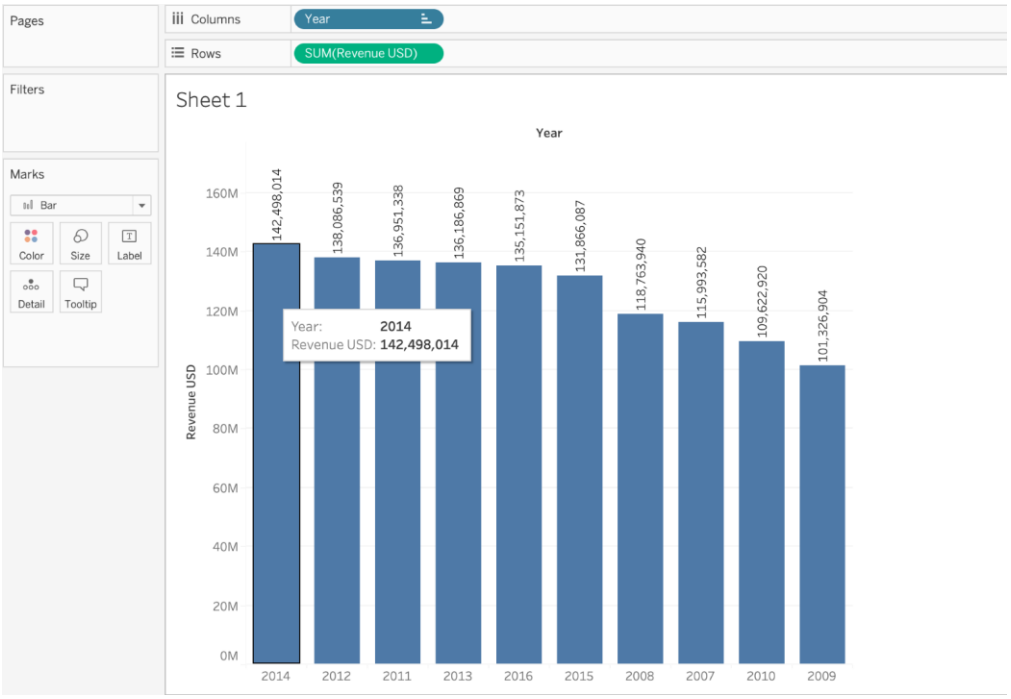


Figure 3.1: Overall Revenue in USD by Year

A: The year with the highest Revenue in USD is 2014 and the Revenue made is $ 142,498,014.

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

a. During this year, what was the gross margin in dollars for Germany and what was it for the U.S.?

b. During this year, what was the gross margin as a percentage of Net sales (the gross ratio) for Germany and the U.S.?

Navigate to the Analysis Menu, then select Create Calculated Field. Input the provided formula by dragging and dropping the specified three measures, while incorporating the required arithmetic operators (minus signs).

Subsequently, apply the same approach as in the prior question to obtain the solution for this particular query. However, this time, drag 'Gross Margin in USD' from Measures and drop it into Rows.

To eliminate sorting by gross margin, right-click on 'Year' within Columns and select Clear Sort. This action will display the Gross Margin in USD periodically over the years.

If you wish to visualize both revenue and Gross Margin in USD on the same chart, drag and drop Revenue in USD into Rows. To create dual axes, right-click on the vertical axis of 'Gross Margin in USD,' then click on 'Dual Axis.'

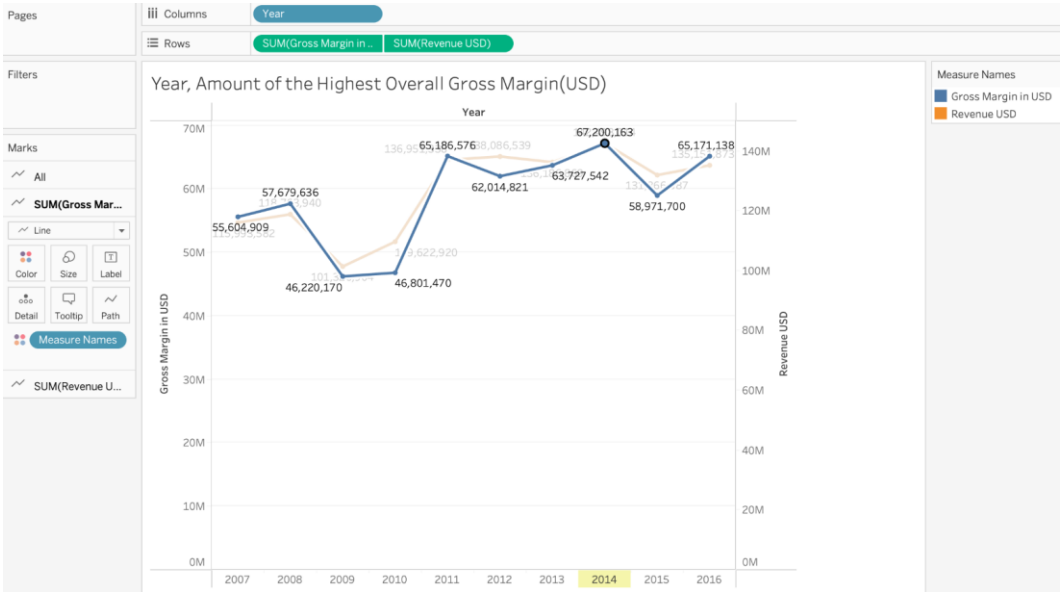


Figure 3.2: Year, Amount of the Highest Overall Gross Margin

A: The year with the highest overall gross margin (in USD) was 2014 and the amount was $65,171,138.

a. During this year, what was the gross margin in dollars for Germany and what was it for the U.S.?

Spot the year with the top Gross Margin in USD, click on it. Then, hit 'Keep only.' After that, drag 'Country' into 'Color.' Next up, go to the 'Show Me' section on the right side and pick the 'Side-by-side' bar chart. Don't forget to tick the 'Labels' option to show them on your chart.

Head to the Analysis Menu, then hit Create Calculated Field. Now, you can whip up the 'Gross Margin Ratio' using the formula below. After that, tweak the aggregation method to 'average' by right-clicking on 'Gross Margin Ratio,' going to Default Properties, selecting Aggregation, and picking 'Average.'

Next, follow the same steps as the hint for Questions 2 and 2.b. to crack this question. Just remember to drag 'Gross Margin Ratio' from Measures and drop it into Rows.

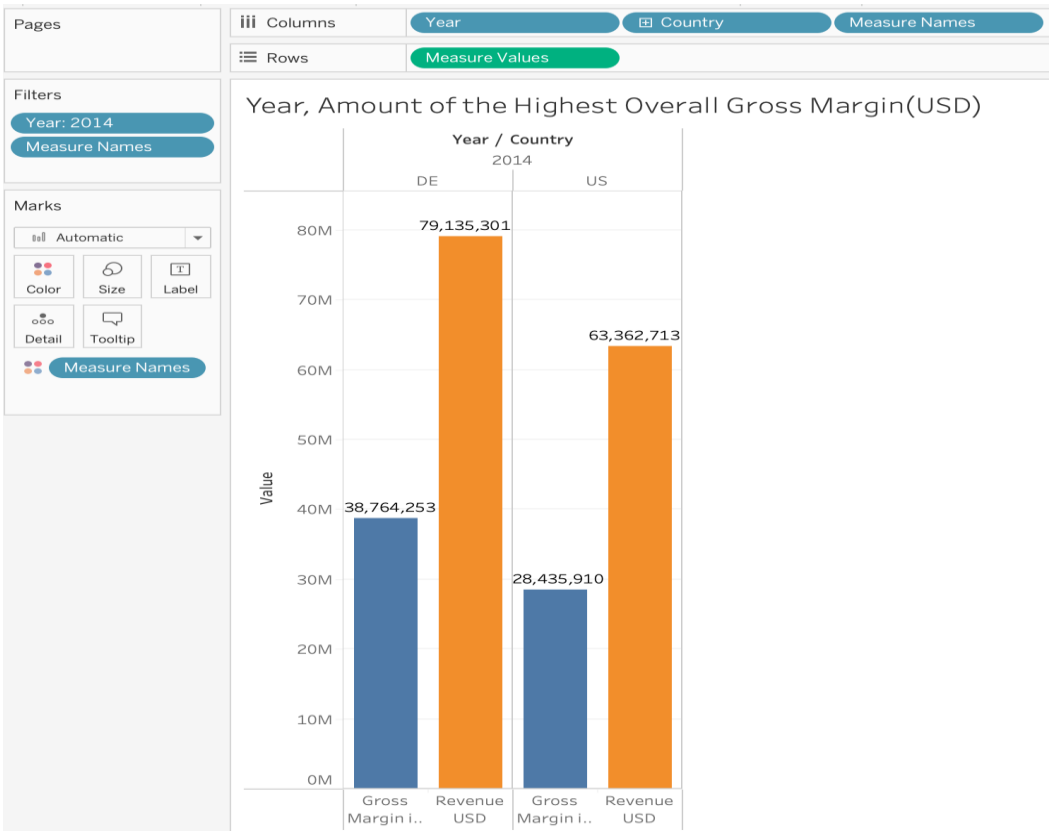


Figure 3.3: Overall Gross Margin for Germany and US in 2014

A: During 2014, the gross margin in dollars for Germany was $38,764253 and for the US was $28,435,910.

b. During this year, what was the gross margin as a percentage of Net sales (the gross margin ratio) for Germany and the U.S.?

Once you've got the answer, switch the format of 'Gross Margin Ratio' to percentage. To do this, right-click on 'Gross Margin Ratio' under Measures, go to Default Properties, select Number format, and opt for Percentage.

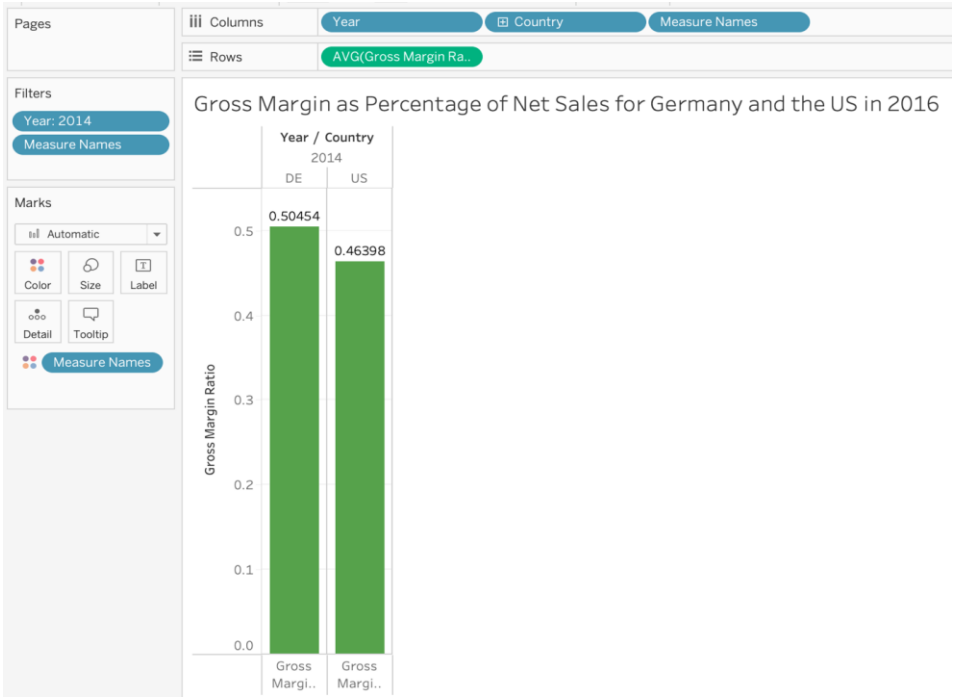


Figure 3.4: Percentage of Gross Margin for Germany and US in 2014

A: During 2014, the gross margin as a percentage of Net sales for Germany was 0.50454 and for the U.S. was 0.46398.

3.4.3 Which sales organization has the highest revenues for 2016?

Choose Year, Sales Org as the columns and Revenue in USD and add a filter on the year dimension.

Under filters à Year à choose the drop-down arrow à edit filter à choose 2016 à ok. The graph will contain only 2016 data.

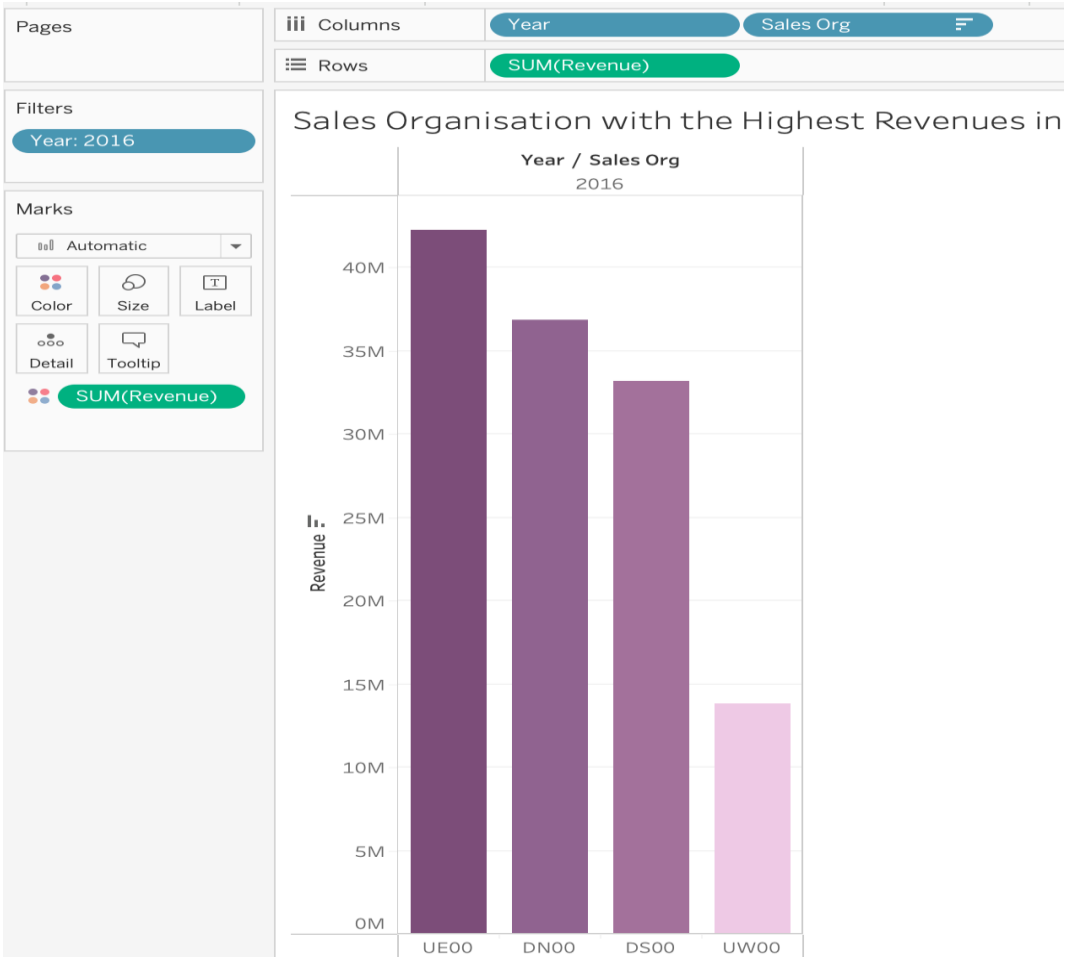


Figure 3.5: Sales Organization with the Highest Revenue in 2016

A: the sales organization with the highest revenue is UE00 with a revenue of $42,214,216.

3.4.4 In the year 2016, which was the product with the highest sales in terms of quantities sold and what was that quantity?

Choose Year, Product as columns and Sales Quantity as rows. Since the question is interested in only 2016.

Under filters à Year à choose the drop-down arrow à edit filter à choose 2016 à ok. The graph will contain only 2016 data.

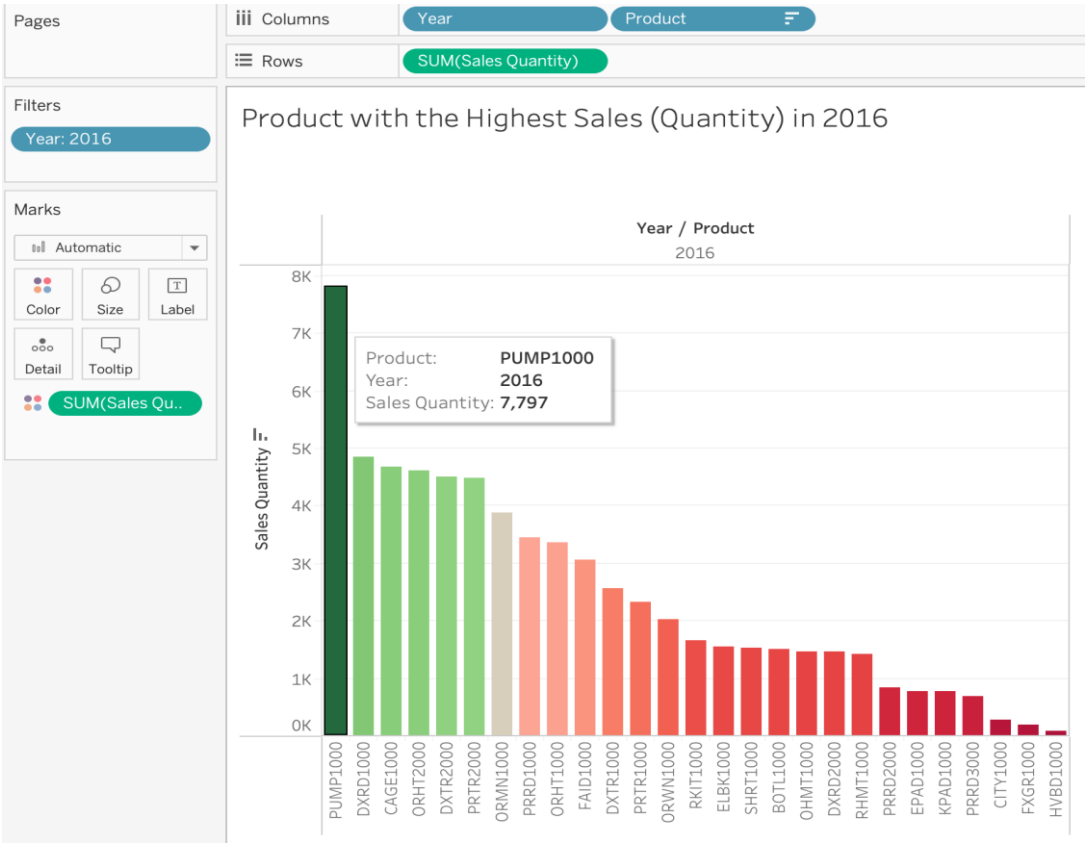


Figure 3.6: Product with the Highest Sales Quantity in 2016

A: The product with the highest sales in terms of quantities sold is PUMP1000. The quantity sold in the year 2016 was 7,797.

3.4.5 Which product (represented by 'product description) was the lowest-selling accessory (Division AS) in terms of sales quantities sold during 2016?

To get the below graph, add Prod Descr in columns and Sales Quantity in rows and add Year and Division as filters:

1. Year = 2016, under filters à Year à choose the drop-down arrow à edit filter à choose 2016 à ok. The graph will contain only 2016 data.
2. Division = AS, Under filters à Division à choose the drop-down arrow à edit filter à choose AS à ok. The graph will contain only AS(accessories) data.

