Final Portfolio

Chapters 1,2,3,4,5,6 & 7

Razia.S.Choudhury- B00868310

CONTENTS

CHAPTER 1 SAP LUMIRA DISCOVERY

CHAPTER 2 MICROSOFT EXCEL PIVOT TABLE

CHAPTER 3 SAP PREDICTIVE ANALYTICS FOR VISUALIZATION

CHAPTER 4 TABLEAU

CHAPTER 5 SAP ANALYTICS CLOUD

CHAPTER 6 SAP PREDICTIVE ANALYTICS FOR DATA MINING

CHAPTER 7 SAP HANA

CHAPTER 1 : SAP LUMIRA DISCOVERY

SECTION 1.1 INTRODUCTION

SAP Lumira Discovery can be described as a business technology platform designed to accommodate the following aspects of Self-Service BI (BI, 2017)

- Data Acquisition - Importing data from disparate sources including SAP HANA, SAP BW, MS Excel, Cloud applications, Salesforce, etc.
- Data Wrangling/Preparation – Includes exploring, cleaning, and harnessing data in real-time; modeling data — custom calculations, groupings, formatting values, etc — with the goal of uncovering valuable insights and support decision-making. The tool also supports integrating and appending datasets from multiple sources.
- Data Visualization – Create data stories with BI-enforced visualisations — storyboards and dashboards; drag-drop feature; several chart types; maps-enabled geo charts; pictograms and shapes; infographics.
- Data Collaboration — Share the analysis reports by exporting datasets, crosstabs, visualizations, excel and pdf files on the local or dedicated servers.

SECTION 1.2 DATASET DESCRIPTION

SAP Lumira Discovery will be used to analyze two datasets maintained by the Government of Canada (Statistics Canada, 2015). These datasets were provided in class as Excel spreadsheets (.csv files)

Population.csv – Contains information about the annual estimates of population by age and sex for Canada, provinces, and territories. It includes 16 columns of data attributes out of which only 7 columns (**Ref Date, GEO, SEX, AGE Group, Vector, Coordinate, Value)** would be utilized in the tool for advanced data processing, analyzing and visualizing.

Sales.csv — Contains information about the value and volume of sales of different alcoholic beverages collected from several provincial and territorial liquor authorities also retail outlets across Canada. It includes 17 columns of examinable data values like date, sales value, sales volume, type of beverages, origin of product, GEO. No data cleaning was required for this dataset.

SECTION 3: BUSINESS PROBLEM AND RESEARCH QUESTIONS

3.1 – Business Use Case

The alcohol consumption in Canada might vary across its several provinces due to several reasons. The sales of alcohol in Canada can be related to the population density of these provinces. Also, there might be many examinable factors effecting the popularity of a specific alcohol beverage in Canada. Therefore, to speculate and visually analyse the alcohol sales and customer behaviour, the below research questions have been formed.

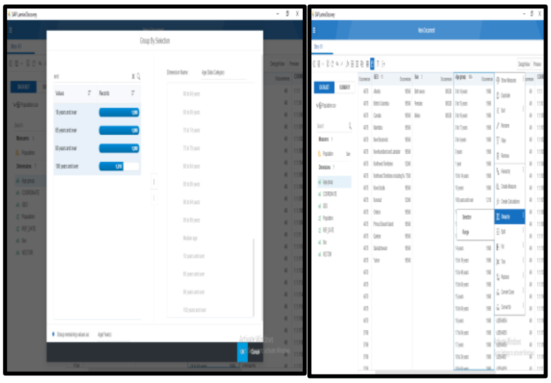
3.2 – Research questions

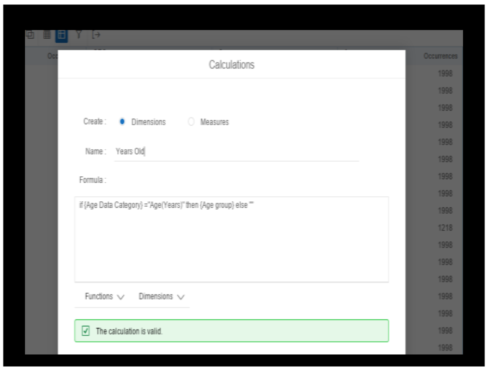
1. What is the number of adults who are legally allowed to purchase alcohol beverages in each province of Canada?
2. How did the per capita sales for beers and wines varied by province for the years 2005 and 2018?
3. What was the average cost per litre of beers and wines across the different provinces of Canada for the years 2005 and 2018?
4. What is the trend for the total volume sales over the years?
5. How did the alcohol purchasing preferences for beers and wines varied for Nova Scotia, for over several years? Do adults buy alcohol based on pricing or preference in Nova Scotia?

SECTION 1.4 ANALYTICAL PROCESS AND KEY FINDINGS

1.4.1 What is the number of adults who are legally allowed to purchase alcohol beverages across Canada?

For this question, the dataset — Population.csv — was loaded and visualized using the tool. In the data view, it was observed that the dataset had a dimension called Age group which had mixed data values (ranges of ages and single years) and lacked consistency. To clearly identify the adults in the different provinces of Canada, a new dimension — Age Data Category — was created using grouping from another dimension, 'Age group'(as shown in Fig 1.4.1(a)) which had two values —Age Range and Age (Years). Furthermore, a new dimension called Years Old was created to populate the age of adults whose Age is in years using the formula - ***if {Age Data Category} ="Age (Years)" then {Age group} else ""***. (Fig 1.4.1(b))





Calculations

Create : ● Dimensions ○ Measures

Name : Years Old

Formula :

```
if {Age Data Category} ="Age(Years)" then {Age group} else ""
```

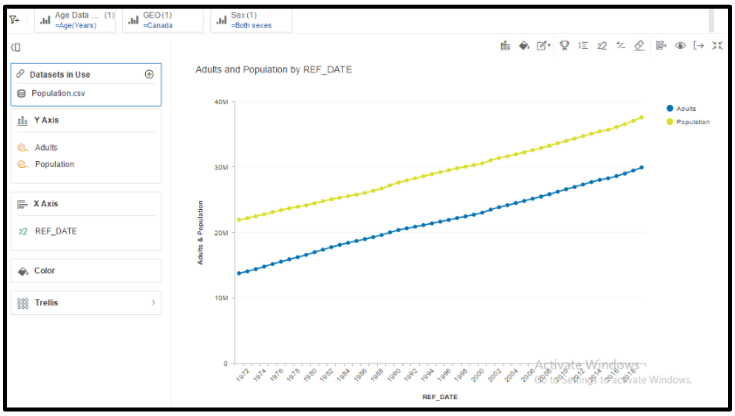
☑ The calculation is valid.