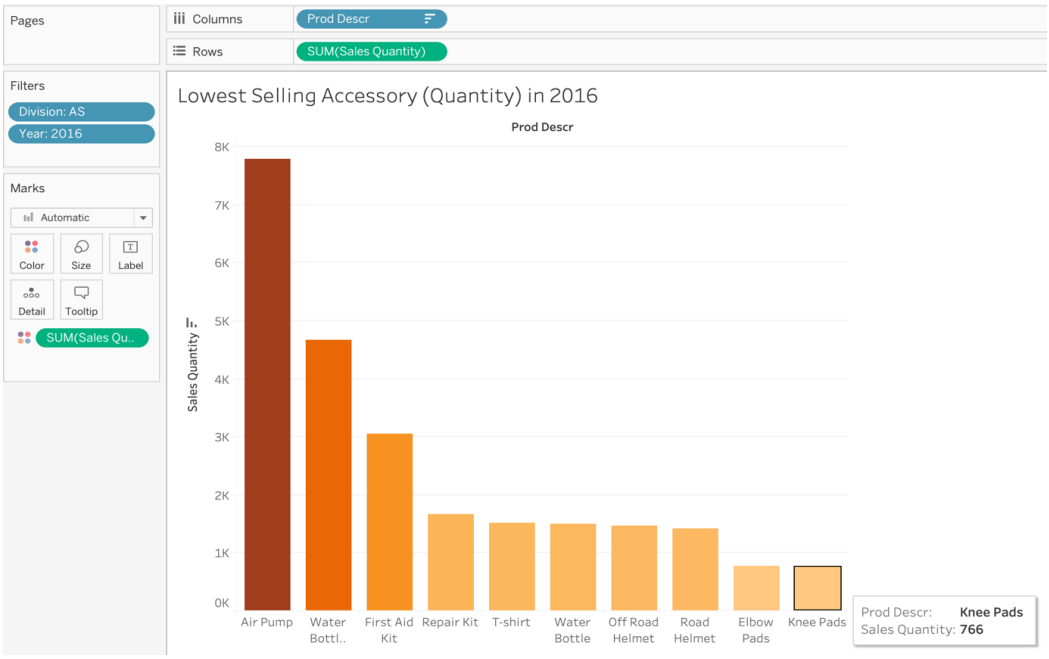


Figure 3.7: Lowest Selling Accessory During 2016

A: The lowest-selling accessory in 2016 was Knee Pads and the quantity sold was 766.

3.5 Conclusion

In conclusion, my exploration of Tableau has provided valuable insights into its role as a robust analytical tool. My personal opinion underscores its transformative impact, going beyond a simple data visualization tool to become a valuable ally in understanding and extracting meaningful information from complex datasets.

The analysis has yielded logical and agreeable conclusions, emphasizing Tableau's effectiveness in converting raw data into meaningful narratives. Notably, the tool's user-friendly features, particularly dynamic dashboards, stand out as powerful assets in simplifying data analysis and enhancing decision-making processes.

My critical thinking throughout the analysis has considered the appropriate use of Tableau in diverse contexts. Recognizing its versatility in corporate and educational settings showcases its adaptive nature. Simultaneously, acknowledging limitations, such as a learning curve for advanced functionalities, reflects a balanced perspective.

Tableau's benefits are clear – it empowers users to unlock the potential of their data, promoting a culture of data-driven decision-making. In reflection, this analysis serves as a testament to

the synergy between Tableau and a well-structured dataset. The interplay of technology and data has not only provided logical and agreeable insights but has also paved the way for a nuanced understanding of the wholesale operations of the fictional company GBI. As a student navigating the realm of data analysis, this experience has underscored the pivotal role of technology in transforming raw data into actionable insights.

In summary, my journey with Tableau as a student navigating the world of data analysis has been intellectually rewarding. It has provided a firsthand understanding of how technology, in the form of Tableau, can serve as a catalyst for informed decision-making.

Chapter 4 - ERPSim SAC

4.1 Introduction to ERPSim

ERP stands for Enterprise Resource Planning, a term that is widely associated with Database Management Systems. ERPSim is a business simulation game developed by SAP to provide hands-on experience with SAP's ERP software. In this software, participants take on various roles within a fictional company and work through various business scenarios to understand the overall business operations, develop business strategies and estimate outcomes.

The goal of SAP ERPSim is to provide participants with a practical understanding of how ERP systems work in real-world business settings, including how data flows between different departments, how transactions are processed, and how decisions impact overall business performance.

The ERP game is played by several teams over several rounds and sell their products in the market. The decisions made by the teams determine the sales, while the simulator automates the operations. When connected to the SAP HANA ERPSim system, the teams can run analytical reports from time to time in the game to monitor activities and strategize the next steps.

Therefore, this software can be used for the following purposes:

1. Training and Education especially in academic programs and professional training for hands-on experience with SAP ERP.

2. Skill Development to improve their proficiency in navigating the system, entering data, and making decisions within a simulated business environment.

3. Business Process Optimization through simulating and testing different processes for optimization.

4. Change Management analysis to simulate the impact of these changes on business processes, helping to manage the transition and mitigate potential risks.

5. Team Building where the participants can work together to foster collaboration and problem-solving skills through teamwork.

6. Decision-making training to help understand the consequences of various choices and how those decisions affect overall business performance.

4.2 Dataset Used and Problem Description

The file used for this exercise is ERPSIM.xlsx. The dataset consists of 6558 rows and 15 columns out of which 3 are measures and 12 are dimensions. Measures are the key figures or facts which are used to make calculations in this case are Price, Quantity and Revenue. The dimension columns give more information regarding a measure. The dimensions in this file are Team, Area, Product, Round, Country, Region, City, Day, Distribution Channel, and Sales_Order. It is used to examine past sales performance and come up with strategies for the future. It is a complete dataset with no cleaning required and hence can be used directly. A few changes were made to the dataset post-uploading to help facilitate better reporting.

To start your story, choose the Stories option in the main menu à Canvas à Classic Design Experience à Create à Add Data.

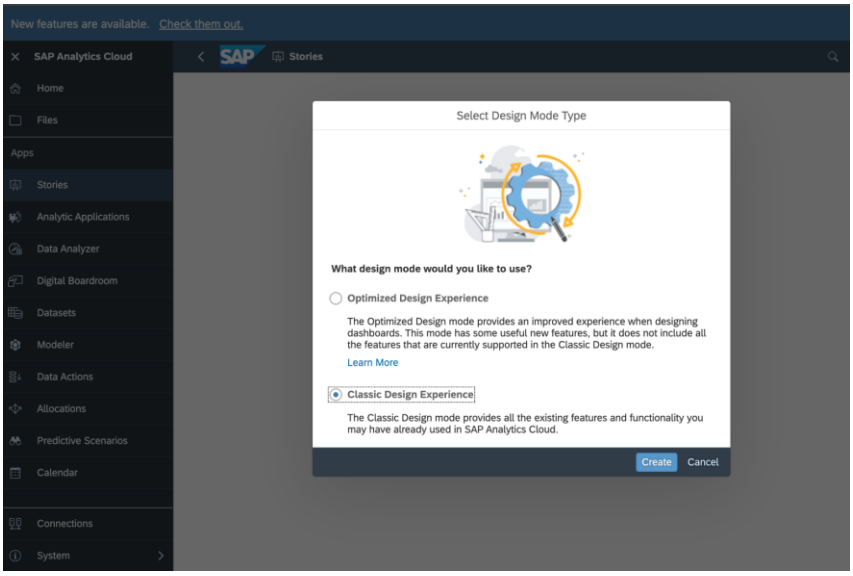


Figure 4.1: Starting the Story.

The data is then uploaded using Select Source File à choose the ERPSIM.xlsx from your local machine à Import.

Figure 4.2: Importing the Dataset

Once the dataset is successfully uploaded, you will see a Data View of your data. Since some of the fields are identified incorrectly, move Round, Day, Distribution Channel, SalesOrder, lat, Lng into Dimensions. Now you will have 3 measures and 12 dimensions. Once done save the work by creating your folder.

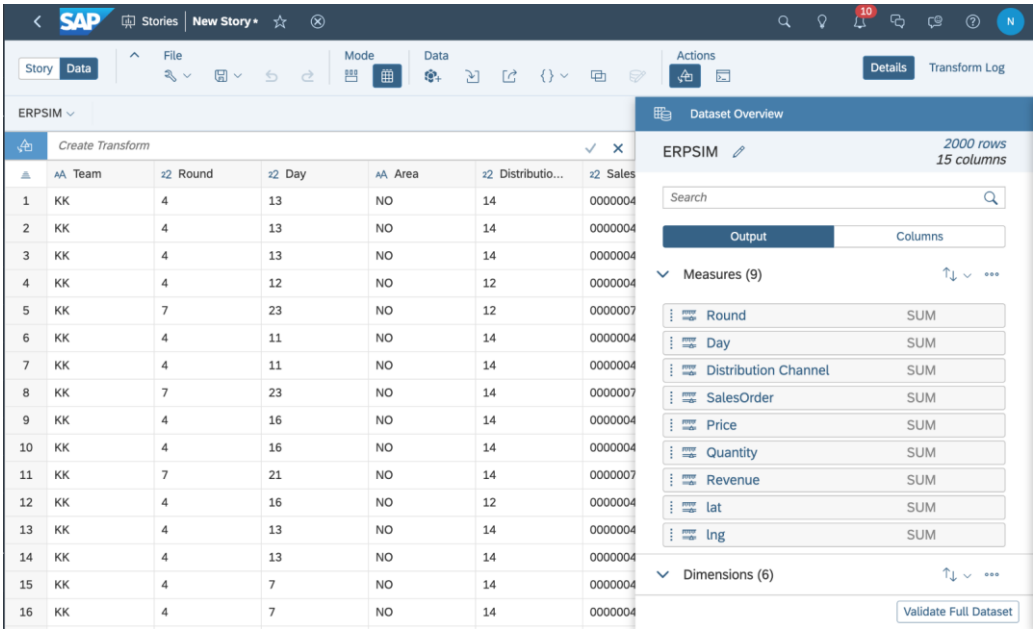


Figure 4.3: Data View with Measures and Dimensions Identified Automatically by SAC.

The next step is to enhance the data by doing the following steps:

1. The values of the dimensions need to be changed. Highlight the Area column. Click on the downward-pointed arrow to show additional icons.
2. Select Create a Transformation à Replace.
3. Replace NO with North, SO with South, WE with West, EA with East.
4. Change the Data Types from the Distribution Channels from Integer to String from the drop-down list and change the statistical type from Continuous to Nominal.
5. Transform the Distribution Channels where 10 as Hypermarkets, 12 as Grocery Chains, and 14 to Convenience Stores using the same Transformation process.

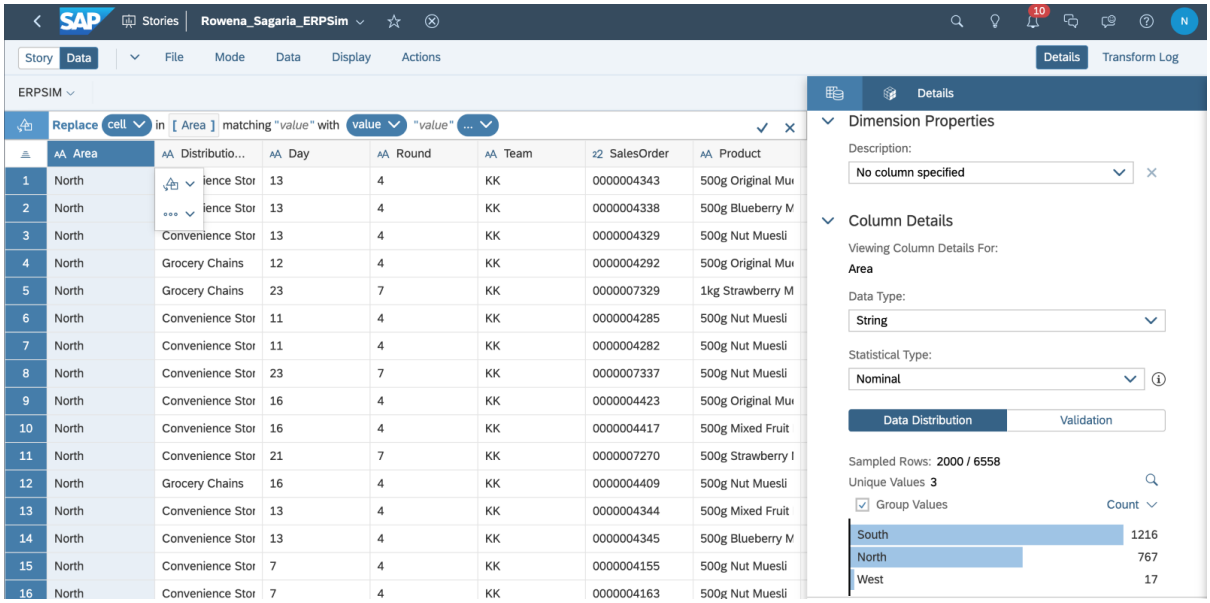


Figure 4.4: Replacing Text Using the Transformation Option

We can examine and get inferences of specific data groups like the City panel; we can get the name of the top 3 regions in terms of the number of sales orders (count).

To get the top 3 regions, combine Round and Day by creating a hierarchy.

1. Select both the Round and Day columns.
2. Select the Create Hierarchy option from the Dimension Actions button in the Dataset Overview.
3. Under the Edit option, rename Hierarchy as Round and Day and press OK.

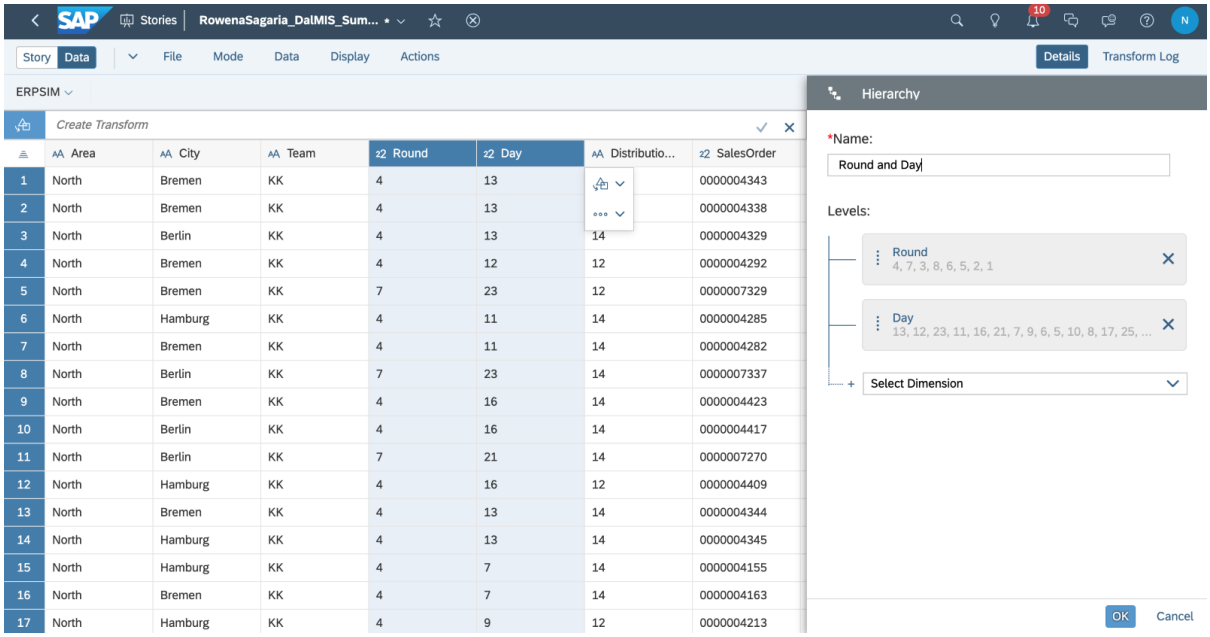


Figure 4.5: Creating a Hierarchy for the City Dimension

A: The top 3 regions in terms of the number of sales orders are Stuttgart, Nuremberg, and Munich accordingly.

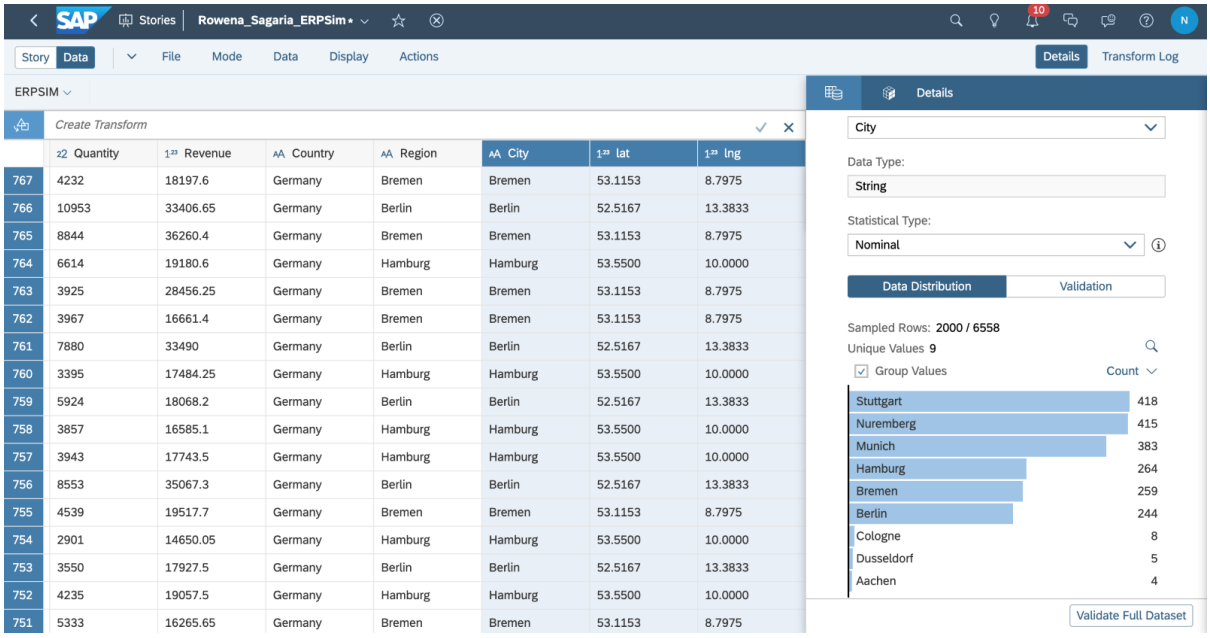


Figure 4.6: Number of Sales Orders Based on Region.

4.3 Research Questions to be Solved

Using ERPSim, the below questions based on the transformations will be addressed:

1. Which team had the third-highest revenue? What was the revenue?

2. Display the trend of revenue and quantity over rounds for each team (only for KK, LL, MM, NN, OO) by the products that are packaged as 500g.
3. Show the market share (in terms of revenue) of regions per each team (with percentage).

4. Illustrate the difference in Revenue and Quantity by different cities on the Geo map chart using two layers (Bubble and Choropleth/Drill layer).

5. What product(s) brought the 2nd highest quantity in the distribution channel 'Convenience Store' by team PP and TT (not combined but per each team)? I.E., the product with the 2nd highest quantity by team PP in the distribution channel 'Convenience Store' and the product with the 2nd highest quantity by team TT in the distribution channel 'Convenience Store'? If possible, try to use one chart to solve this problem.

4.4 Analysis of the Research Questions using ERPSim and Outcomes

Using the above data, switch to the story tab à add an object such as a chart, Geo Map or table to start the data visualization.

4.4.1 Which team had the third-highest revenue? What was the revenue?

Using a bar chart, choose Measures as Revenue and Dimensions as Team. Change the graph orientation to vertical and then hit save.

A: The team with the third-highest revenue is KK. The revenue of Team KK is 24,372,421 euros.

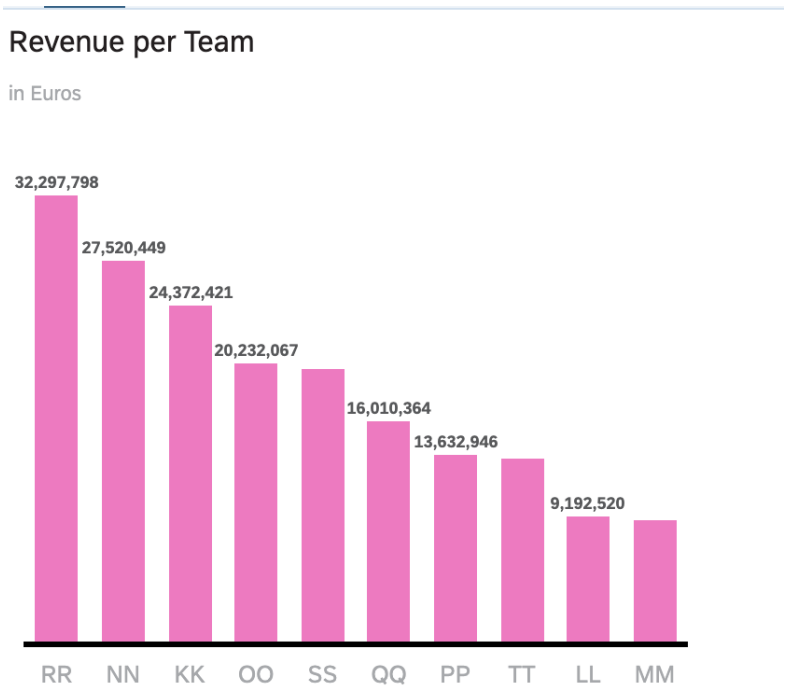


Figure 4.7: Bar Graph – Revenue per Team

4.4.2 Display the trend of revenue and quantity over rounds for each team (only for KK, LL, MM, NN, OO) by the products that are packaged as 500g.

A: Using a line chart, choose Revenue(measure) for the Left Y-axis and Quantity(measure) for The Right Y-Axis. Choose Dimension – Round, Color – Team. Use 2 filters, one for Products à select only one 500gms and Team à tick KK, LL, MM, NN.

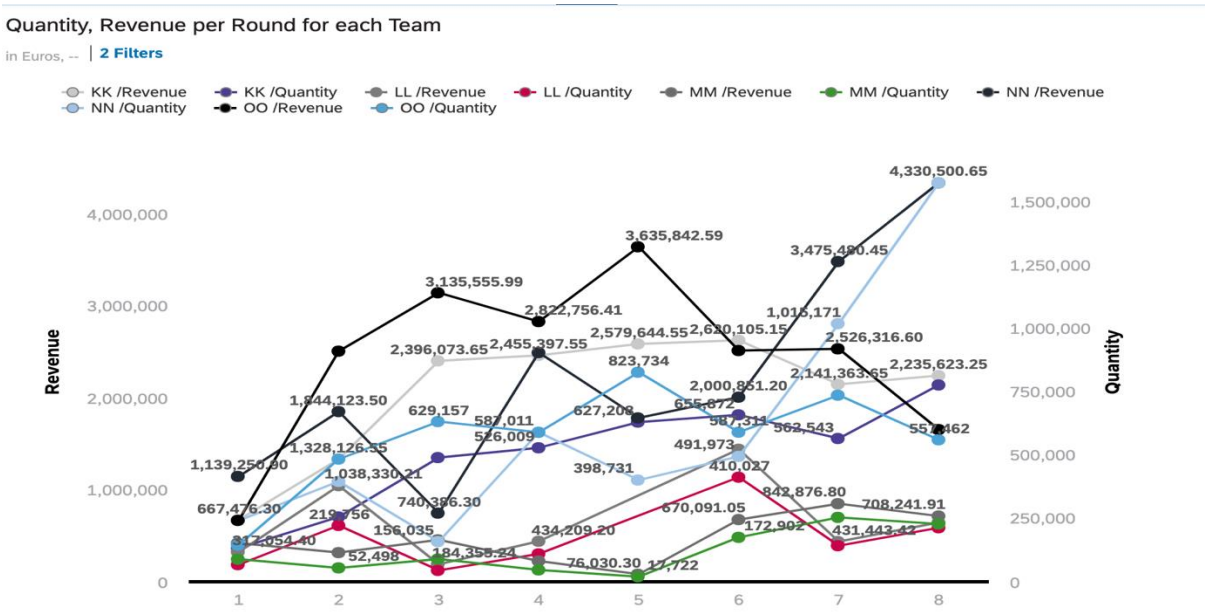


Figure 4.8: Line Graph – Quantity, Revenue per Round for Chosen Teams

4.4.3 Show the market share (in terms of revenue) of regions per team (with percentage).

A: Using a Stacked Bar/Column Chart, choose Measure – Revenue, Dimension – Team, Color – Region. Under the Chart Orientation, check the option “Show Chart as 100%”.

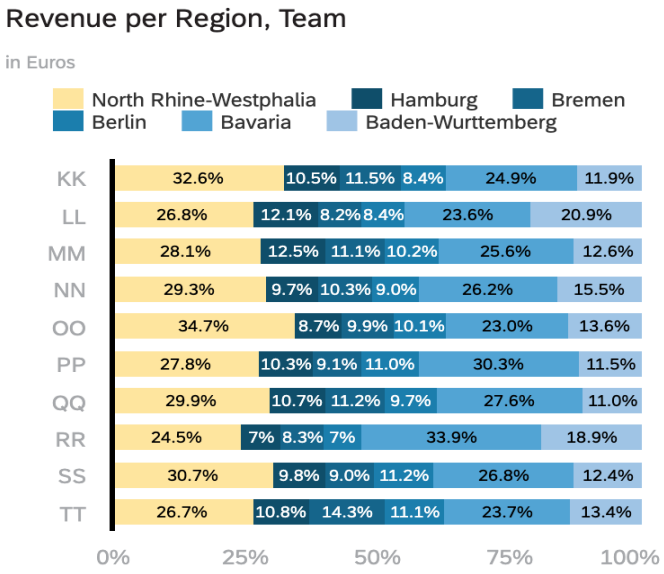


Figure 4.9: Stacked Bar Chart – Revenue per Region, Team

4.4.4 Illustrate the difference in Revenue and Quantity by different cities on the Geo map chart using two layers (Bubble and Choropleth/Drill layer).

A: Initiate Canvas and generate a 'Geo Map.' Click on +Add layer, keeping the bubble layer as default. Further, add a location dimension, choose City, and set Quantity as 'Bubble Size.' Decrease the 'Bubble Size' to 50% and confirm with OK. Enlarge the chart using the 'Zoom to Data' button located on the top-left side of the map.

Add another layer by clicking +Add layer. Opt for the 'Heat map layer' as the layer type. Integrate a location dimension, selecting City, and set Revenue as the parameter for 'Heat map Color.' Confirm with OK.

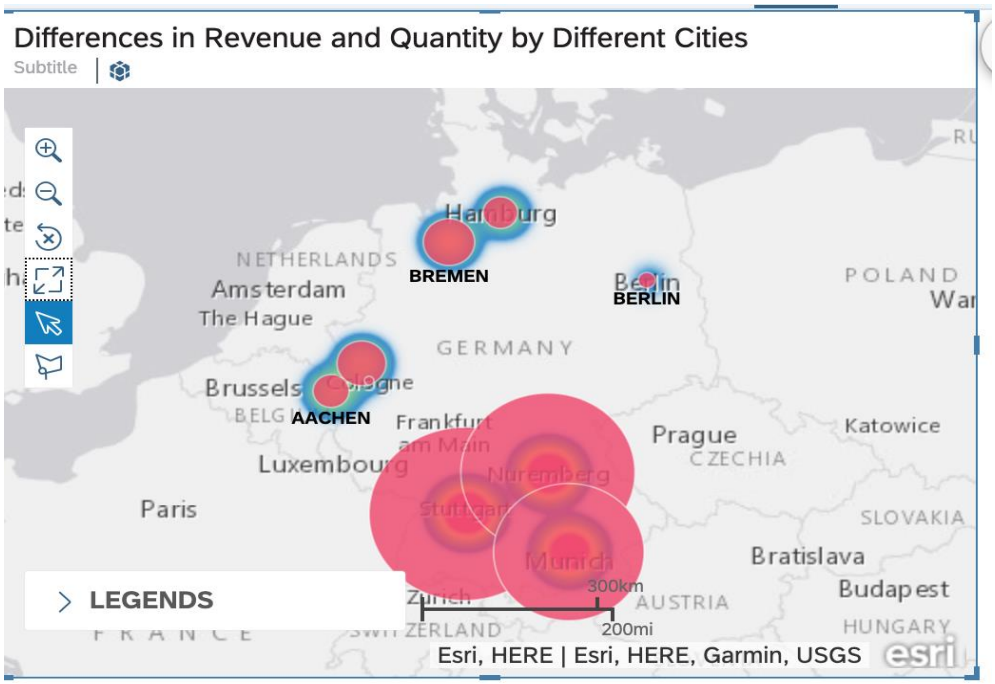


Figure 4.10: Bubble Chart – Difference in Revenue and Quantity by Cities

4.4.5 What product(s) brought the 2nd highest quantity in the distribution channel ‘Convenience Store’ by team PP and TT (not combined but per each team)? I.E., the product with the 2nd highest quantity by team PP in the distribution channel ‘Convenience Store’ and the product with the 2nd highest quantity by team TT in the distribution channel ‘Convenience Store’? If possible, try to use one chart to solve this problem.

Create a 'Responsive' Page in SAC:

a. Initiate a New Page by clicking on the + symbol adjacent to the page number and choosing 'responsive.' The default setup includes two lanes on the 'Responsive' page, and you can eliminate one by right-clicking on the desired lane and selecting 'Remove.'

b. Incorporate a chart by clicking on the 'chart' icon.

c. Now, craft a responsive vertical bar chart:

- Include a title, such as 'A small dashboard for ERPsim game performance.'
- Select the vertical bar chart, then opt for 'Measures' and choose '+ create measure input control' for Measures. Select 'All measures.'
- For Dimensions, choose '+ create dimension input control' for Dimensions. Select Area, City, Distribution Channel, Product, Round, Day, and Team.
- The resulting chart will be visible below. Now, effortlessly visualize the desired dimension and measure by clicking the buttons on the responsive chart's side.

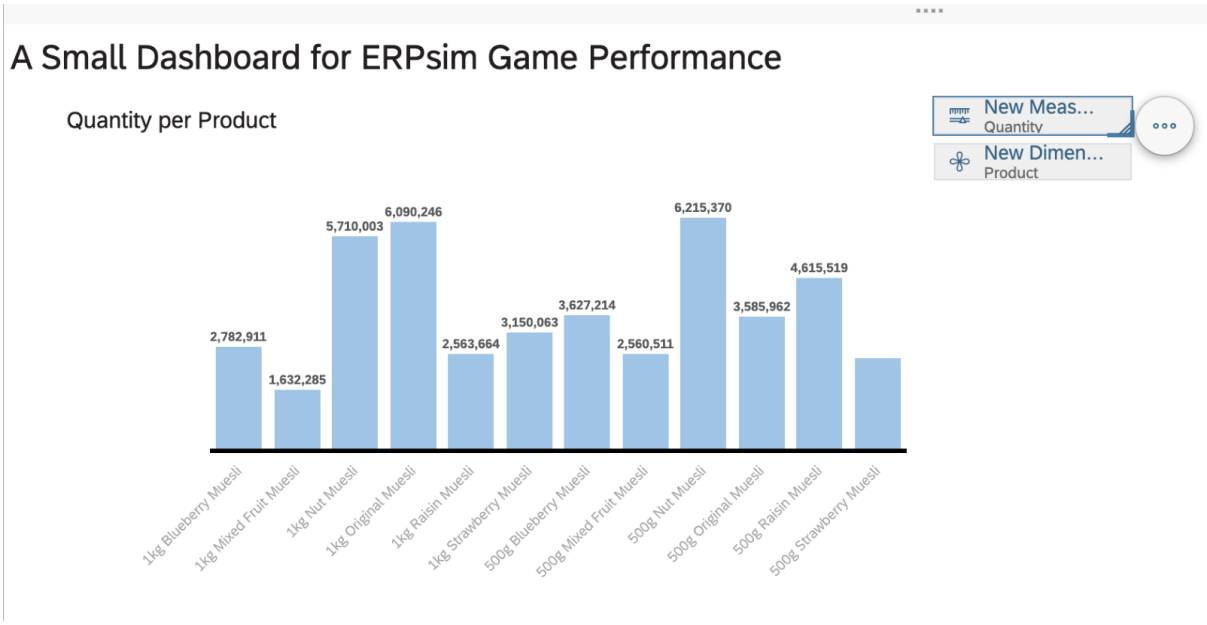


Figure 4.11: Responsive Graph – Quantity per Product

A: The product with the second highest quantity by team PP in the distribution channel – Convenience Store is 500g Strawberry Muesli with a quantity of 175,455.

The product with the second highest quantity by team TT in the distribution channel - Convenience Store is 500g Original Muesli with a quantity of 366,216.

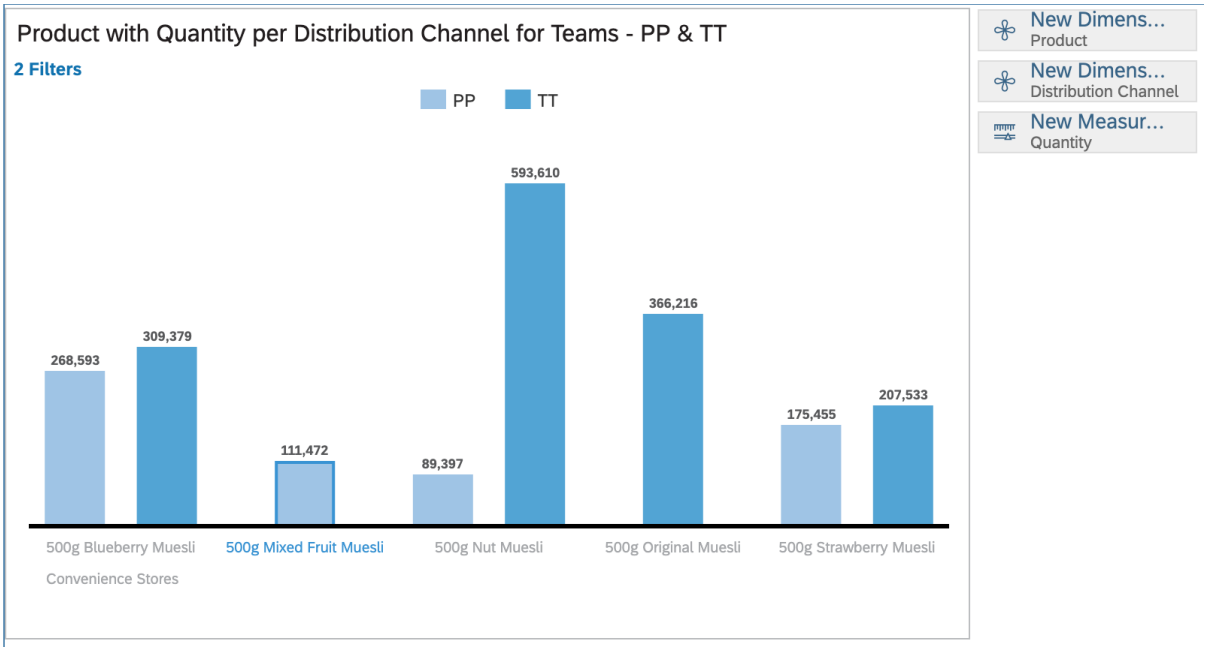


Figure 4.12: Product with Quantity per Distribution Channel for Teams – PP and TT.

4.5 Conclusion

In conclusion, ERPsim is a game-changer for businesses aiming to boost operational efficiency and strategic decision-making. Through its simulation features, organizations can experiment with and improve their enterprise resource planning (ERP) strategies in a low-risk setting. This not only helps spot potential weaknesses but also fosters the development of more robust and flexible business processes. Personally, I've found ERPsim's hands-on experience enhances user understanding and proficiency with ERP systems, leading to smoother implementations and better overall system utilization.

The game simulation feature in ERPsim adds a dynamic and engaging element to the learning and testing process, offering a realistic environment for users to apply and experiment with ERP strategies. I personally find this feature highly effective in providing practical insights and reinforcing theoretical knowledge. The interactive nature of the simulation allows users, like me, to navigate complex scenarios, make informed decisions, and witness the direct impact on business operations.

Concerning its suitability for the task at hand, ERPsim proves to be a versatile and efficient tool. Its ability to replicate real-world ERP challenges gives users, including university students like me, valuable hands-on experience. The addition of a responsive charting interface further boosts the tool's adaptability to diverse business needs, enabling users to customize simulations for specific scenarios. This flexibility, along with its user-friendly design, positions ERPsim as an invaluable asset for businesses and students alike, providing effective training and optimizing ERP systems for maximum efficiency and competitiveness.

Chapter 5: Data Wrangling and Analysis using AQUASTAT and FAOSTAT

5.1 Introduction to Data Wrangling and Analysis using AQUASTAT and FAOSTAT

In the realm of data analysis and research, the process of data wrangling and cleaning serves as the foundation upon which valuable insights are built. This report delves into the essential techniques of data wrangling and cleaning, focusing on the utilization of two prominent websites, AQUASTAT and FAOSTAT, as primary data sources. Through these platforms, we explore the intricacies of obtaining raw data pertaining to water resources, agricultural production, and related statistics, setting the stage for comprehensive analysis and interpretation.

Data Wrangling with AQUASTAT and FAOSTAT

AQUASTAT, maintained by the Food and Agriculture Organization of the United Nations (FAO), provides extensive datasets related to water and agriculture, offering a wealth of information crucial for understanding global water usage, irrigation practices, and associated challenges. Similarly, FAOSTAT serves as a comprehensive database for agricultural statistics, encompassing production, trade, and consumption data for various crops and livestock across countries and regions. Leveraging the capabilities of these platforms, we embark on a journey to collect and wrangle pertinent data, laying the groundwork for subsequent analysis and visualization.

Introduction to Data Cleaning Using Excel Pivot Tables

Effective data analysis hinges upon the cleanliness and accuracy of the dataset. However, raw data often arrives in formats fraught with inconsistencies, errors, and missing values, necessitating meticulous cleaning procedures to ensure reliability. In this report, we delve into the fundamentals of data cleaning, employing Excel pivot tables as a powerful tool to streamline the process. By harnessing the functionalities of pivot tables, we aim to rectify discrepancies, eliminate outliers, and prepare the data for meaningful analysis, thereby uncovering actionable insights and facilitating informed decision-making.

Throughout this report, we will explore the methodologies, challenges, and best practices associated with data wrangling and cleaning, equipping aspiring analysts and researchers with

the necessary skills to navigate the complexities of real-world data and extract valuable insights with confidence.

5.2 Explanation of the Datasets Used

There are three different datasets used for the exercise below.

1) The first dataset was taken from AQUASTAT – FAO's Global Information System on Water and Agriculture. https://www.fao.org/aquastat/en/databases/maindatabase/. It offers free access to 180+ water-related variables and indicators by country since 1960. For this assignment, data from 2011-2020 for Canada, China, France, Russia, and the USA were collected, resulting in a final dataset with 40 columns. Despite the valuable data on the website, challenges arose during downloading, particularly with errors occurring when selecting items and simultaneous table updates. Additionally, the dataset required cleansing to address issues like duplications, errors, and omissions.

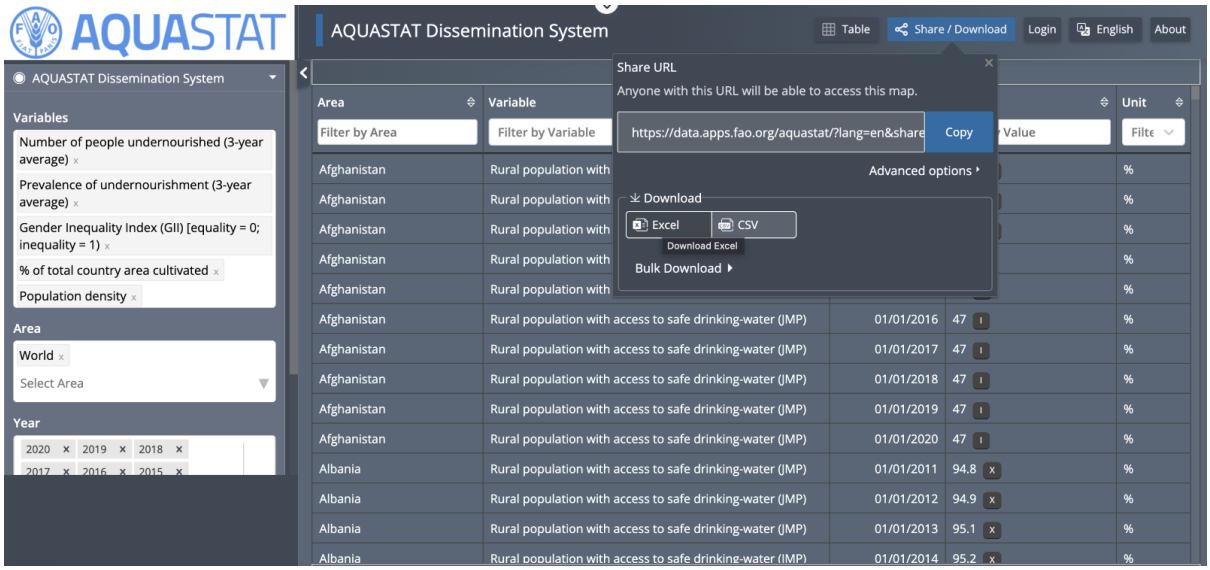


Figure 5.1: Dataset from AQUASTAT - Global Information System on Water and Agriculture

2) The second dataset was taken from FAOSTAT – Suite of Food Security Indicators. https://www.fao.org/faostat/en/#data/FS
Categorized into four dimensions (availability, access, utilization, and stability), data from 2013-2022 for Algeria, Bangladesh, Brazil, Kenya, and the USA were collected

for this assignment. The final dataset, comprising 17 columns, underwent cleansing due to issues such as duplications, errors, and omissions.