Fig 1.4.1(a): Grouping data; Create Dimension Age Data Category

Fig1.4.1(b): Creating the Calculated Dimension- Years Old

Given the minimum age limit for alcohol purchasers is 19 years across Canada except Quebec, Alberta and Manitoba where the legal age is 18 years. The date was further wrangled in the data view to accommodate a new dimension called "Adults" created by applying the above-mentioned constraints. A new measure was also created from the dimension Adults. To illustrate this, a line chart was used in the design view. The measures Adults and Population was dropped on the Y-AXIS and the dimension REF-DATE on the X-AXIS. Filters GEO=Canada, Age Data Category=Age (Years) and SEX=Both sexes are applied to determine the number of adults who are legally allowed to purchase alcohol beverages across Canada as seen in Fig 1.4.1(c). To explore the data sharing functionality of SAP Lumira, a 'Cross Tab' (Chart type table) was created with Adults (from 'Measures') in the Measure, Ref_Date in Columns and GEO in Rows. Applied filters were Age Data Category=Age (Years), SEX=Both sexes and GEO

Exclude "Northwest Territories including Nunavut". This was exported as an Excel file (Adults.xlsx) on the local computer using the Export function.



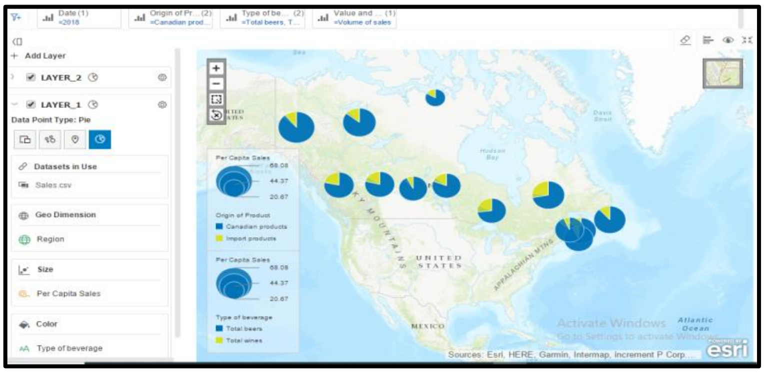
Age Data ... (1) =Age(Years)
GEO (1) =Canada
Sex (1) =Both sexes

Fig 1.4.1(c) Canadian population and number of adults

Key Findings*:* As the population is increasing with time, alcohol revenue is likely to increase in Canada considering a greater number of adults are legally permitted to purchase alcohol beverages across the country.

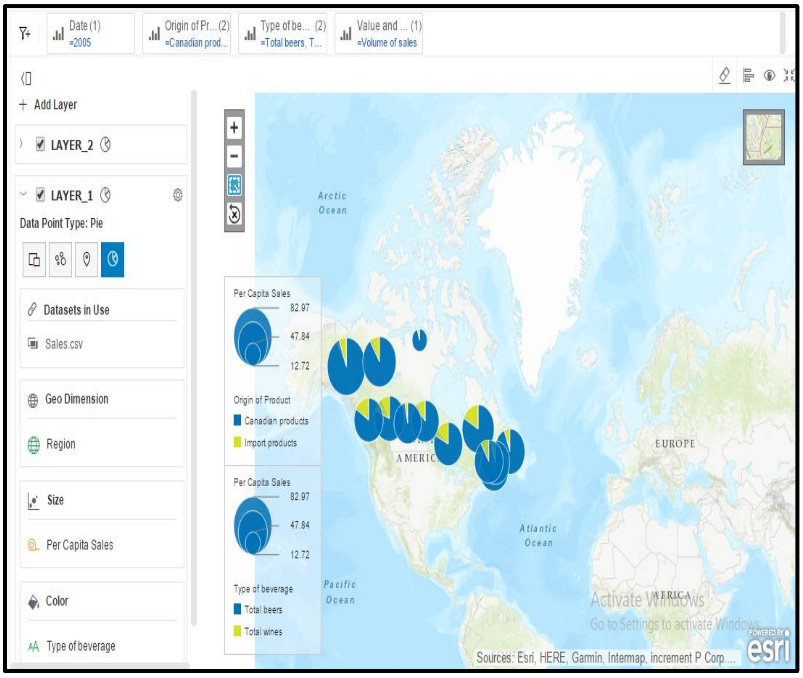
1.4.2 How did the per capita sales for beers and wines vary by province for the years 2005 and 2018?

As seen, alcohol consumption in Canada changes with respect to different provinces. It is important to analyse how much volume of alcohol can be purchased by each adult across Canada. For this question, the datasets — Population.csv and Adults.csv (generated by SAP Lumira – 1.4.1 for reference) — were loaded and merged using the common attributes GEO and Date. In the data view, a new calculated field called '*Per Capita Sales'* (volume of alcohol per adult) was created from the measure *VALUE* using formula **{VALUE}/{Adults}*1000.** To visualize the variance of provincial per capita sales in 2018 against the trend in the year 2005, GEO map chart feature of the tool was used. The map chart was created having two layers **—** Layer 1: Data point type: Pie / Per Capita Sales as the Size / Region as Geo Dimension / Type of beverage as color. Layer 2: Data point type: Pie / Per Capita Sales as the Size / Region as Geo Dimension / Origin of Product as color. The filters Type of beverage = Total Beer and Total Wine, Date = 2018, Origin of Product = Canadian and Import, and Value and Volume = Volume of Sales. Fig 1.4.2(a) shows the illustration for the same. The Date filter value was changed to 2005 which generated the geo chart as shown in Fig 1.4.2(b) .



Date (1) =2018
Origin of Pr... (2) =Canadian prod...
Type of be... (2) =Total beers, T...
Value and ... (1) =Volume of sales

Fig 1.4.2(a). Provincial per capita sales in 2018



Key Findings: The litres of alcohol consumed per adult for the alcoholic beverages — beers and wines — was higher in 2005 as compared to the year 2018. One of the reasons for this could be that these beverages were sold at a much lower price in 2005 than in 2018. Therefore, it is important to find the average cost per litre of these beverages to support this analysis.

Fig 1.4.2(b): Provincial per capita sales in 2005

1.4.3 What was the average cost of beers and wines across the different provinces of Canada for the years 2005 and 2018?

Knowing the per capita sales of the different provinces of Canada, it is important to measure the average cost of alcohol per litre to understand the purchasing characteristics of adults in Canada. We know the per capita sales of beers in 2005 is much higher than 2018. One of the reasons could be that beer was much cheaper in 2005 than it is in 2018. To visualize this, two separate dimensions — Sales Value and Sales Volume were created from the Value and volume column in data view of SAP Lumira using the formulas — ***Sales Value = if {Value and volume} ="Volume of sales" then {VALUE} else "", Sales Volume = if {Value and volume} ="Value of sales" then {VALUE} else "".*** These dimensions were converted to measures (sum of all the values) from which a calculated field called – **Average Cost** was created using the formula – *Sales Value/Sales Volume*. In the design view, the measure Average Cost was dropped on the Y-AXIS and the dimension Date was dropped on the X-AXIS. Filters GEO= Alberta, Nova Scotia, Ontario etc. (Selected all provinces and not territories), Date = 2005, 2018, Type of beverage = Total beers, total wines, and Origin of Product = Canadian products, import products were applied to illustrate the average cost of beers and wines across the different provinces of Canada for the years 2005 and 2018 as seen in Fig1.4.3(a).

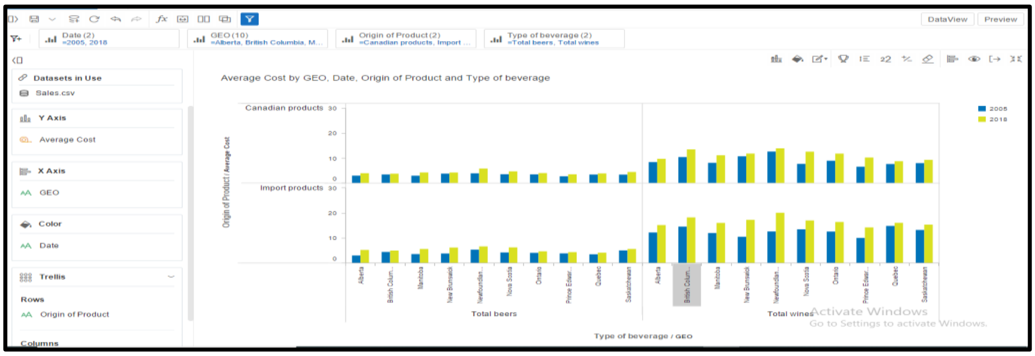


Fig1.4.3(a) : Average cost of beers and wines; Canadian vs imported;2005 vs 2018

***Key Findings**:* The average cost of beers and wines are higher in 2018 when compared to those in 2005 for both Canadian and Imported products. This trend is observed for all the provinces but there was not much difference in the yearly costs of beers and wines for Quebec. Hence, it can be concluded that one of the reasons for the gradual decline in the per capita sales of beers and wines in 2018 was due to the increase in product price.

1.4.4 What is the trend for the total volume sales over the years?

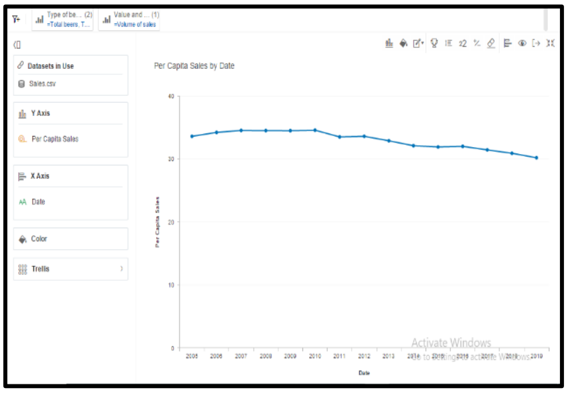


Fig 1.4.4(a) Per capita sales vs years

It is important to analyse if the total volume sales for beers and wines was low in 2018 because as seen in Fig 1.4.4(a), the per capita sales was low for that year. To illustrate this, a funnel chart type was used. Sales Volume was added to Values and Date was dropped in the Color section. The type of beverage was filtered to only include Total Beers and Total Wines. This resulted in a funnel shaped visualization as shown in Fig 1.4.4(b). with the total sales volume arranged in descending order with the respective years.

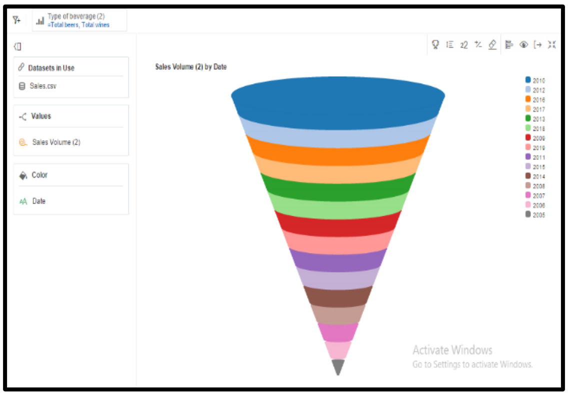


Fig 1.4.4(b) : Total Volume of Sales vs Date

Key Findings: Canadians bought higher volumes of beers and wines in the year 2018 than in 2005.Although the per capita sales were lower in 2018, but it did not affect the total sales as there was a greater number of adults in Canada at that time in comparison to 2005 (reference 4.1.1) However, the trend shows the highest sales happened in 2010 which was when the per capita sales of beers and wines was maximum (as seen in Fig 2).

<u>SUMMARIZED FINDINGS</u>

- The population of adults has increased from 2005 to 2018 as seen in question 1.4.1
- The per capita sales of beers and wines has reduced from 2005 to 2018 as seen in question 1.4.2.
- The average cost of beers and wines has increased from 2005 to 2018 as seen in 1.4.3
- The total volume of sales for all beers and wines is higher in 2018 in comparison to 2005 as seen in 1.4.4.