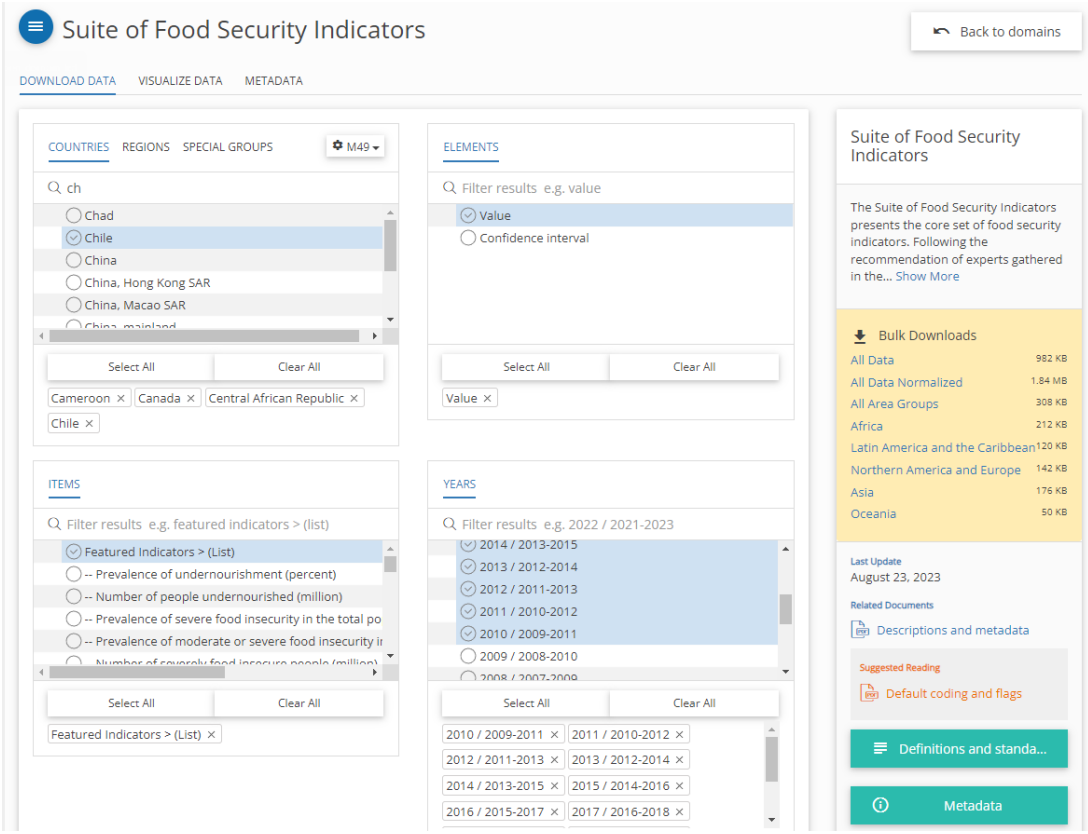


Figure 5.2: Dataset from FAOSTAT – Suite of Food Security Indicators

3) The second dataset was taken from FAOSTAT – Suite of Food Security Indicators. https://www.fao.org/faostat/en/#data/GW. It covers greenhouse gas (GHG) emissions and related activity data from pre- and post-agricultural production stages within agri-food systems. Data from 2012-2021 for Canada, China, France, Russia, and the USA were collected, resulting in a final dataset with 49 columns. Similar to the other datasets, despite the valuable data on the website, cleansing was necessary to address issues like duplications, errors, and omissions.

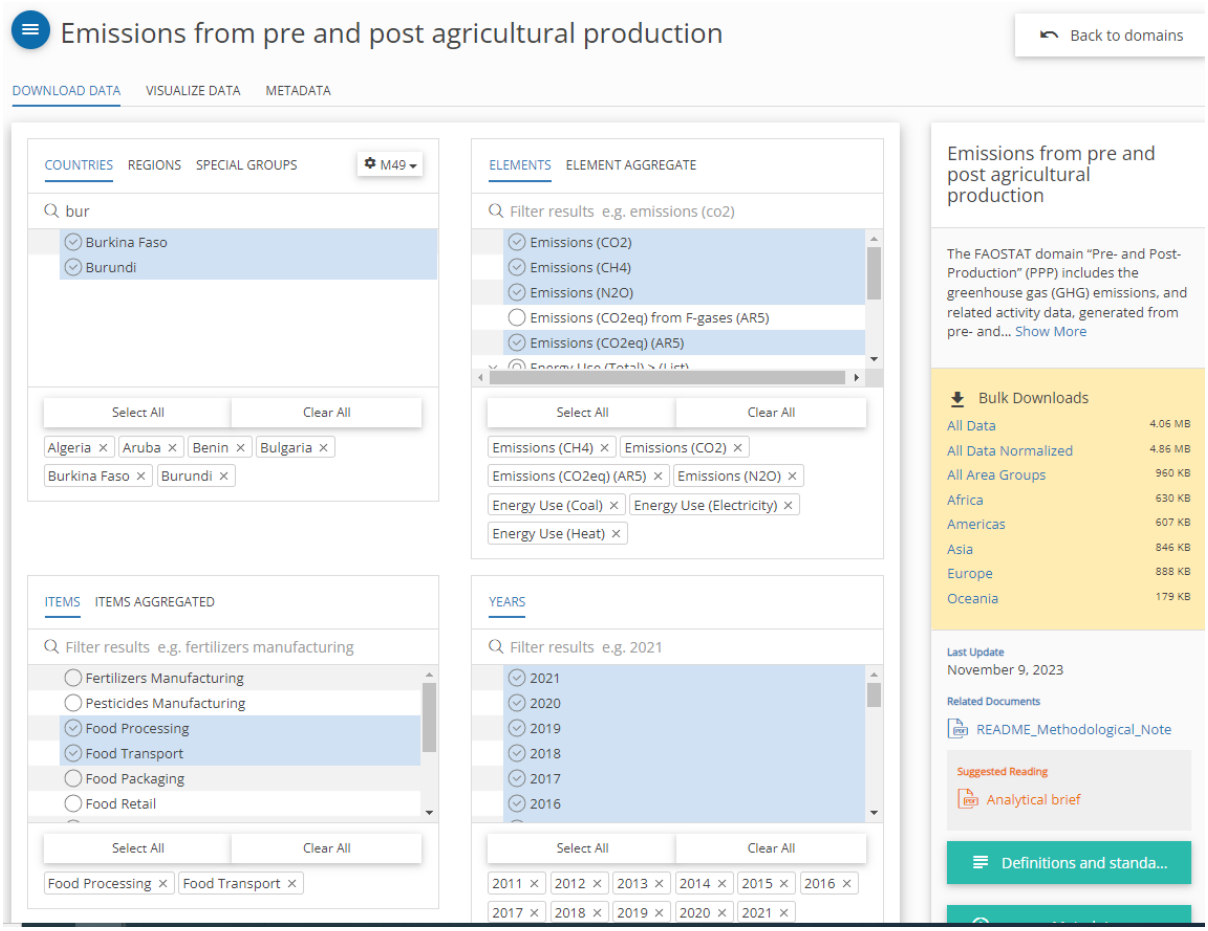


Figure 5.3: Dataset from FAOSTAT – Suite of Food System Waste Disposal

5.3 Research Questions to be Solved.

All three datasets underwent an intensive cleaning process of deleting unwanted rows, reformatting data and using pivot.

Examples of cleaning given below:

a. The Data was first opened in excel and after first glance, all the unwanted and irrelevant columns were deleted.

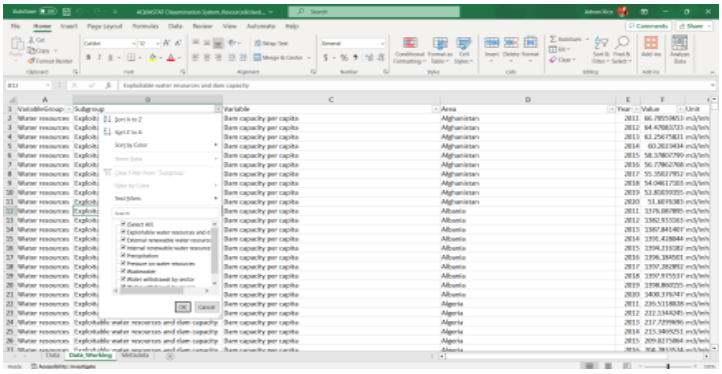


Figure 5.4: Deleting Columns from AQUASTAT

The non-numerical values (#NUM!) were replaced with blank characters and arithmetic formulas were used to aggregate columns.

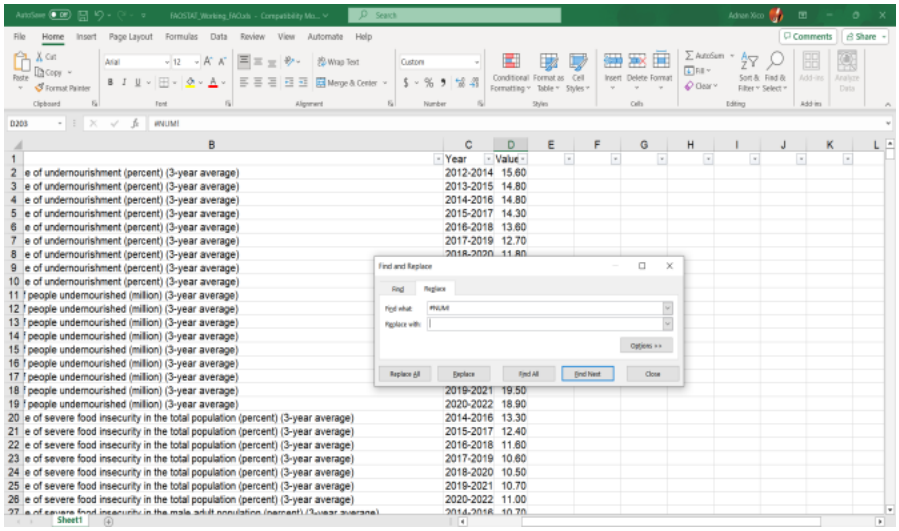
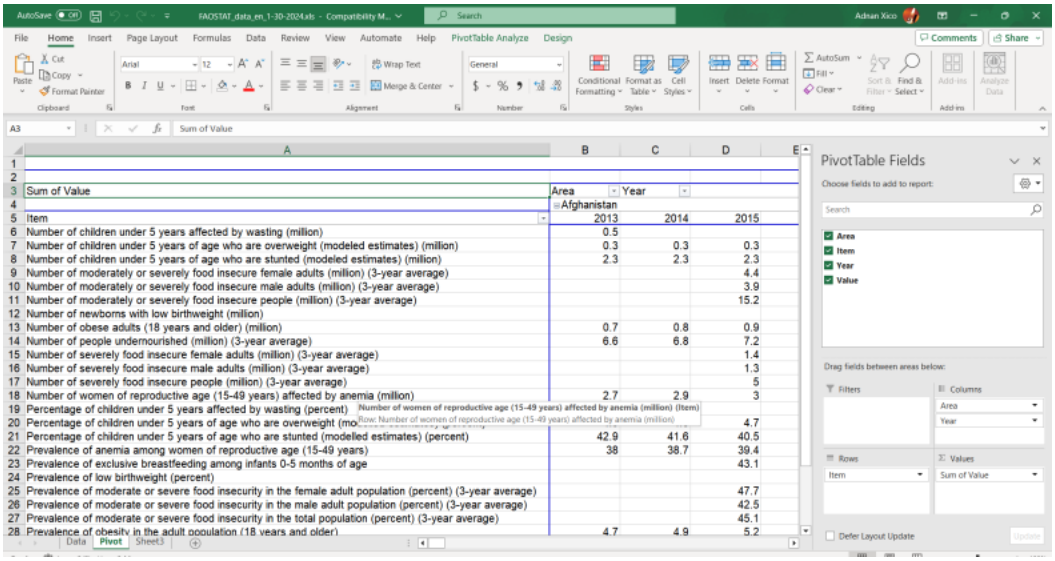


Figure 5.5: Replacing Non-Numerical Value

b. Creating pivot tables after data anomalies are fixed to make better sense of the data at hand and make analysis easier.



Sum of Value	Area	Year	
	Afghanistan		
Item	2013	2014	2015
Number of children under 5 years affected by wasting (million)	0.5		
Number of children under 5 years of age who are overweight (modeled estimates) (million)	0.3	0.3	0.3
Number of children under 5 years of age who are stunted (modeled estimates) (million)	2.3	2.3	2.3
Number of moderately or severely food insecure female adults (million) (3-year average)			4.4
Number of moderately or severely food insecure male adults (million) (3-year average)			3.9
Number of moderately or severely food insecure people (million) (3-year average)			15.2
Number of newborns with low birthweight (million)			
Number of obese adults (18 years and older) (million)	0.7	0.8	0.9
Number of people undernourished (million) (3-year average)	6.6	6.8	7.2
Number of severely food insecure female adults (million) (3-year average)			1.4
Number of severely food insecure male adults (million) (3-year average)			1.3
Number of severely food insecure people (million) (3-year average)			5
Number of women of reproductive age (15-49 years) affected by anemia (million)	2.7	2.9	3
Percentage of children under 5 years affected by wasting (percent)			
Percentage of children under 5 years of age who are overweight (mo			4.7
Percentage of children under 5 years of age who are stunted (modelled estimates) (percent)	42.9	41.6	40.5
Prevalence of anemia among women of reproductive age (15-49 years)	38	38.7	39.4
Prevalence of exclusive breastfeeding among infants 0-5 months of age			43.1
Prevalence of low birthweight (percent)			
Prevalence of moderate or severe food insecurity in the female adult population (percent) (3-year average)			47.7
Prevalence of moderate or severe food insecurity in the male adult population (percent) (3-year average)			42.5
Prevalence of moderate or severe food insecurity in the total population (percent) (3-year average)			45.1
Prevalence of obesity in the adult population (18 years and older)	4.7	4.9	5.2

Figure 5.6: Creating Pivot Table for the Datasets

Finally for all three datasets, the Pivot Table from each of the three datasets was duplicated and transferred to a new sheet, utilizing the "Values & Transpose" option under Paste Special. Subsequently, Excel's Conditional Formatting function was employed to identify and highlight blank values. Following this step, columns (representing variables) and rows (depicting countries) with insufficient data were removed (refer to Figure 8). With these actions concluded, the prepared dataset was now devoid of unnecessary entries and ready for analysis.

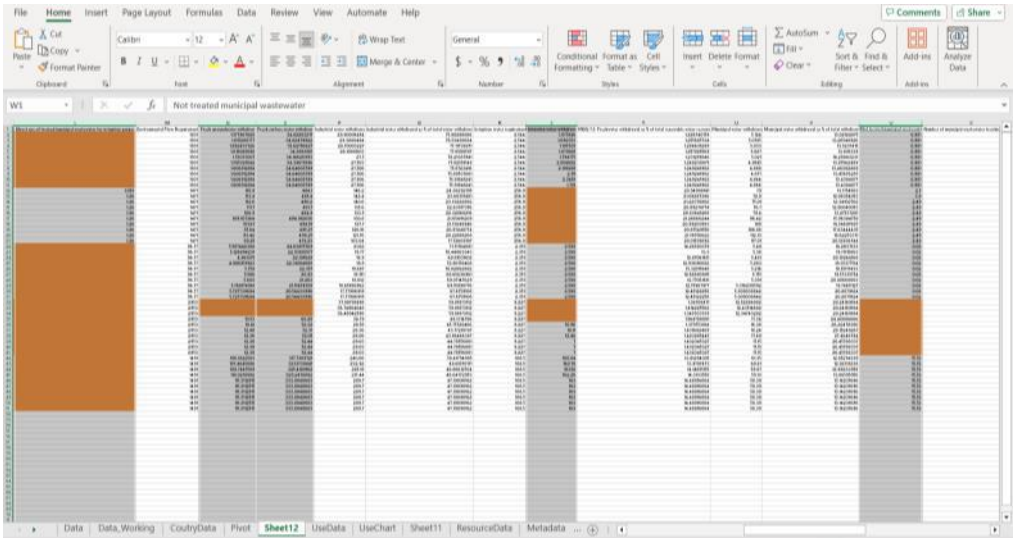


Figure 5.7: Deleting Columns and Rows Lacking Sufficient Data.

5.4 Analysis of the Research Questions using Excel Pivot Tables

a. For the AQUASTAT dataset

Question 1: What is the treated municipal wastewater trend for Canada, China, France, Russia, and the USA?

Choose the legend series to be country, Axis to be year and value to be the sum of treated municipal wastewater. Choose a line graph to showcase the visualization.

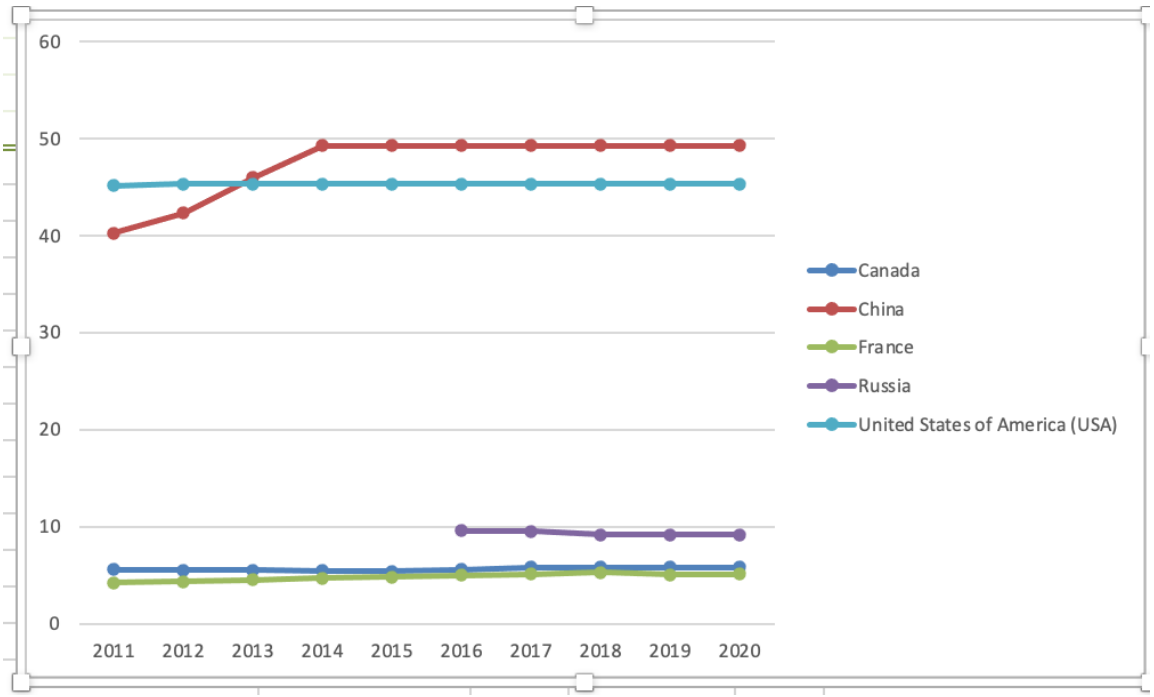


Figure 5.8: Treated Municipal Wastewater Trend for Canada, China, France, Russia, USA

A: The graph shows that China has the highest trend while France has the lowest trend in terms of their treated municipal wastewater.