Total volume of sales = number of adults * per capita

Over the years, the number of adults has increased in Canada which resulted in the increase of the total volume of sales despite reduction in per capita sales. The reduction in per capita sales could be due to many reasons but one analysed finding is due to the increase in average product price.

1.4.5 How did the alcohol purchasing preferences for beers and wines vary for Nova Scotia, over the last several years? Do adults buy alcohol based on pricing or preference in Nova Scotia?

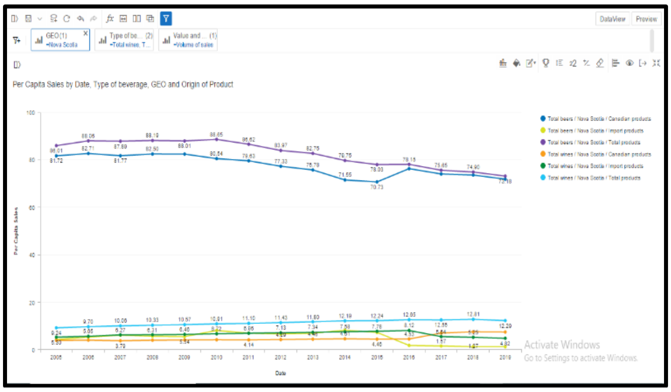

DataView
Preview
GEO (1) =Nova Scotia
Type of be... (2) =Total wines, T...
Value and ... (1) =Volume of sales
Per Capita Sales by Date, Type of beverage, GEO and Origin of Product
Per Capita Sales
Date
100
80
60
40
20
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Total beers / Nova Scotia / Canadian products
Total beers / Nova Scotia / Import products
Total beers / Nova Scotia / Total products
Total wines / Nova Scotia / Canadian products
Total wines / Nova Scotia / Import products
Total wines / Nova Scotia / Total products
86.01
88.06
87.69
88.19
88.01
88.65
86.62
83.97
82.75
79.75
78.03
78.15
75.65
74.90
81.72
82.71
81.77
82.50
80.54
79.63
77.33
75.78
71.55
70.73
73.18
9.24
9.70
10.06
10.33
10.57
10.81
11.10
11.43
11.80
12.19
12.24
12.65
12.81
12.29
3.79
4.14
4.46
Activate Windows
Go to Settings to activate Windows.

Fig 1.4.5(a): **Sales per Capita Nova Scotia; Beers vs Wines; Canadian vs**

To examine the sales trend for beers and wines in Nova Scotia, a line chart was selected in the design view as seen in Fig 1.4.5(a). The measure 'per capita sales' was added to Y-axis, date to X-axis, Type of Beverage, GEO and Origin of Product in Color. The applied filters were GE0 = Nova Scotia, Type pf Beverage = Total Beers and Total Wines, Value, and volume = Volume of sales. It was found that Nova Scotians from 2015 mostly preferred buying local beers and wines. On the other hand, during the same time the sales for imported beers and wines declined. To determine if price acted as a factor to make such preferences, line chart illustration was used having average cost added to the Y-axis, Date to X-axis, Type of Beverage to Color and Origin of product in the column section of Trellis as seen in Fig 1.4.5(b)

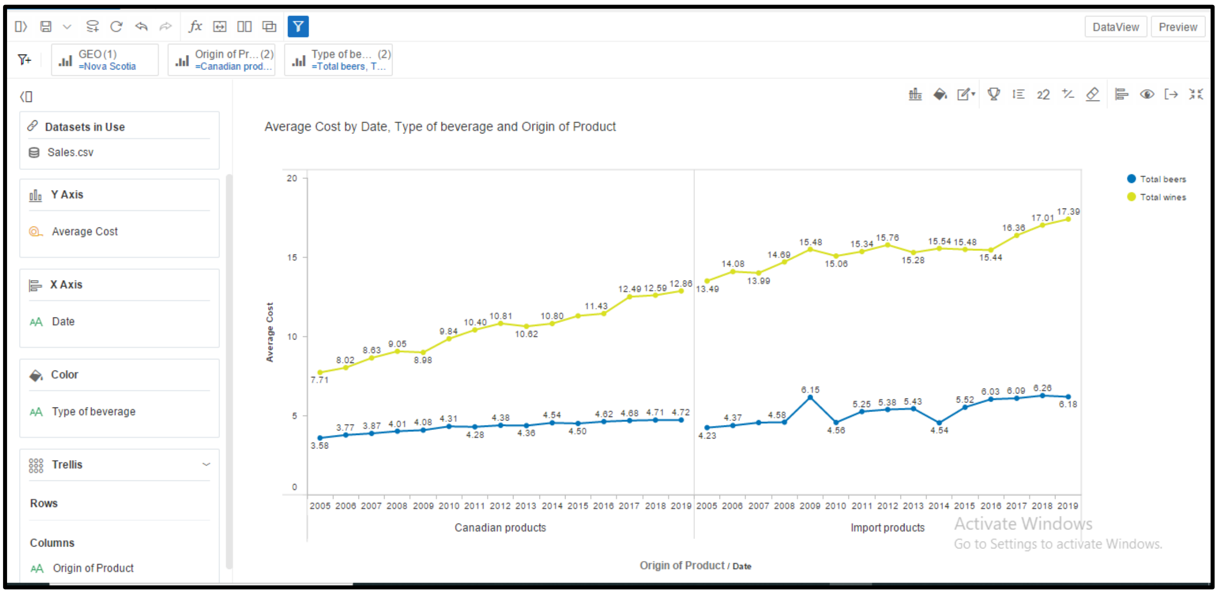


Fig 1.4.5(b) Average cost Nova Scotia; beers vs wines; Canadian vs Imported

Key Findings: From 2015, the average price of imported beers and wines kept rising and thus could have been the reason for decline in respective sales. On the other hand, from 2016, the average prices of Canadian beers and wines started rising slightly. But the increase in pricing did not affect the alcohol preference of Nova Scotians adults.

SECTION 5: ANALYSIS AND CRITIQUE OF THE TOOL

As a first-time user of SAP Lumira Discovery, I did not find the working and functionality of this tool hard to follow. The tool facilitates advanced data discovery and analysis which made it easier to implement data manipulation techniques (grouping, formulas, cleaning, etc) smoothly. I found the tool user-friendly in developing illustrations of several types also enjoyed creating interactive maps with the double-layered features. The tool can be used to Integrate on-premise data discovery, dashboards, and analytic applications with business user access to all analytics across the organization using SAP Analytics Hub (SAP LUMIRA, n.d.). The drag and drop feature of the tool enabled creation of BI charts with ease. The tool is equipped with effective sharing options which can help deliver impactful BI solutions across the enterprise or with the other products of SAP (BI, 2017).

Although, I could explore most of the chart styles for the visualization process, I am looking forward to creating dashboards using scatter/heat/tree/KPI illustrations in the future. I am also looking forward to exploring SAP Lumira Discovery's interoperability with SAP Lumira Designer and live data connectivity with SAP HANA.

One of the challenging parts of this tool, is that the values added to any dimensions or measures are case-sensitive. So, we need to input the correct case in the formula fields otherwise it produces invalid results. Being a data analyst, I have worked with several visualization tools in the past like Power BI, Tableau, Spotfire, etc and therefore have certain expectations form on-premise, self-service BI tools. The speed at which SAP Lumira performs in real-time is quite disappointing. Many times, the interface got freeze and slow. When I would change the filters of the BI charts, the new changes would not be reflected unless I changed the chart style. Another disappointment was when the geolocation maps could not be created in an online mode. I had to change the Network and proxy settings to create the geo maps in offline mode. Overall, SAP Lumira is an efficient analytics tool, but the functional processes are not as smooth as one would expect.

SECTION 6: CONCLUSION

SAP Lumira Discovery is no doubt efficient in merging, manipulating, analysing, and visualizing the underlying datasets. However, due to its slowness and high memory consumption, it cannot be considered as one of the most powerful BI tools. The real-time performance and live location features can be improved to make this tool more user-friendly. Moreover, considering the launch and productivity of several Cloud-based SAP products, this tool can also be migrated to the server to avoid excessive memory consumption issue and the extra effort of downloading this software.

CHAPTER 2 MICROSOFT EXCEL PIVOT TABLE

SECTION 1: INTRODUCTION TO THE TOOL

One of the universally adopted BI tool is Microsoft Excel due to its simplicity and customization benefits. Pivot tables are the most powerful and useful features in Excel. They provide an interactive view of the data in the form of cross-tabulated structure called crosstab which allows grouping of data into categories, filtering to include and exclude data columns, sorting, formatting, performing data calculations and build effective visualizations using Pivot Charts (Fransen, 2021).

SECTION 2: DATASET DESCRIPTION

Excel pivot tables will be used to analyse and visualize the sales data for the years 2007-2011 of an organization called GBI. The dataset was provided in class as an Excel Workbook (.xlsx) and include 18 columns of data attributes – calendar year, month, material, country, revenue, net sales, customer, sales quantity, etc — which can be examined, evaluated, and manipulated to support decision-making for GBI.

SECTION 3: BUSINESS PROBLEM AND RESEARCH QUESTIONS

3.1 – Business Use Case

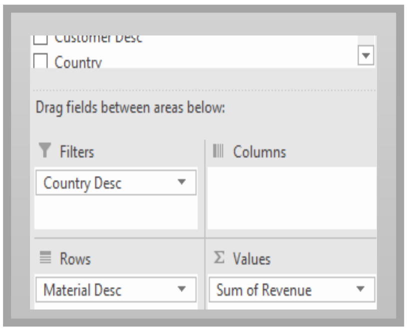
Profitability Analysis helps organizations to assess the growth of their business, keep track of business performance, and identify the areas where profits can be maximised. This business scenario is focussed on the profitability analysis of a company called Global Bike Inc. Therefore, to speculate and visually analyse the company's profits for different products, customers, regions, and years the following research questions have been formed.

3.2 – Research questions

1. What is the trend of overall revenue generation by different materials in USA and Germany? Did the same materials generate highest and lowest revenue in both countries?
2. Which customer has the highest percentage contribution to the total revenue? What has been the trend of that customer's percentage contribution over the years?
3. Are US and Germany buying and selling products at the same price?
4. Which year had the highest gross margin? How does the profitability differ between US and Germany for the year with the highest gross margin?
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?

SECTION 2.4 ANALYTICAL PROCESS AND KEY FINDINGS

2.4.1 What is the trend of overall revenue generation by different materials in USA? Is the trend same for Germany? Did the same materials generate highest and lowest revenue in both countries?



Customer Desc
□ Country

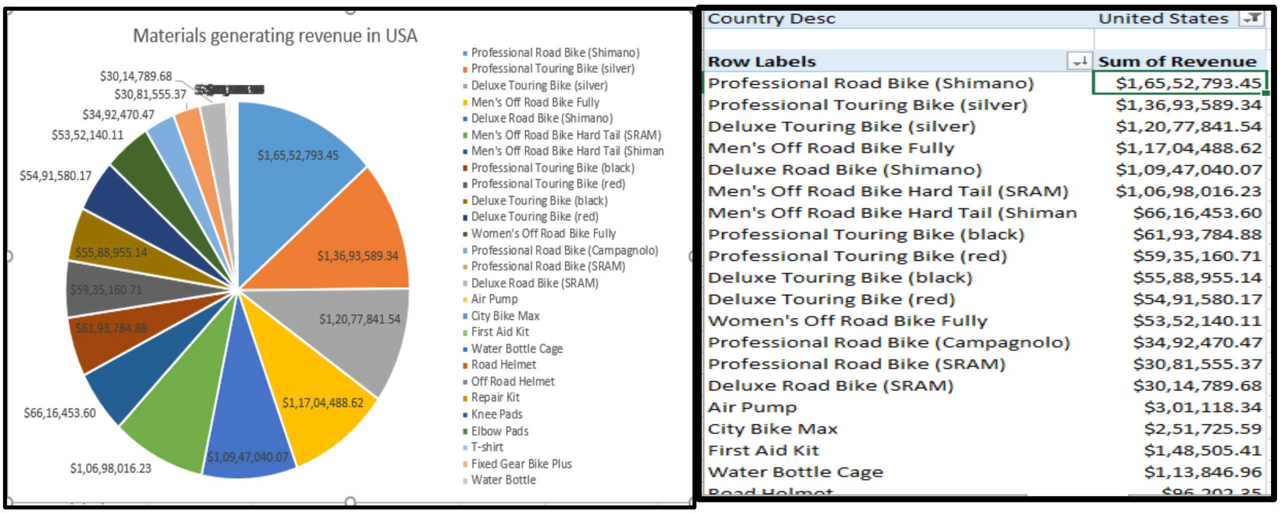
Drag fields between areas below:

Filters	Columns
Country Desc	

Rows	Values
Material Desc	Sum of Revenue

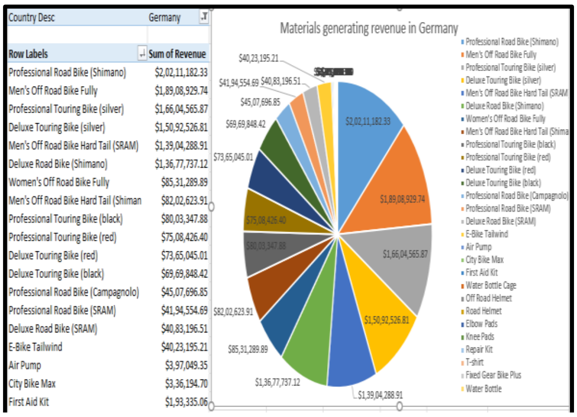
Fig 2.4.1(a) : Pivot table builder

A pivot table for the entire dataset was inserted in a new worksheet. As seen in Fig 2.4.1(a), in the PivotTable Builder, *Material Desc* was added to Rows, Revenue to Values and Country Desc to Filters (to filter out the country -USA first). This resulted in a crosstab as shown in Fig 2. The Rows were first sorted by descending sum of revenues to determine the materials contributing towards the highest of GBI. To illustrate this, a pie chart was created from the table as shown in Fig 2.4.1(b).



Country Desc	United States
Row Labels	**Sum of Revenue**
Professional Road Bike (Shimano)	$1,65,52,793.45
Professional Touring Bike (silver)	$1,36,93,589.34
Deluxe Touring Bike (silver)	$1,20,77,841.54
Men's Off Road Bike Fully	$1,17,04,488.62
Deluxe Road Bike (Shimano)	$1,09,47,040.07
Men's Off Road Bike Hard Tail (SRAM)	$1,06,98,016.23
Men's Off Road Bike Hard Tail (Shiman	$66,16,453.60
Professional Touring Bike (black)	$61,93,784.88
Professional Touring Bike (red)	$59,35,160.71
Deluxe Touring Bike (black)	$55,88,955.14
Deluxe Touring Bike (red)	$54,91,580.17
Women's Off Road Bike Fully	$53,52,140.11
Professional Road Bike (Campagnolo)	$34,92,470.47
Professional Road Bike (SRAM)	$30,81,555.37
Deluxe Road Bike (SRAM)	$30,14,789.68
Air Pump	$3,01,118.34
City Bike Max	$2,51,725.59
First Aid Kit	$1,48,505.41
Water Bottle Cage	$1,13,846.96
Road Helmet	$96,202.35

Fig 2.4.1(b): Pivot table and pie chart showing revenues generated by GBI materials in USA

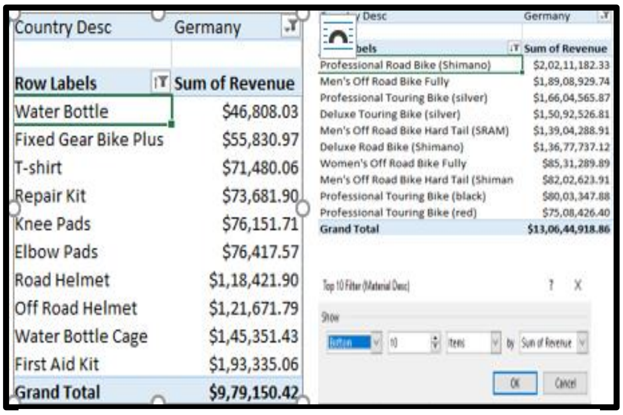


Country Desc	Germany
Row Labels	**Sum of Revenue**
Professional Road Bike (Shimano)	$2,02,11,182.33
Men's Off Road Bike Fully	$1,89,08,929.74
Professional Touring Bike (silver)	$1,66,04,565.87
Deluxe Touring Bike (silver)	$1,50,92,526.81
Men's Off Road Bike Hard Tail (SRAM)	$1,39,04,288.91
Deluxe Road Bike (Shimano)	$1,36,77,737.12
Women's Off Road Bike Fully	$85,31,289.89
Men's Off Road Bike Hard Tail (Shiman	$82,02,623.91
Professional Touring Bike (black)	$80,03,347.88
Professional Touring Bike (red)	$75,08,426.40
Deluxe Touring Bike (red)	$73,65,045.01
Deluxe Touring Bike (black)	$69,69,848.42
Professional Road Bike (Campagnolo)	$45,07,696.85
Professional Road Bike (SRAM)	$41,94,554.69
Deluxe Road Bike (SRAM)	$40,83,196.51
E-Bike Tailwind	$40,23,195.21
Air Pump	$3,97,049.35
City Bike Max	$3,36,194.70
First Aid Kit	$1,93,335.06

Materials generating revenue in Germany
$40,23,195.21
$41,94,554.69
$40,83,196.51
$45,07,696.85
$69,69,848.42
$73,65,045.01
$75,08,426.40
$80,03,347.88
$82,02,623.91
$85,31,289.89
$1,36,77,737.12
$1,39,04,288.91
$1,50,92,526.81
$1,66,04,565.87
$1,89,08,929.74
$2,02,11,182.33
Professional Road Bike (Shimano)
Men's Off Road Bike Fully
Professional Touring Bike (silver)
Deluxe Touring Bike (silver)
Men's Off Road Bike Hard Tail (SRAM
Deluxe Road Bike (Shimano)
Women's Off Road Bike Fully
Men's Off Road Bike Hard Tail (Shima
Professional Touring Bike (black)
Professional Touring Bike (red)
Deluxe Touring Bike (red)
Deluxe Touring Bike (black)
Professional Road Bike (Campagnolo)
Professional Road Bike (SRAM)
Deluxe Road Bike (SRAM)
E-Bike Tailwind
Air Pump
City Bike Max
First Aid Kit
Water Bottle Cage
Off Road Helmet
Road Helmet
Elbow Pads
Knee Pads
Repair Kit
T-shirt
Fixed Gear Bike Plus
Water Bottle

The filter value was then changed to Germany to analyse and visualize the revenue trends of the country. Furthermore, the sorting order was then reversed (ascending sum of revenues) to determine the material with the lowest revenue in each of these countries seen in Fig 2.4.1(c). Fig 2.4.1(d) demonstrates the exploration of more sorting capabilities in Pivot tables facilitating the selection of Top 10 and Bottom 10 values revealing the respective revenue makers.