Question 2: Which country had the highest and lowest dam capacity in 2016?

The field names include Country, Year and Dam capacity per capita. Apply a filter on the year to show only the 2016 data. Assign country to legend, year to the axis and sum of dam Capacity to values. Choose a bar graph to represent the visualization.

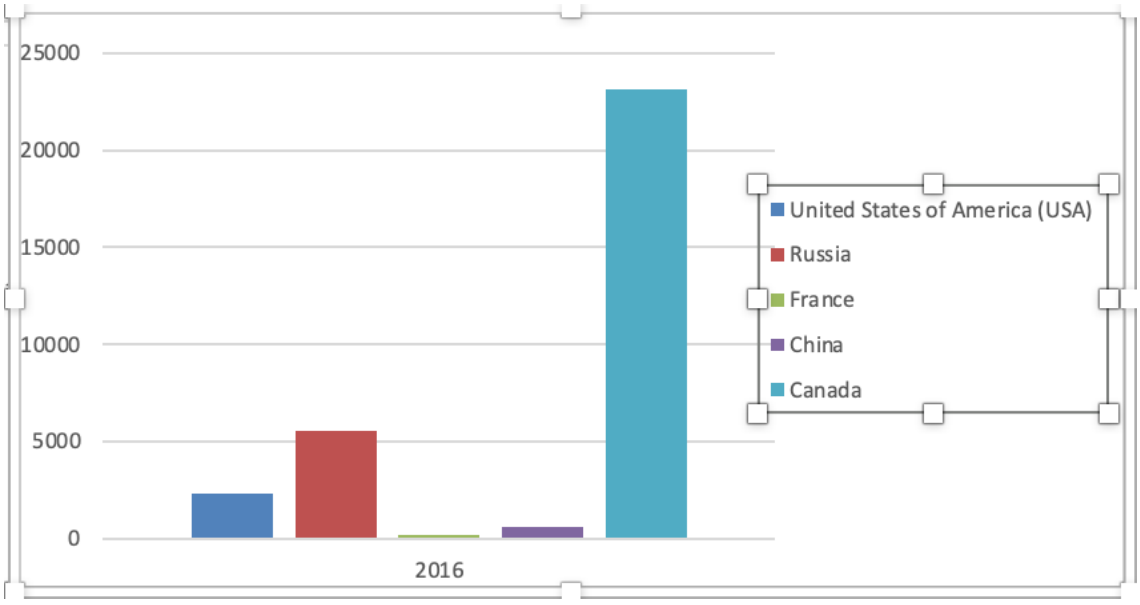


Figure 5.9: Dam Capacity for Canada, China, France, Russia, USA in 2016.

A: The highest dam capacity belongs to the USA and the country with the lowest dam capacity is France.

b. For the FAOSTAT – Food System Waste Disposal dataset

Question 1: Which are the top 3 countries that have the highest emission of gases in 2017?

To get this bar visualization, I chose count of emissions as values, the year as legend and the area as the axis. I filtered the year to show only 2021 data.

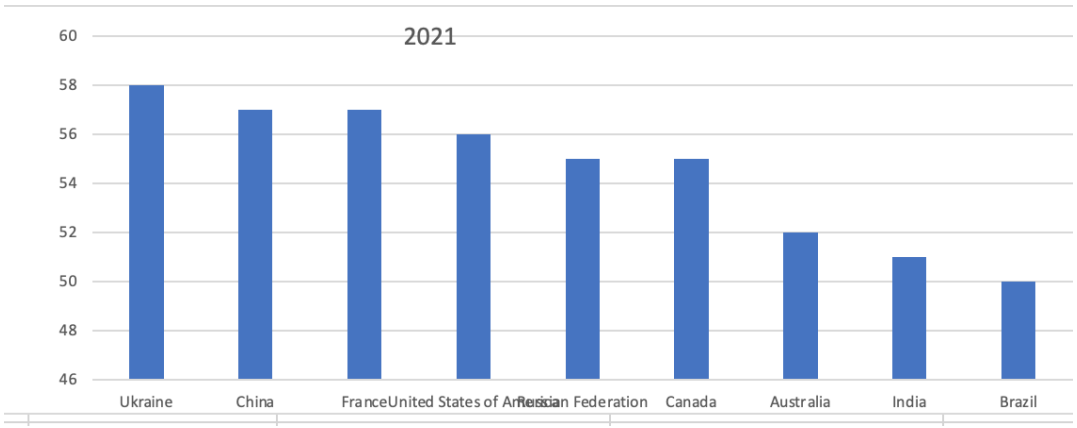


Figure 5.10: Countries that had the Highest Emission of Gases in 2017

A: The top 3 countries that have the highest emission of gases in 2017 are Ukraine, China, and France.

Question 2: Name the largest and the lowest sector using electricity in China.

To generate the visualization below, the "Country" variable was assigned to Filters, the variable name "Source of Gas and Energy usage" was included in Rows, and the variables "Sum of AR5 and Electricity value" were placed in the Values. To enhance clarity, a Pie Chart was created by selecting the pertinent data.

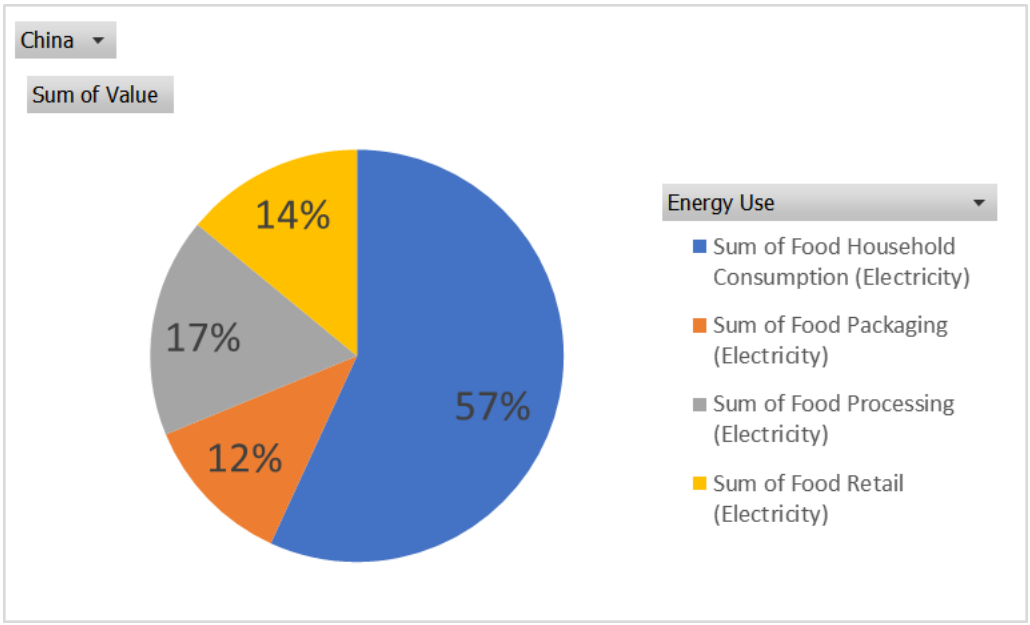


Figure 5.11: Source of Electricity Usage in China

A: The largest usage is by the Food Household sector at 57% while the lowest usage is by Food Retail at 12%.

c. For the FAOSTAT – Food Security Indicators dataset

Question 1: Which country has the highest percentage of undernutrition individuals?

To get the visualization, choose the columns as a country, rows as year and values as sum of the prevalence of undernourishment. I chose a line graph to show the rise and fall of numbers pertaining to Algeria, Bangladesh, Brazil, Kenya, and the US.

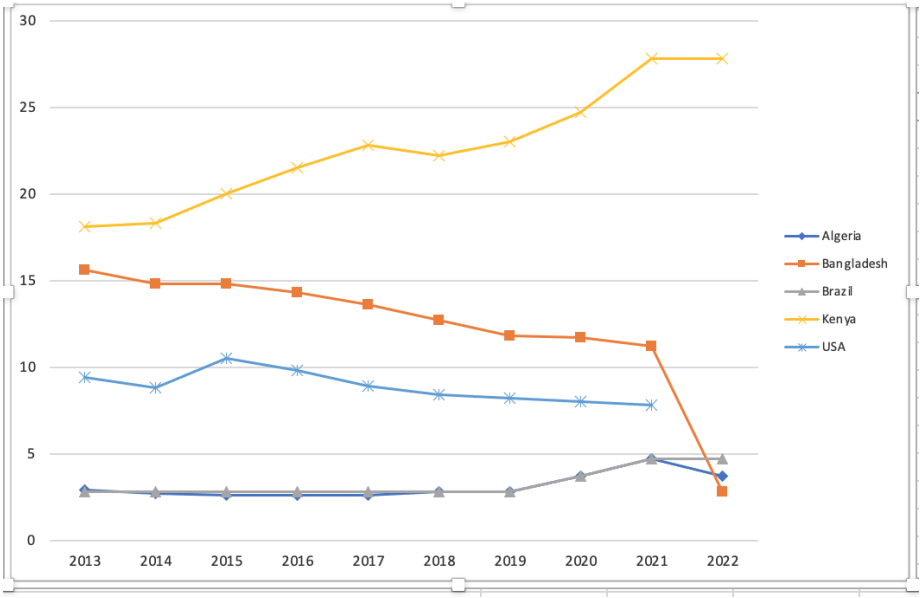


Figure 15.2: Percentage of Undernutrition Individuals

A: The country with the highest percentage of undernutrition individuals is Kenya, which is rising exponentially from 2013 to 2022.

Question 2: What is the trend of the average number of overweight children under 5 years in the US?

Choose country in columns, year in rows and values as the number of children under 5 years of age who are overweight. Change the value to the average of the numbers. For the visualization, I used a line graph to show the rise and fall through the years.

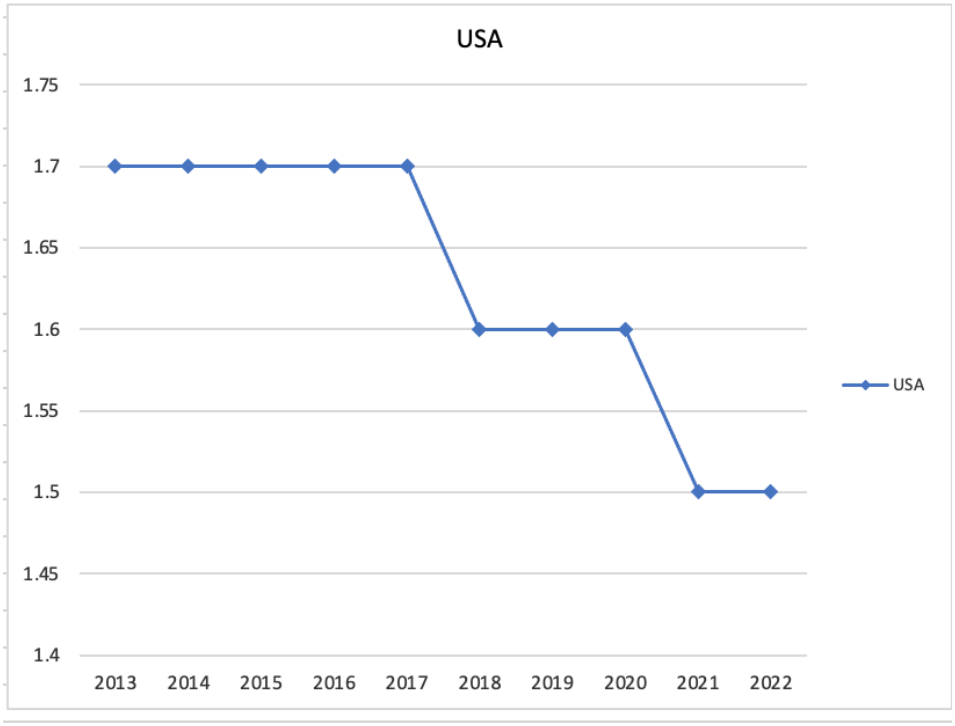


Figure 5.13: Average Number of Overweight Children under 5 Years in the US

A: The highest average of overweight children under 5 years was in 2013 and the lowest is in 2022. This shows that health consciousness has increased in the US over the decade.

5.5 Conclusion

As a student who has delved into the realms of data analysis using AQUASTAT and FAOSTAT, alongside harnessing the power of Excel pivot tables for data cleaning, this journey has been a transformative exploration into the intricate world of business analytics. Navigating through these online platforms, I have unearthed a treasure trove of data, providing invaluable insights into global water resources and agricultural trends.

The practical application of Excel pivot tables in data cleaning has been a game-changer. These tools have not only streamlined the process of refining raw data but have also allowed for a comprehensive analysis that goes beyond the surface. The agility and versatility of pivot tables in addressing inconsistencies and outliers have been particularly noteworthy, paving the way for a meticulously curated dataset ready for deeper exploration.

From my perspective, the use of Excel pivot tables in business analysis presents a spectrum of advantages. Firstly, the user-friendly interface makes it accessible even to those with limited technical expertise, fostering a quick learning curve. The visual representation of data through pivot tables enhances interpretability, making it easier to communicate findings to stakeholders. Furthermore, the flexibility of Excel allows for iterative analyses and adjustments, enabling a dynamic and responsive approach to changing data landscapes.

However, as with any tool, there are considerations to be mindful of. The dependence on Excel may pose challenges when dealing with exceptionally large datasets, potentially impacting performance and efficiency. Additionally, while pivot tables offer robust functionalities, more complex analyses may necessitate the integration of additional tools or programming languages.

In conclusion, this hands-on experience with AQUASTAT, FAOSTAT, and Excel pivot tables has not only enriched my technical skill set but has also underscored the multifaceted nature of business analysis. The advantages in terms of accessibility and visualization are substantial, providing a solid foundation for data-driven decision-making. Nevertheless, a nuanced understanding of the limitations and potential challenges ensures a balanced and informed approach to utilizing these tools for effective data analysis in real-world scenarios. This journey has illuminated the dynamic interplay between theory and practice, empowering me as a student to navigate the evolving landscape of data science with confidence and proficiency.

Chapter - 6: Practical Analysis Using SAP

(Unsupervised Learning and Supervised Learning)

6.1 Introduction to SAP (Unsupervised Learning Method)

SAP Analytics Cloud (SAC) is a cloud-based analytics platform offered by SAP that integrates business intelligence, augmented analytics, and planning capabilities. It allows users to visualize, analyze, and share insights from various data sources. While SAC is not primarily designed for machine learning tasks like unsupervised learning, it does offer functionalities that can support such analyses, including k-means clustering.

In SAP Analytics Cloud (SAC), users can harness a range of capabilities for both supervised and unsupervised learning, enabling them to extract actionable insights from their data. Here's a structured approach to utilizing SAC effectively for advanced analytics tasks:

6.1.1 Supervised Learning:

For supervised learning tasks, such as regression or classification, users can select relevant features from the dataset and train predictive models within SAC. These models can then be evaluated for performance metrics like accuracy, precision, and recall. Once satisfied with the model's performance, it can be deployed for real-time predictions, facilitating data-driven decision-making.

Time analysis, in the context of analytics, refers to the process of analyzing data over time to identify trends, patterns, and insights. Time analysis is crucial for understanding how data evolves and performs over different periods, which is essential for making informed decisions and forecasting future trends.

In SAP Analytics Cloud (SAC), time analysis allows users to:

1. Analyze time series data for trends and patterns.
2. Filter and slice data based on specific time periods.
3. Create custom calculations and forecasts.

4. Compare data across different time periods.

5. Create dynamic reports and dashboards with time-based narratives.

6.1.2 Unsupervised Learning:

SAC provides robust options for unsupervised learning, including clustering and anomaly detection techniques. Users can reduce dimensionality and identify underlying patterns in the data through these methods. Visualizations within SAC help in interpreting these patterns, providing valuable insights into hidden trends or anomalies.

Integration and Monitoring:

Integration of insights derived from both supervised and unsupervised learning with other analytics processes is seamless within SAC. Users can incorporate these insights into reports, dashboards, or further analysis. Moreover, SAC facilitates continuous monitoring of model performance, enabling timely updates and refinements as needed to ensure the accuracy and relevance of insights over time.

K-means clustering: It is a popular unsupervised machine learning algorithm used for clustering data into groups based on similarity. It aims to partition n data points into k clusters in which each point belongs to the cluster with the nearest mean. The value of k is predetermined and represents the number of clusters the algorithm will create.

It has the following features:

- Cluster data is partitioned based on similarity.
- Centroid-based which means it iteratively assigns points to nearest centroids.
- Requires initial selection of k centroids.
- Minimizes within-cluster sum of squares (WCSS) by iteratively optimizing the cluster until convergence.
- Efficient and simple, suitable for large datasets.
- Deterministic; results depend on initial centroids.
- Sensitive to initialization; multiple runs may help.
- Assumes spherical clusters, struggles with irregular shapes.
- Scalable but may struggle with high dimensions.
- Requires predefined k, limiting flexibility if the optimal number of clusters is unknown.

SAC can therefore be effectively used for unsupervised learning in the following ways:

1. Data Preparation: SAC enables users to connect to diverse data sources and preprocess data for analysis, including cleaning, transformation, and manipulation.

2. Visualization: SAC offers interactive visualization tools, facilitating visual exploration of data to discern patterns, trends, and potential clusters before conducting k-means clustering.

3. K-Means Clustering: While lacking native support, SAC allows users to employ custom scripting or integrate external tools like Python or R within its data modelling capabilities for implementing k-means clustering.

4. Integration with ML Platforms: SAC integrates seamlessly with SAP and third-party machine learning platforms that offer k-means clustering algorithms, enabling users to conduct clustering analysis on SAC data.

5. Advanced Analytics: SAC includes advanced analytics features like predictive analytics and smart insights, enhancing clustering analysis with supplementary insights and context.

6. Collaboration and Sharing: SAC facilitates collaboration and knowledge sharing within organizations by enabling users to disseminate clustering results, visualizations, and dashboards, fostering data-driven decision-making.

6.2 Explanation of the Dataset used

6.2.1 Dataset used for Unsupervised Learning

The file used for this exercise is Stores.csv. The dataset consists of 150 rows and 5 columns out of which 4 are measures and 1 is a dimension. Measures are the key figures or Measures are the key figure or facts which are used to make calculations in this case are the Sales Turnover, Store Size, Staff Size and Profit Margin.

The Dimension columns give more information regarding a measure, which in this case is Store (Store Name). It is used to examine 150 stores and come up with promotion strategies for the future by dividing the stores into 3 segments based on the measures. A few changes were made to the dataset post-uploading to help facilitate better reporting.

The given dataset is organized into meaningful groups to mine the data efficiently using a measure of association. The result of the analysis is a set of clusters using K- Means clustering which is a method of finding clusters and their centres (R) given a choice in the number of clusters (K). It is often used for market segmentation. The goal is to make
the inter-cluster difference (distance) high and the intra-cluster difference (distance) low.

Follow the below steps to analyse the dataset using K-Means Clustering:

The file is uploaded to SAC by creating a new project canvas and selecting the “Classic Design Experience” option.

Once the data is imported successfully → select Story View → insert Chart → Bubble chart.

- Add the following Measures.
- Add Sales Turnover to the X-Axis.
- Add Staff Size to the Y-Axis.
- Add Profit Margin to Size.
- Add Store to the Dimensions.
- Add a Tooltip Measure as shown in Figure

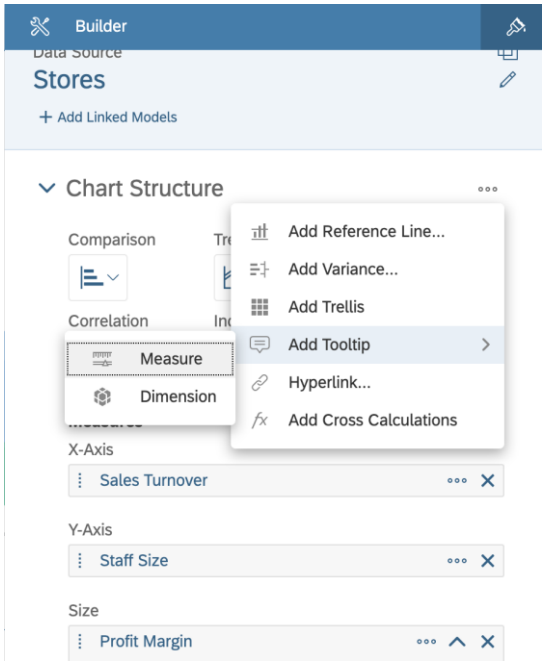


Figure 6.1: Adding a Tooltip

The Tooltip Measures will now show as a Chart Structure option. Add Store Size to Tooltip Measures which results in a bubble chart of the first three measures by Store.