Fig 2.4.1(c): Pivot table and pie chart showing revenues generated by GBI materials in Germany



Country Desc	Germany
Row Labels	**Sum of Revenue**
Water Bottle	$46,808.03
Fixed Gear Bike Plus	$55,830.97
T-shirt	$71,480.06
Repair Kit	$73,681.90
Knee Pads	$76,151.71
Elbow Pads	$76,417.57
Road Helmet	$1,18,421.90
Off Road Helmet	$1,21,671.79
Water Bottle Cage	$1,45,351.43
First Aid Kit	$1,93,335.06
Grand Total	**$9,79,150.42**

y Desc	Germany
bels	**Sum of Revenue**
Professional Road Bike (Shimano)	$2,02,11,182.33
Men's Off Road Bike Fully	$1,89,08,929.74
Professional Touring Bike (silver)	$1,66,04,565.87
Deluxe Touring Bike (silver)	$1,50,92,526.81
Men's Off Road Bike Hard Tail (SRAM)	$1,39,04,288.91
Deluxe Road Bike (Shimano)	$1,36,77,737.12
Women's Off Road Bike Fully	$85,31,289.89
Men's Off Road Bike Hard Tail (Shiman	$82,02,623.91
Professional Touring Bike (black)	$80,03,347.88
Professional Touring Bike (red)	$75,08,426.40
Grand Total	**$13,06,44,918.86**

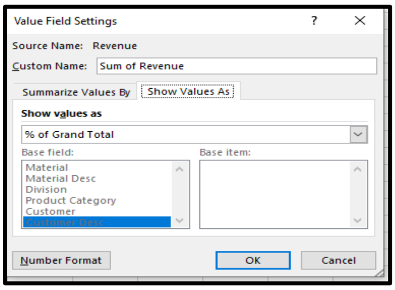
Top 10 Filter (Material Desc)

Show: Bottom, 10, Items, by Sum of Revenue

Fig 2.4.1(d): Revenue of the top 10 and bottom 10 Materials of GBI in Germany

Key Findings: Almost all the materials generate higher revenue in Germany in comparison to USA. However, the material — E-Bike Tailwind generates sales and revenue only in Germany so it can be assumed that this material is only sold in Germany making it a profit-making attribute for GBI. Also, this is one of the reasons why the annual revenue of Germany has been higher than USA in the past years. Overall, the material — Professional Road Bike (Shimano) had the highest total revenue for both USA and Germany whereas Water Bottle had the lowest aggregated revenue for both the countries. However, the amount of revenue generated in Germany by both these materials is comparatively higher than USA. To maximise profitability, the production or distribution of the material Professional Road Bike (Shimano) can be increased in both these countries to potentially achieve high levels of net sales.

2.4.2 Which customer has the highest percentage contribution to the total revenue? What has been the trend of that customer's percentage contribution over the years?



Value Field Settings

Source Name: Revenue

Custom Name: Sum of Revenue

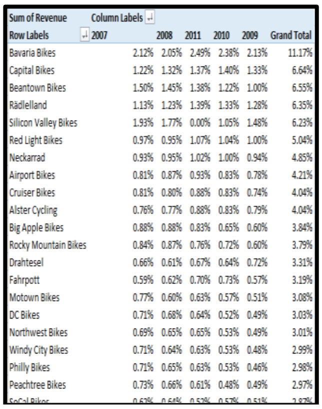
Show values as: % of Grand Total

Base field: Material, Material Desc, Division, Product Category, Customer, Customer Desc

Base item:

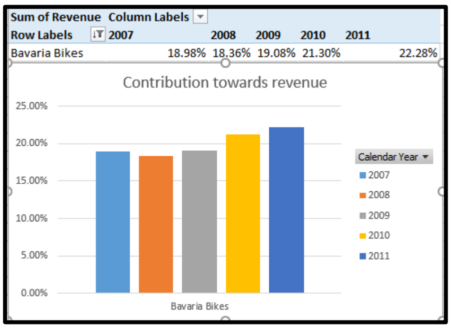
Fig 2.4.2(a). Formatting Values as % of grand total

The Customer Desc was added to rows, calendar year to columns, and sum of Revenue remained in Values in the pivot table builder. All existing filters were removed. The row labels were sorted in descending order of sum of revenue to reveal the top revenue makers. The **sum of revenue** Values field was formatted to show values as **% of Grand total** as shown in Fig 2.4.2(a). This resulted into a crosstab with each customer's total revenue generated over the years as shown in Fig 2.4.2(b). To check the trend of the customer with the highest percentage contribution to the total revenue, the data was further sorted to hide all the other customer information except the top one and to remove the Grand Total. After this, a pivot chart was created to check that customer's contribution over the years as shown in Fig 2.4.2(c).