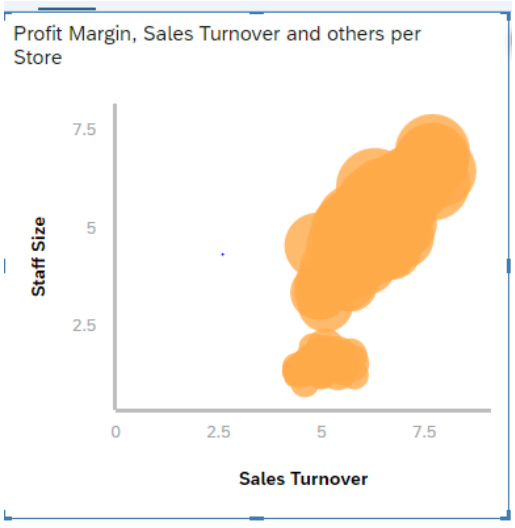


Figure 6.2: A Bubble Chart of Store Data

1. Creating the Cluster analysis:

Using K-means the stores of the above graph are grouped using Smart Grouping:

- Toggle on Smart Grouping (near the bottom of the Builder
- panel).
- Change the Number of Groups to 3 where 3 is k in the k-means algorithm.
- Change the Group Label to “Cluster” just to be consistent with our understanding of cluster analysis.
- Select Include Tooltip Measures in grouping so that all four measures are considered in the cluster analysis.

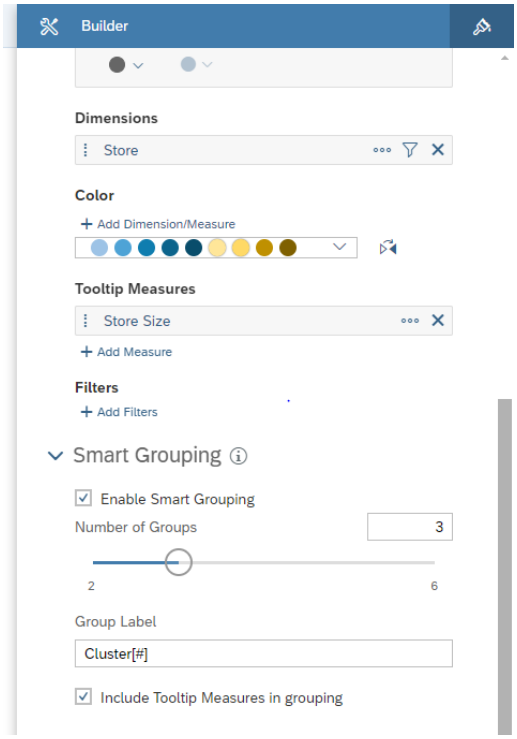


Figure 6.3: Configure Smart Grouping for creating clusters.

- Change the colour pallet so that filters the different groups of the cluster can be distinguished and examined.
- Change the range of the X-axis to 4~9 to shorten it.

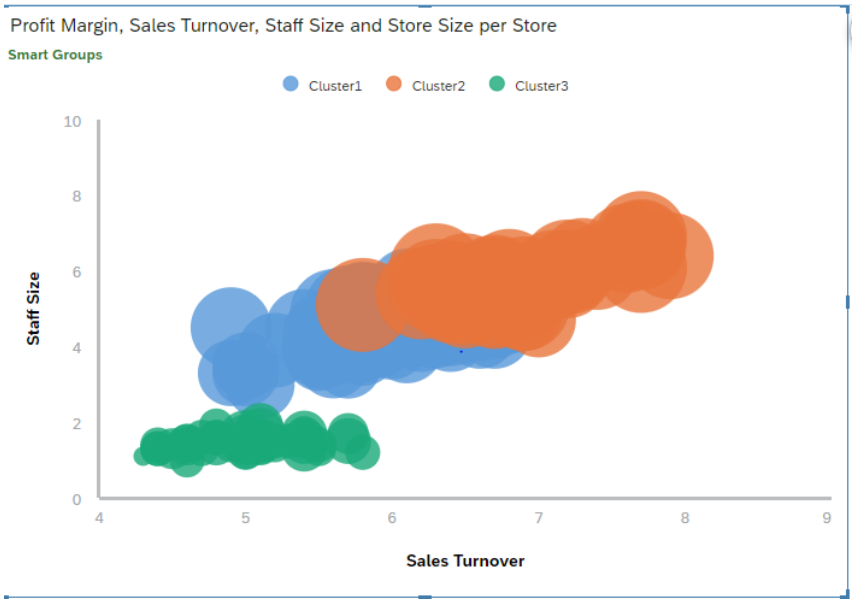


Figure 6.4: Clustered Bubble Chart with Measures per Store

2. Visualization and Interpretation:
   - Use the filter option on each cluster to filter them separately and export each one as a .csv file to your computer.
   - Follow this 3 times to get 3 files called cluster 1.csv, cluster 2.csv and cluster 3.csv.
   - These files will contain 4 columns (Measure, Sales Turnover, Staff size and Profit Margin). Add another column manually called Cluster and for all rows add the cluster number as 1,2 or 3 accordingly.
   - The resultant file will be as shown below.

	A	B	C	D	E
1	Store	Sales Turnover	Staff Size	Profit Margin	Cluster
2	Anaheim	5.5	4	1.3	1
3	Anchorage	6.1	4.7	1.4	1
4	Arlington	5.8	5.1	1.9	1
5	Aurora	5.7	4.5	1.3	1
6	Baton Rouge	5.4	4.5	1.5	1

Figure 6.5: The .csv file after Data Wrangling.

- Now merge all the three .csv files by going to the Data View's grid mode → Select Add New Data → Select Data uploaded from a file → Select the cluster 1.csv → import → save → Open with Basic Data Preparation (to append the other 2 files) → OK.
- Reimport the data from the Data ribbon → Select cluster.csv → Select Append from the dialog box → Finish and repeat the above steps for cluster 3.csv → Save.
- To visualize the Custer data, go to Story View → Add a new page → Add a chart → add a calculated measure for Count of Stores.

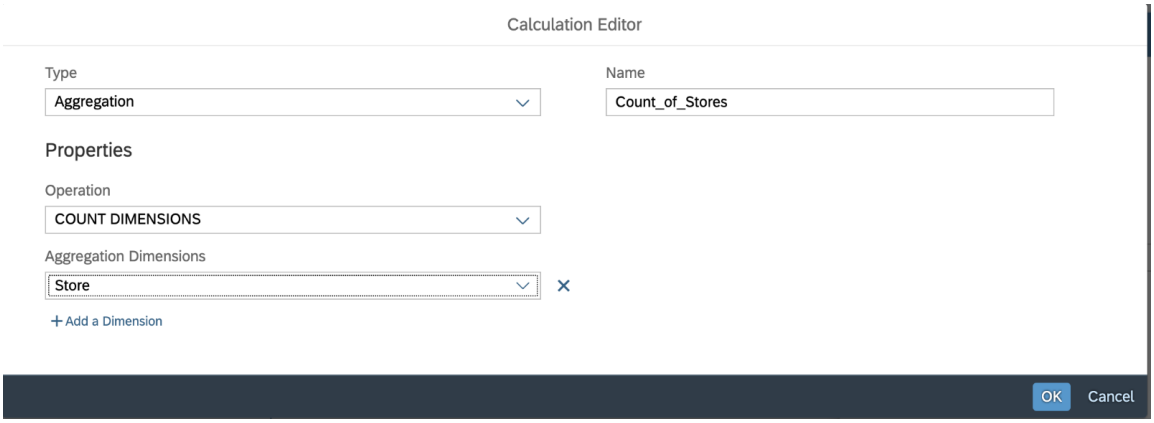


Figure 6.6: Count of Stores Using the Calculation Editor.

- In Builder, select the Link Dimensions icon to the right of Data Source. You should see Clusters as a choice.
- Select Clusters by clicking on the box and choose the matching dimensions from each data set.
- In our case, the "Store" dimension in the Stores data matches the "Store" dimension in the Cluster 1 data.
- Select Data Samples > ID to see samples of the values that will be linked. The Link
- Dimension settings are shown below. → 'Set.'

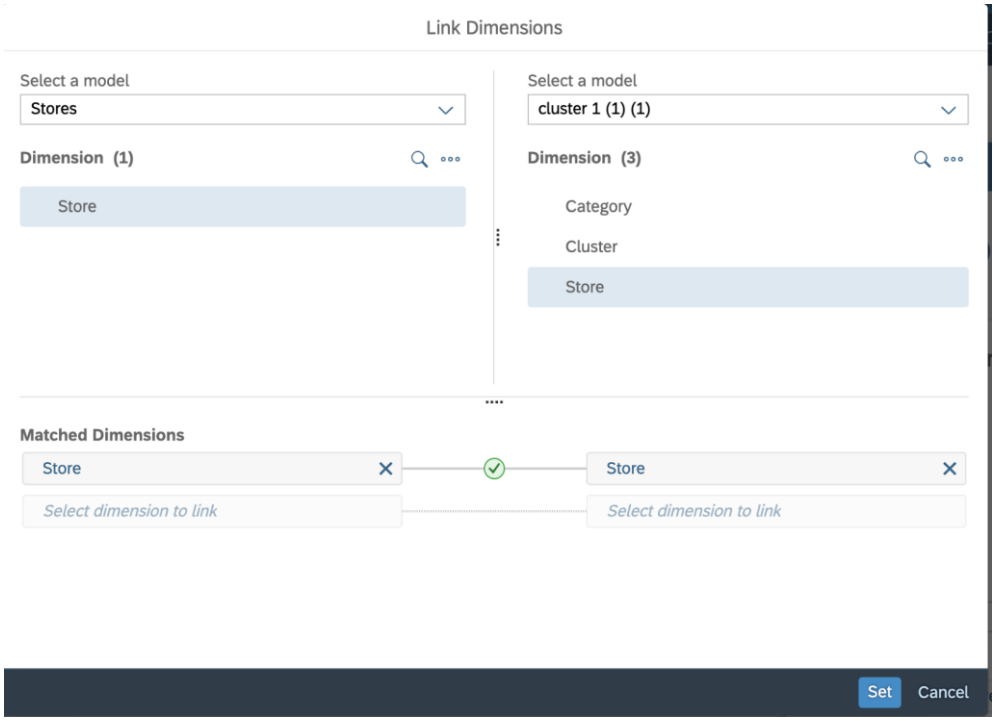


Figure 6.7: Link Dimensions

- Add variables to the Column chart, which is called a blended data chart in SAP.
  - Add Count of Stores from the Store data set to Measures.
  - Add Cluster from the Clusters data set to Dimensions.

Figure 6.8: Cluster with the Highest Number of Stores

The resultant graph shows that the cluster with the highest number of stores is cluster 1, having 61 stores.

6.2.2 Dataset used for Supervised Learning

The dataset used for the time analysis exercise comprises sales transactions spanning from 2008 to 2020. To generate a forecast, it's essential to aggregate the details into monthly periods, a task easily accomplished within a private model in SAP Analytics Cloud (SAC). Here's a summarized guide:

1. Navigate to "Stories" in the left menu.
2. Create a new canvas with the "Classic Design Experience."
3. Add data by uploading the file "GBSales_transactions.xlsx."
4. On the Story's Data tab: Change "Year" and "Month" to dimensions where the data type is set to "String" and statistical type is "Continuous" for both "Year" and "Month."
5. Concatenate "Month" and "Year" using the transformation function, resulting in a column named "Month_Year."

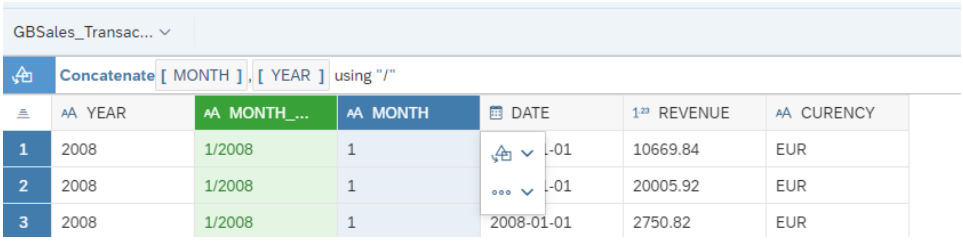


Figure 6.9: Concatenating the Month and Year Column

6. Create a Table view in the Story, including "Revenue" as the measure and "Currency" and "Month_Year" as rows.

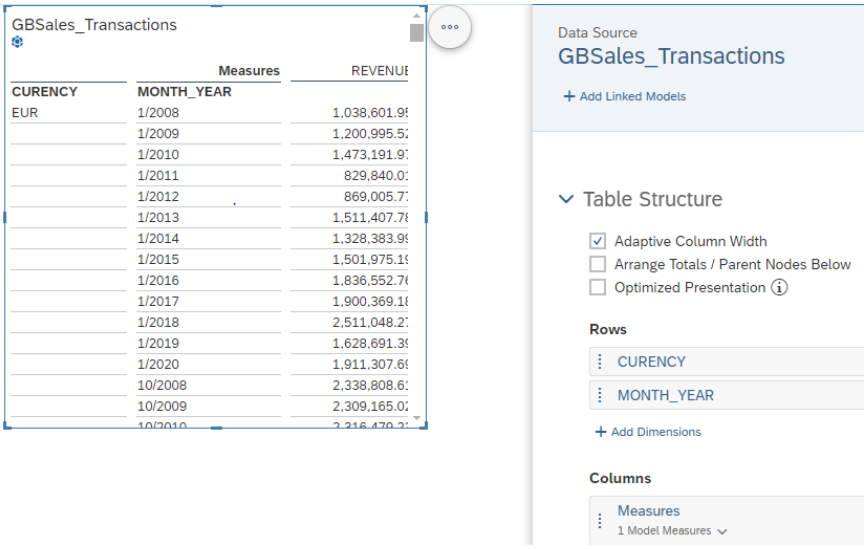


Figure 6.10: Creating the Table

7. Export the table results using the Export function under the table's More Actions menu. The downloaded file is as shown below (Figure 6.10).

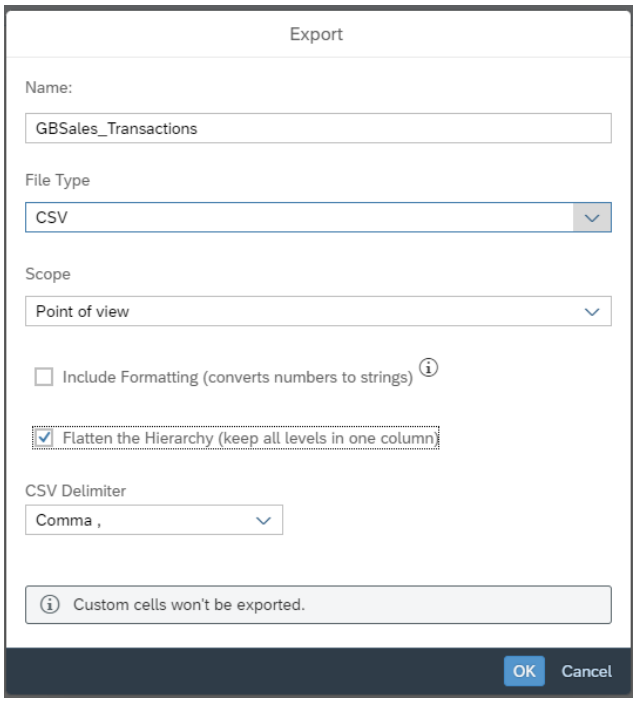


Figure 6.11: Exporting the Table Results

Open the downloaded .csv file, resembling Figure 4. Follow these steps:

1. Move the "REVENUE" heading to row 2 and delete row 1.

2. Save the file as an Excel spreadsheet, optionally as a .csv. Example: GBSales_transactions_aggregated.xlsx.

	Measures	REVENUE
CURENCY	MONTH_YEAR	
EUR	Jan-08	1038602
EUR	Jan-09	1200996
EUR	Jan-10	1473192
EUR	Jan-11	829840
EUR	Jan-12	869005.8

Figure 6.12: The .csv File from SAC

To create the aggregated data set and forecasting model in SAC Predictive Scenario, follow these steps:

To Create Aggregated Data Set:

- Go to Datasets from the left menu.
- Select "Create New From a CSV or Excel File."
- Choose the source file "GBSales_transactions_aggregated.xlsx" and import it.
- Select your folder, give it a name like "Global Bike Forecast Dataset," and add a description.
- Save the dataset for future analyses.

To Create Forecasting Model:

- Navigate to Predictive Scenarios --> Choose "Time Series Forecast."
- Select your folder and name the model, e.g., "Sales Forecast." Then add a description like "Forecast of Global Bike sales" and proceed.
- Configure settings: Describe the model --> Select the dataset created earlier as the Time Series Data Source.
- Set "Revenue" as the Signal (target) --> Choose "Month_Year" for Date.
- Set the Number of Forecasts to 12 (for 12 months).

- Select "Currency" for Entity to differentiate between U.S. and German sales.
- Click on Train & Forecast. Be patient while the model is trained. Sometimes it takes a minute or two. The results will be shown on three tabs or pages: Overview, Forecast, and Explanation. Be sure to expand the Explanation visuals to include all components.

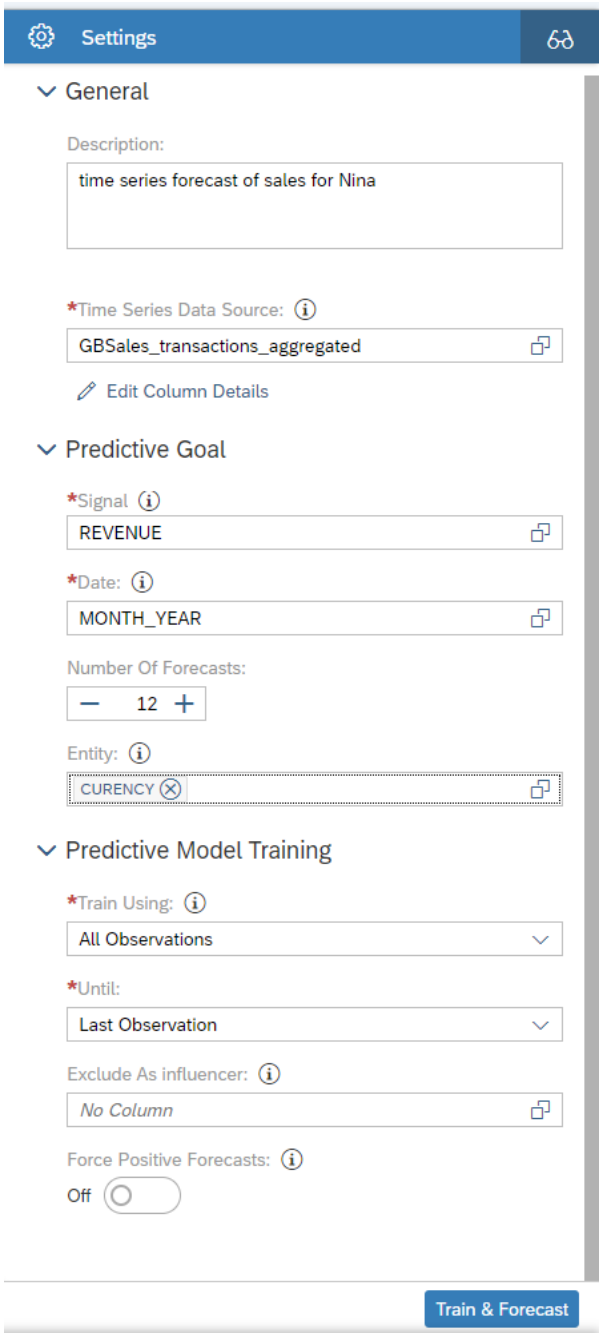


Figure 6.13: Predictive Model Training

- Save the Sales Forecast Results in your folder. The Save Forecast function is circled in

the figure below. Provide a name of your choosing for the saved forecast. I named it "Sales Forecast Results".

6.3 Research Questions to be solved

6.3.1 Research Questions to be Solved for Unsupervised Learning

Question 1: List the name of one store in each cluster.

Question 2: How does the average profit margin, average sales turnover, and average staff size compare amongst the clusters?

Question 3: What can the manager do with these segmentation results?

Question 4: What does the density of a cluster measure? Why is it important?

6.3.2 Research Questions to be Solved for Supervised Learning

Question 1: Define MAPE. What is the significance of MAPE in this forecast?

Question 2: What are the forecasted sales for December 2021 in Germany? What are the forecasted sales for December 2021 in the U.S.?

Question 3: Are there any signal outliers in the US data? If so, what are they? How might these types of outliers affect this analysis?

Question 4: Use the model Explanation for the U.S. forecast and the information from the textbook to explain what the triple exponential smoothing is. How are alpha, beta, and gamma represented in the Explanation? Include a screenshot.

6.4 Outcome Analysis of the Research Questions using SAP Analytics Cloud

6.4.1 Outcome Analysis of the Research Questions for Unsupervised Learning

Question 1: List the name of one store in each cluster.

To answer this question, a cluster bubble was created where the Count_of_Stores was chosen as the measure and the Store as the dimension. This resulted in a bubble graph containing all the names of the stores color-coded to represent the cluster they belong to.

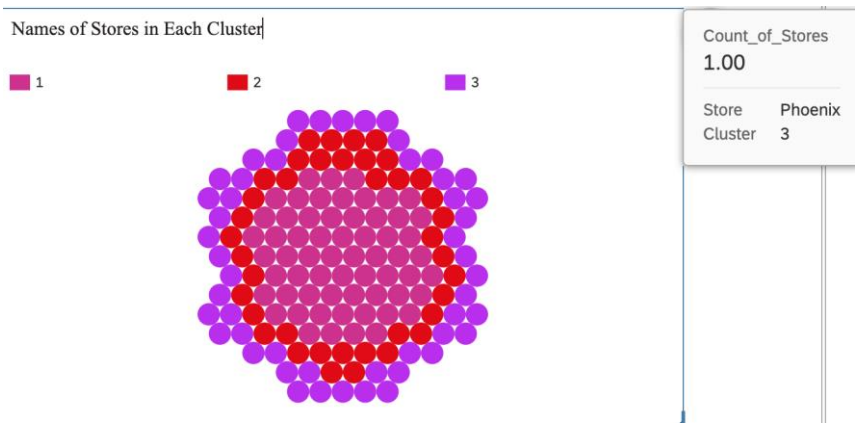


Figure 6.14: Cluster Bubble that represents all the Stores in different clusters.

Answer: The name of a store from cluster 1 is Riverside.

The name of a store from cluster 2 is Lancaster.

The name of a store from cluster 3 is Phoenix.

Question 2: How does the average profit margin, average sales turnover, and average staff size compare amongst the clusters?

To answer this question, a bar graph was used to represent the outcome where the Profit Margin, Sales Turnover and Staff Size were added as measures and the Cluster was chosen as the dimension. The resultant graph has 9 vertical bars, 3 bars for each cluster.

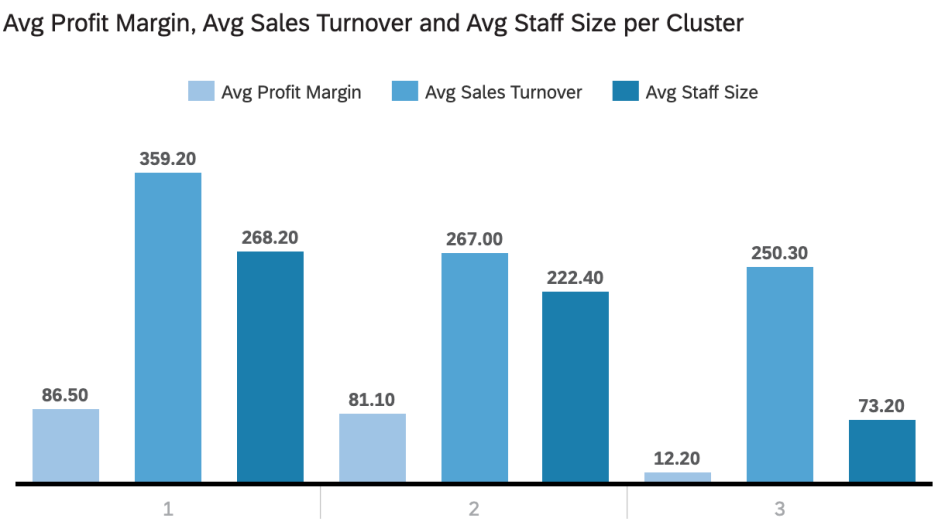


Figure 6.15: Avg Profit Margin, Avg Sales Turnover and Avg Staff Size for Each Cluster.

Answer: The Average Profit Margin is highest for Cluster 1 (86.50) and lowest for Cluster 3 (12.20).

The Average Sales Turnover is highest for Cluster 1 (359.20) and the lowest for Cluster 3(250.30).

The Average Staff Size is highest for Cluster 1 (268.20) and lowest for Cluster 3 (73.20).

Question 3: What can the manager do with these segmentation results?

Answer: Once a manager obtains segmentation results, they can:

1. Understand customer groups.
2. Customize marketing campaigns.
3. Inform product development.
4. Optimize pricing strategies.
5. Improve customer experience.
6. Allocate resources effectively.
7. Predict churn and retention.
8. Track performance for ongoing optimization.

Question 4: What does the density of a cluster measure? Why is it important?

Answer: The density of a cluster measures how tightly packed or concentrated the data points within that cluster are. It quantifies the closeness of points to each other within the cluster space. This measure is important because it provides insights into the coherence and homogeneity of the cluster.

Density in clustering is crucial for various reasons:

- Quality Assessment: It helps gauge cluster quality, with higher density indicating well-defined clusters.
- Cluster Separation: Density aids in distinguishing between clusters, making it easier to identify distinct groups.
- Interpretation: Understanding density helps interpret cluster characteristics, with high density indicating dominant patterns.
- Algorithm Performance: Density influences algorithm performance, aiming for high intra-cluster density and low inter-cluster density.
- Data Exploration: Visualizing cluster densities helps users understand data structure through techniques like heatmaps.

6.4.2 Outcome Analysis of the Research Questions for Supervised Learning

Question 1: Define MAPE. What is the significance of MAPE in this forecast?

Answer: MAPE stands for Mean Absolute Percentage Error. It is a metric used to evaluate the accuracy of a forecasting model by measuring the average absolute percentage difference between actual and predicted values.

The significance of MAPE in this forecast lies in its ability to provide a standardized measure of forecasting accuracy, allowing analysts to assess the reliability of the sales forecast generated by the predictive model. A lower MAPE indicates higher accuracy, while a higher MAPE suggests greater deviation between predicted and actual values

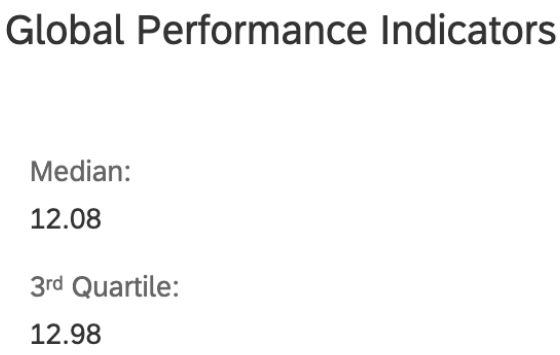


Figure 6.16: Expected MAPE Percentage

Here the median = 12.08 suggests that when the dataset is sorted, If the median is close to the forecasted values, it indicates that the model's predictions are centered around the typical value of the dataset. In this case, a lower MAPE would suggest that the model is accurately capturing the central tendency of the data.

The value at the midpoint 3rd quartile = 12.98, it means that 75% of the dataset's values are below or equal to 12.98 when sorted in ascending order. The 3rd quartile provides information about the spread or variability of the dataset. If the forecasted values align closely with the 3rd quartile, it suggests that the model is capturing the upper range of the dataset accurately.

Question 2: What are the forecasted sales for December 2021 in Germany? What are the forecasted sales for December 2021 in the U.S.?

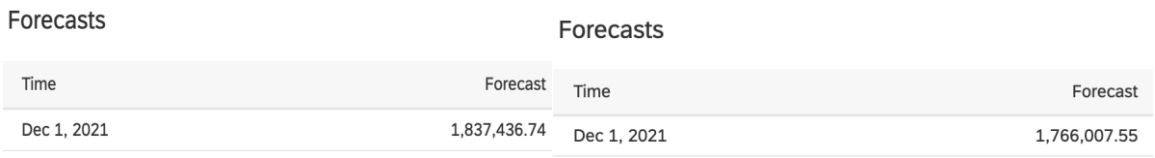


Figure 6.17: Forecasted sales for December 2021 for Germany and the U.S Respectively

Answer: The forecasted sales for December 2021 for Germany is 1,837,436.74 and the forecasted sales for December 2021 for the US is 1,766,07.55

Question 3: Are there any signal outliers in the US data? If so, what are they? How might these types of outliers affect this analysis?

Outliers (Past)

Time	Actual	Forecast
May 1, 2010	6,187,308.11	9,771,399.91
Jun 1, 2010	8,579,709.29	12,310,322.69
Jun 1, 2011	9,343,302.09	12,355,871.78

Figure 6.18: Signal Outliers in the US Data

Answer: There are 3 outliers in the US Data. They are shown in Figure 18.

Signal outliers in time series data can distort trends, bias estimates, and reduce model performance, leading to inaccurate forecasts and misleading insights. They can impact analysis by distorting patterns, biasing estimates, reducing model accuracy, and misleading interpretations. Proper identification and handling of outliers are crucial to maintain the integrity and validity of time series analysis results.

Question 4: Use the model Explanation for the U.S. forecast and the information from the textbook to explain what the triple exponential smoothing is. How are alpha, beta, and gamma represented in the Explanation? Include a screenshot

Figure 6.19: The Model Explanation for the U.S. Forecast

Answer: Triple Exponential Smoothing, also known as Holt-Winters Triple Exponential Smoothing, is a forecasting method used for time series data with seasonality and trend. It extends Holt-Winters method by incorporating three smoothing parameters: level (α), trend (β), and seasonality (γ). It combines these components to forecast future values. This method is valuable for capturing complex patterns in data, like seasonal variations, and is commonly used in industries such as finance and retail for forecasting sales and demand.

6.5 Conclusion

In conclusion, leveraging SAP Analytics Cloud (SAC) for both supervised and unsupervised learning presents a powerful opportunity for organizations to derive actionable insights from their data. While SAC may not inherently support certain unsupervised learning techniques like k-means clustering, its robust suite of data preparation, visualization, and integration capabilities empowers users to conduct advanced analyses effectively. Through SAC's customizable scripting options and seamless integration with external machine learning platforms, such as Python or R, users can implement a wide range of algorithms, including both supervised and unsupervised learning models.

Moreover, SAC's advanced analytics features, including predictive analytics and smart insights, enhance the analysis process by offering additional context and understanding. This seamlessly integrates with the forecasting process, where users leverage SAC's powerful tools for data

analysis, visualization, and predictive analytics. Key steps such as data preparation, model creation, training, and evaluation, supported by metrics like Mean Absolute Percentage Error (MAPE), aid in assessing performance and identifying areas for improvement across both supervised and unsupervised learning tasks.

Understanding descriptive statistics like median and quartiles further enriches the interpretation of model accuracy, providing valuable insights into the distribution and trends within the data. Ultimately, SAC facilitates data-driven decision-making by providing insights into trends, patterns, and forecasts, thereby enabling organizations to optimize processes, drive innovation, and achieve strategic objectives across a spectrum of learning paradigms. The collaborative nature of SAC enhances its utility, enabling organizations to share results and visualizations across teams, fostering informed decision-making in both supervised and unsupervised learning contexts. By harnessing the capabilities of SAP Analytics Cloud, organizations can unlock the full potential of their data to drive organizational success.

Chapter - 7: Tableau Using Text Analysis

7.1 Introduction

Text analysis, crucial for understanding customer sentiment and market trends, is empowered by Tableau, a renowned data visualization tool.

a. Text Parsing and Analysis: Tableau's connectivity enables the seamless integration of diverse textual data sources, aiding researchers in extracting valuable insights into consumer behaviors and preferences.

b. Sentiment Analysis: Tableau employs NLP algorithms for sentiment analysis, enabling businesses to assess customer satisfaction levels and refine strategies accordingly.

c. Topic Modeling and Keyword Analysis: Through integration with advanced techniques, Tableau uncovers latent themes and prevalent topics within textual datasets, driving informed decision-making processes.

d. Text Visualization: Tableau's visualization techniques facilitate the intuitive exploration of textual insights, empowering stakeholders to glean actionable insights dynamically.

e. Integration with Machine Learning Models: Tableau's extensibility allows integration with external machine learning models, enhancing the sophistication of text analysis tasks.

Netlytic is a cloud-based tool for analyzing text and social networks, capable of automatically summarizing and uncovering communication networks within publicly accessible social media content.

7.2 Description of the Tool and the Dataset Used

This exercise focuses on text mining and analysis using an open-sourced wine reviews dataset to create Tableau visualizations, explore customer behavior and suggest strategic recommendations. Key concepts include natural language processing techniques like bag of words and sentiment analysis.

The dataset is sourced from Kaggle and contains about 130,000 wine reviews, with attributes such as country, province, description, rating points, price, title, variety, and winery. A reduced dataset (Wine_Description_3Continents.csv) of 6000 samples with an added 'continent' column is used for the exercise. Additionally, polarity scores(polarity_scores.csv) derived from the NLTK library are provided for sentiment analysis.

7.2.1 Data visualization exercise with sampled dataset using Tableau

Open Tableau and navigate to the Text file option to locate and select the "Wine_Description_3Continents.csv" file. Open the file and proceed to the worksheet by clicking 'Go to Worksheet' at the bottom of the screen.

Upon opening, Tableau automatically generates a Country & Province Hierarchy, Latitude, and Longitude, and classifies measures (Points and Price) and Dimensions. To include Continent in the location hierarchy, drag Continent and drop it within the 'Country, Province hierarchy' (above Country) and rename it to 'Location'.

Change the aggregation of 'Points' and 'Price' to 'Average' instead of 'Sum'. Right-click 'Points' (or 'Price') -> Default properties -> Aggregation -> Average.

Drag 'Location' into Columns and 'Price' and 'Points' into Rows. Click the '+' button next to 'Continent' to expand the bar charts into 'Country' or 'Province' hierarchies. Sort by 'Price' within the location hierarchy 'Country'.

Show the value by selecting 'Label Marks' and choosing 'Show mark labels'. Additionally, make the chart more colourful by dragging 'Price' into Colour Marks.

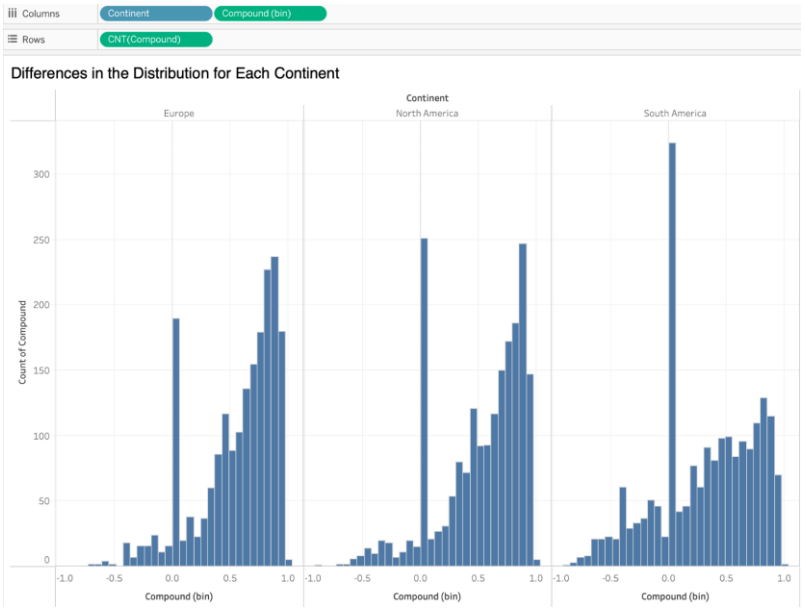


Figure 7.1: Average Prices and Average Rating Scores of Sampled Wines by Continents and Countries

7.2.2 Scatter Plot

1. Create a new worksheet.

2. Place 'Price' into Columns, 'Points' into Rows, and 'Country' into 'Colour'. Assess the correlation between wine price and review points.

3. Add 'Province' just below 'Country' in the Marks to further analyse the relationship between wine price and review points.

4. Right-click the 'Points' Axis, edit axis, and change the scale range of the Y-axis from 80 to 100. Evaluate the correlation again.

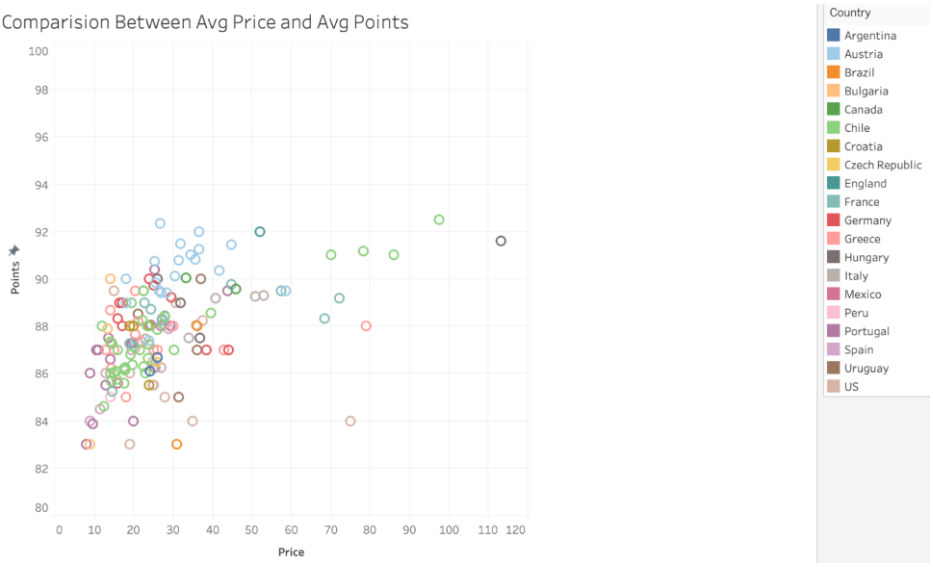


Figure 7.2: Price and Rating Points by Countries and Provinces of Origins

7.2.3 Geo-map chart

1. Create a new worksheet.

2. Drag and drop 'Country' onto the chart to display the map.

3. Place 'Price' over 'Colour' Mark. Click 'Label' to display the average price of each country. Use the '+' button next to 'Country' to examine average prices at the province level.

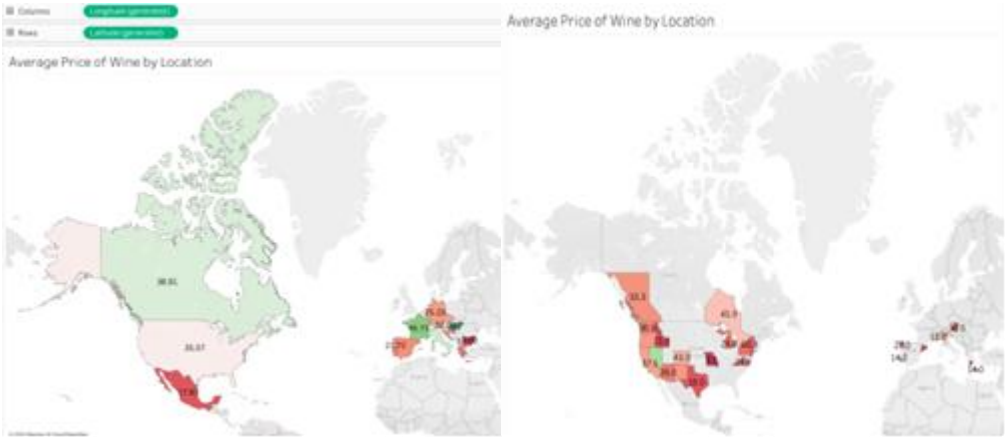


Figure 7.3: Average Prices by Different Countries Producing Wines

7.2.4 Creating an interactive Dashboard.

1. Create a new dashboard by clicking 'New dashboard' at the bottom of the screen.

2. Set the Size scale to 'Automatic'.

3. Arrange the visualizations by dragging and dropping the bar chart (Sheet 1), scatterplot (Sheet 2), and geo map (Sheet 3) onto the dashboard. Position the geo map to the right of the scatterplot and under the bar chart.

4. Customize the titles of the three charts in the dashboard.

5. Utilize the bar chart as a filter by clicking on it, then accessing the 'Funnel' button located on the right side of the chart.

6. Interact with the dashboard by clicking on different areas of the bar chart to observe corresponding changes in the scatter plot and geo map.

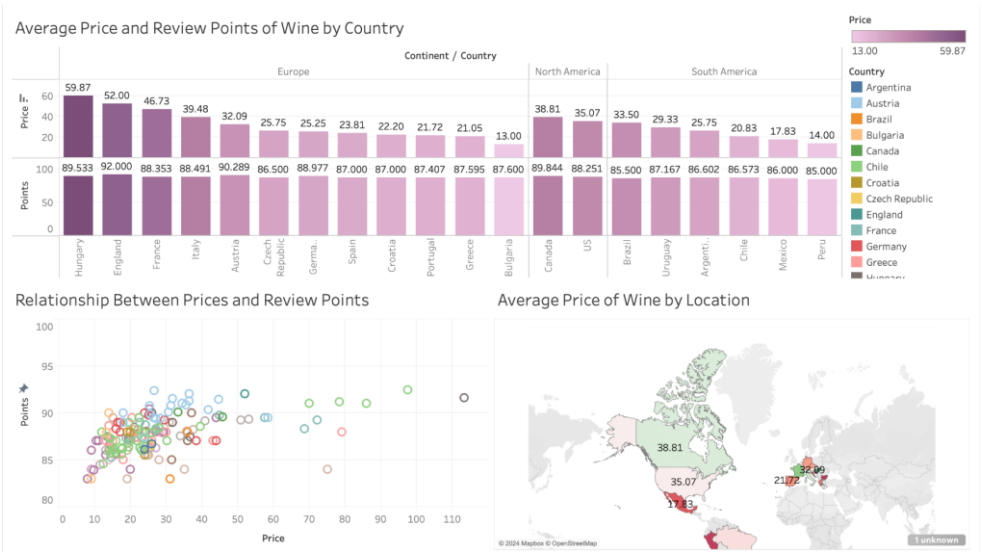


Figure 7.4: Interactive Dashboard

7.3. Data splitting and creating a bag of words with Netlytic

Netlytic is a cloud-based tool for text and social network analysis, capable of summarizing and discovering communication networks from publicly available social media posts.

A "bag of words" approach involves converting unstructured reviews into unique word values, facilitating word frequency analysis. This method helps understand common words in reviews by continent, aiding in understanding customer behavior.

7.3.1 To split sampled data by continent:

1. Open "Wine_Description_3Continents.csv" in MS Excel and save it as a separate file (e.g., Wine_Description_Temp.csv).
2. Retain only the "Continent" and "Description" columns.
3. Filter the Continent column for 'Europe'.
4. Copy the filtered Description column into a new CSV file named 'EU_Descriptions'.
5. Repeat steps 3 and 4 for 'North America' and 'South America', saving them as 'NA_Descriptions' and 'SA_Descriptions' respectively.
6. Each resulting file contains 2000 wine review descriptions specific to its continent.

7.3.2 To run Netlytic and create a bag of words

1. Register an account on Netlytic.

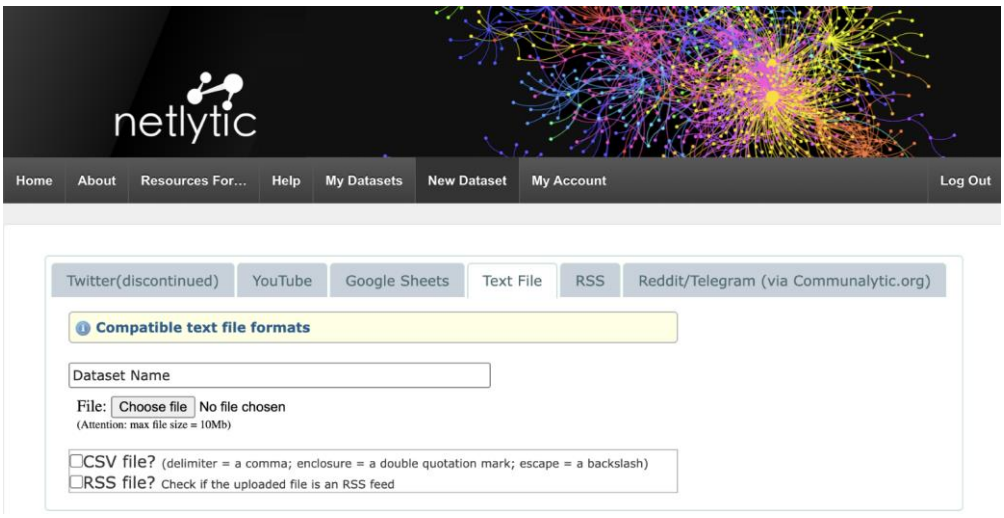


Figure 7.5: Uploading a .csv File in Netlytic

2. Navigate to the "New Dataset" tab.

3. Choose "Text File", provide a name, and upload one of the three .csv files created earlier.

4. If the file isn't recognized as a CSV, ensure to mark the required checkbox.

5. After uploading, click "Import", then proceed to the next step.

6. Preview the data on the Preview tab, then navigate to the "Text Analysis" tab.

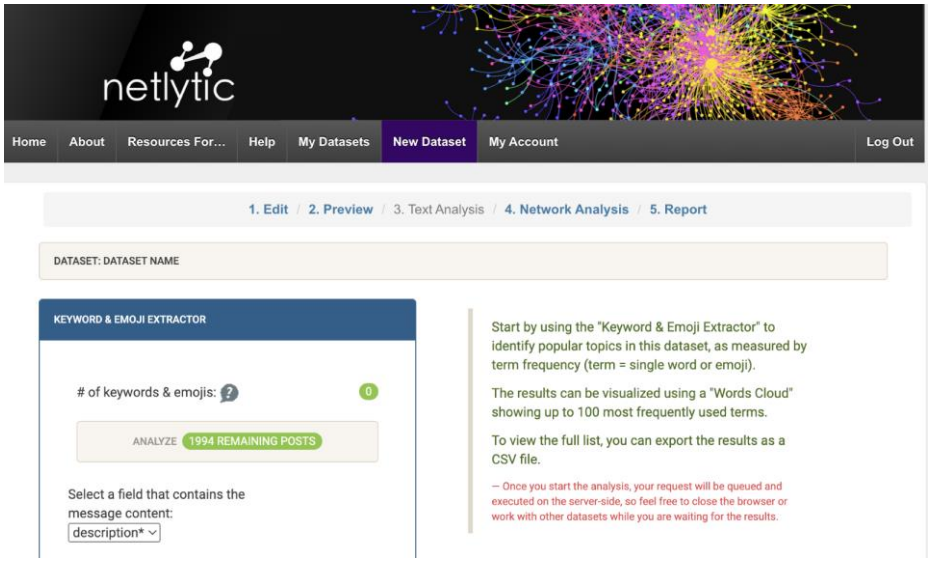


Figure 7.6: Analysing a .csv File in Netlytic

7. Click on the "Analyse" text to initiate processing, which will change the file status to "Queued".

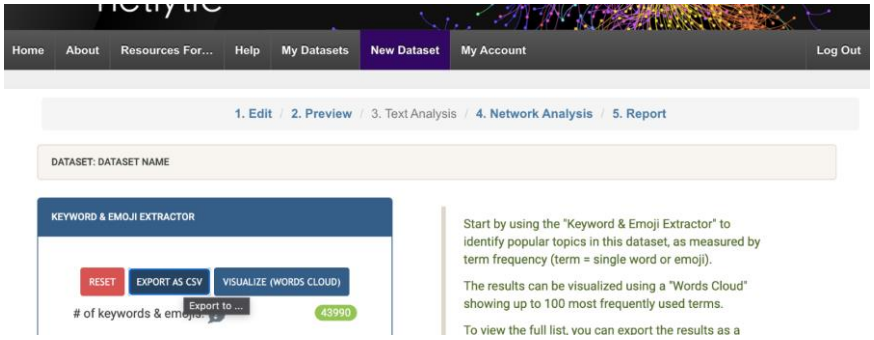


Figure 7.7: Exporting a .csv File in Netlytic

7.3.3 To combine three "Bags of Words" files into one

1. Open the newly created files containing three columns: "term" for words found in reviews, "#messages" for the number of reviews the word appeared in, and "#instances" for the total occurrences of the word across all reviews.

2. In each file, add a new column labelled "continent" and populate it with the respective continent values (Europe, North America, and South America) across all rows.

	A	B	C	D
1	term	#messages	#instances	continent
2	wine	1083	1431	Europe
3	fruit	731	821	Europe
4	palate	726	749	Europe
5	acidity	682	703	Europe
6	aromas	677	685	Europe
7	Drink	584	592	Europe

Figure 7.8: Changes done to the Exported Netlytic Files

3. Combine all three datasets into one file named "Netlytics_Wine_BagOfWords_Total.csv".

4. The resulting file should contain four columns: "term", "#messages", "#instances", and "continents". The "continents" column should have three repeated unique values for Europe, North America, and South America.

8. Netlytic may take time to process. To expedite, open new tabs and repeat the process for the remaining .csv files.

9. Keep reloading the page until the keyword extractor panel changes. Click "export as csv" to obtain the file.

7.4 Research Questions to be Solved.

Q1: What are the most frequent words used in review descriptions of wine production in different countries?

Q2: What are the differences in distributions for each continent?

Q3: Evaluate the Sentiment Distribution for this dataset.

Q4: Show the top 10 positive reviews for all continents.

Q5: Show the top 10 negative reviews for all continents.

7.5 Outcome Analysis of the Research Questions using Tableau.

Q1: What are the most frequent words used in review descriptions of wine production in different countries?

To draw tag-cloud charts comparing the top 100 (or 200) most recurrent terms in wine reviews of three continents using the Netlytic data ("Netlytics_Wine_BagOfWords_Total.csv"):

Open Tableau, navigate to Text, then click Next. Find and select the "Netlytics_Wine_BagOfWords_Total.csv" file and open it to create a new worksheet --> Drag 'Term' into Rows. A pop-up will appear prompting you to use a filter.

Select 'Filter and then add', then click the 'Top' tab. Choose 'By field' and set it to 'Top 150 by #instances (sum)'.

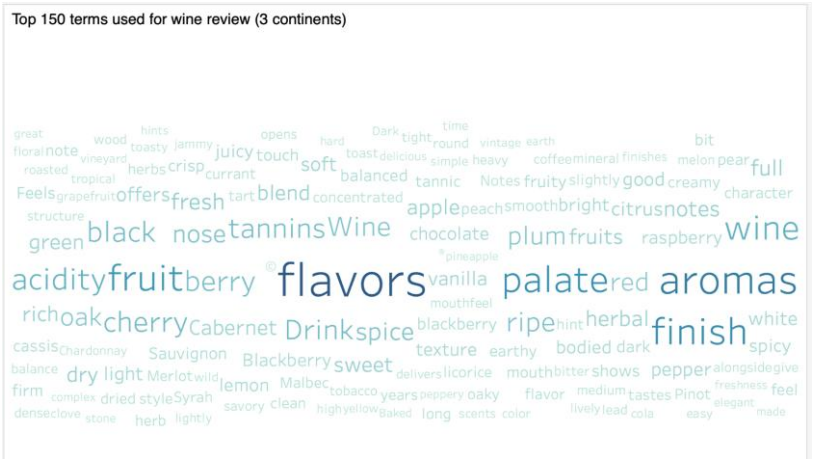


Figure 7.9: Top 150 Terms Used for Wine Review for Europe, North America, South America

Place '#instances' over the 'Size' Mark and '#messages' over the 'Color' Mark. Use 'Show Me' on the top-right side and select 'Tree Map'. Change the Checkbox of the Marks from 'Automatic' to 'Text'.

Adjust the title to "Top 150 terms used for wine review (3 continents)" and Create three more tag-cloud charts for each continent by duplicating the worksheet and using a filter for 'Continent'.

Drag and drop 'Continent' to 'Filter' and select 'Europe', 'North America', and 'South America'. Change the title accordingly for each chart.

Figure 7.10: Top 150 Terms Used for Wine Review for Europe

Once all four charts are completed, create a dashboard to display them.

Figure 7.11: Top 150 Terms Used for Wine Review – A Dashboard

Q2: What are the differences in distributions for each continent?

Sentiment analysis is conducted using Tableau, initially focusing on understanding the score distribution across each continent:

Open Tableau and upload the "polarity_scores.csv" file (Click 'Text File' → Choose 'polarity_scores.csv'). Then, open a new sheet to commence the analysis.

Drag the 'Compound' field and drop it into Rows. Next, navigate to and select 'Histogram'. This provides the score distribution for all continents.

To obtain a histogram for each continent, drag and drop 'Continent' into the Columns shelf in front of 'Compound (bin)'.

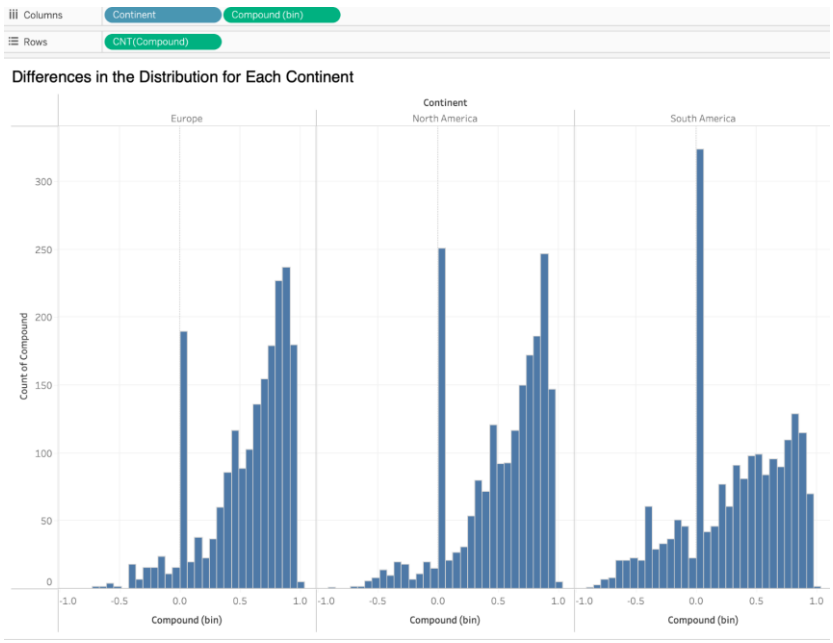


Figure 7.12: Top 150 Terms Used for Wine Review – A Dashboard

Q3: Evaluate the Sentiment Distribution for this dataset.

To analyze review sentiment based on compound scores:

1. Click the ‘New worksheet’ icon at the bottom to create another worksheet. Then, create a calculated field named “Sentiment” using the following conditional statements:

IF [compound] > 0.2 then 'Positive'

ELSEIF [compound] < 0.2 and [compound] > -0.2 THEN 'Neutral'

ELSE 'Negative'

END

2. Create a stacked bar chart to evaluate sentiment distribution. Drag 'Continent' into Columns and 'Compound' into Rows. Next, drag 'Sentiment' dimension over the Color Marks. Change the measure 'Sum(Compound)' to 'Count' to count the number of positive, negative, and neutral reviews. Adjust the graph size from 'Standard' to 'Fit width' to widen the chart.

Figure 7.13: Top 150 Terms Used for Wine Review – A Dashboard

Q4: Show the top 10 positive reviews for all continents.

To analyze the top 10 positive reviews using a bubble chart:

1. Click the ‘New worksheet’ icon at the bottom to create another worksheet. Place the Compound measure under Rows and the Headline dimension under Columns. A ‘Warning’ dialog box may appear, recommending you add a filter by default. Click the button to proceed.

2. Place the Headline dimension under Filters (this step might not be needed if you were recommended to use a filter) and click on the Top tab. Choose the “By field” option and select "Top 10 by Compound Sum" to filter for the top 10 positive reviews.

3. Once the filter has been applied, select the bubble chart to visualize the top 10 positive reviews.

Figure 7.14: Top 10 Positive Reviews

Q5: Show the top 10 negative reviews for all continents.

To analyze the top 10 negative reviews using a bubble chart:

1. Click the ‘Duplicate option at the bottom to create another worksheet that contains the exact copy of the positive bubble chart.

2. Click on the Top tab. Choose the “By field” option and select "Bottom 10 by Compound Sum" to filter for the top 10 negative reviews.

3. Once the filter has been applied, select the bubble chart to visualize the top 10 negative reviews.

Figure 7.15: Top 10 Negative Reviews

7.6 Conclusion

Upon completion of this text analytics exercise, facilitated by Tableau and supplementary tools, participants acquire a nuanced understanding of unravelling insights from unstructured text data, particularly in the realm of wine reviews. Utilizing sentiment analysis techniques, individuals discern sentiment distributions across diverse continents, illuminating varying consumer perceptions and attitudes. Through the creation of tag-cloud charts, participants unveil recurrent terms within reviews, unveiling prevailing consumer preferences and emerging trends, thereby empowering informed decision-making and strategic planning.

Moreover, this exercise underscores the versatility of Tableau in transforming raw data into actionable insights, manifesting in the creation of diverse visualizations such as histograms and stacked bar charts. Such proficiency in data visualization equips individuals with the tools to effectively communicate findings and drive informed decisions, thereby fostering a culture of data-driven decision-making within organizations.

Work Citations:

1. IBM. (n.d.). Data modeling. Retrieved from https://www.ibm.com/topics/data-modeling#:~:text=Data%20modeling%20is%20the%20process,between%20data%20points%20and%20structures.

2. Baumeister, A., Harrer, C., & Sträßer, U. (2011). ERPsim – A Simulation Game for Teaching SAP ERP. International Institute of Informatics and Systemics. https://www.iiis.org/CDs2011/CD2011SCI/EISTA_2011/PapersPdf/EA259ET.pdf

3. Léger, P.-M., et al. (2009). HEC Montréal ERP Simulation Game, Manufacturing Game, Participants Guide. Pearson Education.

4. Sharma, K. (2023, August 22). 26 Tableau Features to Know from A to Z. Tableau. https://www.tableau.com/blog/26-tableau-features-know-a-to-z

5. Intellipaat. (n.d.). What is Tableau? https://intellipaat.com/blog/what-is-tableau/

6. Gupta, A. (2023, February 12). An Introduction to Pivot Tables in Excel. https://www.simplilearn.com/tutorials/excel-tutorial/pivot-table

7. Simplilearn. (2023, June 6). What is Data Wrangling? Benefits, Tools, Examples and Skills. https://www.simplilearn.com/data-wrangling-article

8. GeeksforGeeks. (n.d.). K-means clustering - Introduction. https://www.geeksforgeeks.org/k-means-clustering-introduction/

9. Wikipedia contributors. (2024, February 21). K-means clustering. In Wikipedia. https://en.wikipedia.org/wiki/K-means_clustering

10. Amazon Web Services. (n.d.). What is Text analysis? Retrieved from

https://aws.amazon.com/what-is/text-analysis/

11. MonkeyLearn. (n.d.). What is Text analysis. A Beginner's Guide. Retrieved from https://monkeylearn.com/text-analysis/

12. Toward Data Science. (2019, July 19). Forecast KPI: RMSE, MAE, MAPE, Bias. Retrieved from https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d#:~:text=The%20Mean%20Absolute%20Percentage%20Error,average%20of%20the%20percentage%20errors.