Sum of Revenue	Column Labels					
Row Labels	2007	2008	2011	2010	2009	Grand Total
Bavaria Bikes	2.12%	2.05%	2.49%	2.38%	2.13%	11.17%
Capital Bikes	1.22%	1.32%	1.37%	1.40%	1.33%	6.64%
Beantown Bikes	1.50%	1.45%	1.38%	1.22%	1.00%	6.55%
Rädlelland	1.13%	1.23%	1.39%	1.33%	1.28%	6.35%
Silicon Valley Bikes	1.93%	1.77%	0.00%	1.05%	1.48%	6.23%
Red Light Bikes	0.97%	0.95%	1.07%	1.04%	1.00%	5.04%
Neckarrad	0.93%	0.95%	1.02%	1.00%	0.94%	4.85%
Airport Bikes	0.81%	0.87%	0.93%	0.83%	0.78%	4.21%
Cruiser Bikes	0.81%	0.80%	0.88%	0.83%	0.74%	4.04%
Alster Cycling	0.76%	0.77%	0.88%	0.83%	0.79%	4.04%
Big Apple Bikes	0.88%	0.88%	0.83%	0.65%	0.60%	3.84%
Rocky Mountain Bikes	0.84%	0.87%	0.76%	0.72%	0.60%	3.79%
Drahtesel	0.66%	0.61%	0.67%	0.64%	0.72%	3.31%
Fahrpott	0.59%	0.62%	0.70%	0.73%	0.57%	3.19%
Motown Bikes	0.77%	0.60%	0.63%	0.57%	0.51%	3.08%
DC Bikes	0.71%	0.68%	0.64%	0.52%	0.49%	3.03%
Northwest Bikes	0.69%	0.65%	0.65%	0.53%	0.49%	3.01%
Windy City Bikes	0.71%	0.64%	0.63%	0.53%	0.48%	2.99%
Philly Bikes	0.71%	0.65%	0.63%	0.53%	0.46%	2.98%
Peachtree Bikes	0.73%	0.66%	0.61%	0.48%	0.49%	2.97%
SoCal Bikes	0.62%	0.64%	0.52%	0.57%	0.51%	2.87%

Fig 2.4.2(b). Revenue trends of GBI customers over the years



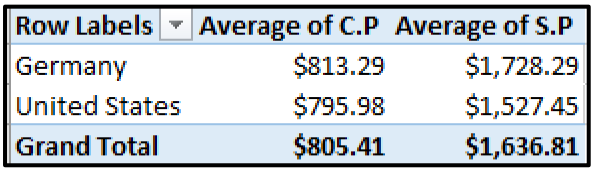
Sum of Revenue	Column Labels				
Row Labels	2007	2008	2009	2010	2011
Bavaria Bikes	18.98%	18.36%	19.08%	21.30%	22.28%

Fig 2.4.2(c): Revenue generated by Bavaria Bikes over the years

Key Findings: Bavaria Bikes is the top customer of GBI producing almost 11.7% of the total revenue. This customer produced revenue for GBI persistently over the years an in a linear increase fashion. In 2011, Bavaria Bikes formed 22.28 % of the total revenue generated by it over the years. Therefore, it can be concluded that Bavaria Bikes is the star customer of GBI and should be highly valued.

2.4.3 Are US and Germany buying and selling products at the same price?

It has been observed that Germany is clearly making more profits for GBI than USA. This might be true for several reasons, but one potential reason is both these countries are buying and selling the GBI products at different prices. This can be measured by calculating average cost price and selling price for each of the products in both regions. For this process, a pivot table was inserted for all the data in a new worksheet. A new calculated field called C.P (Cost price of materials) was created using Pivot Analyser - Fields, Items, & Sets → Calculated Field**.** The formula for this field ***C.P ='Cost of Goods M USD' /'Sales Quantity'***. Another calculated field called S.P (Selling price of materials) was created in the PivotTable Builder. ***S.P = ='Net Sales' /'Sales Quantity'.*** The Country Desc was added to Rows while S.P and C.P to Values in the PivotTable fields. Since the prices and sales quantity varies for different materials, average of C.P and S.P is calculated by formatting the Value Field Settings and summarizing the values as average which resulted into a crosstab as seen in Fig 2.4.3(a).



Row Labels	Average of C.P	Average of S.P
Germany	$813.29	$1,728.29
United States	$795.98	$1,527.45
Grand Total	**$805.41**	**$1,636.81**

Fig 2.4.3(a). The average Cost price and Selling price of products in USA and Germany

Also, the Values are formatted as $ (USD) for normalization. To illustrate this comparison of prices between these two countries, a pivot bar chart is inserted as shown in Fig 2.4.3(b).

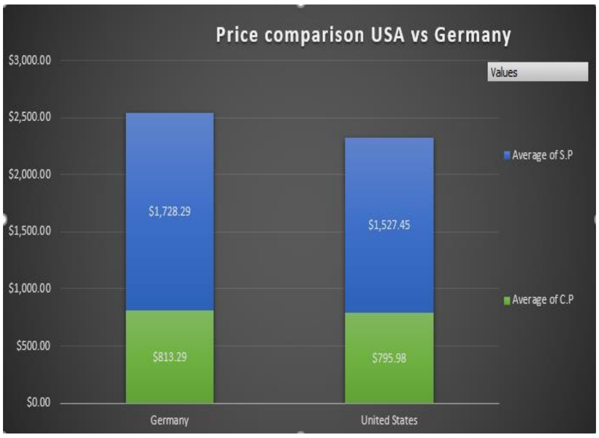
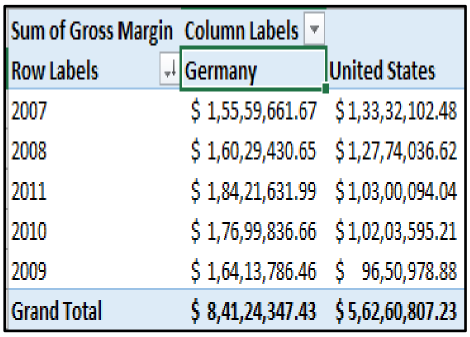


Fig 2.4.3(b): Product price of USA vs Germany

Key findings: Germany is buying each material at an average price of $813.29 which is $7 higher than USA but selling each of the products at an average price of $1728.29 which is $200 more than USA. Since USA is buying and selling products at a lower price than Germany therefore the later becomes a huge contributor to the revenue/profitability of GBI.

2.4.4 Which year had the highest gross margin? How does the profitability differ between US and Germany for the year with the highest gross margin?



Sum of Gross Margin	Column Labels	
Row Labels	Germany	United States
2007	$ 1,55,59,661.67	$ 1,33,32,102.48
2008	$ 1,60,29,430.65	$ 1,27,74,036.62
2011	$ 1,84,21,631.99	$ 1,03,00,094.04
2010	$ 1,76,99,836.66	$ 1,02,03,595.21
2009	$ 1,64,13,786.46	$ 96,50,978.88
Grand Total	$ 8,41,24,347.43	$ 5,62,60,807.23

Fig 2.4.4(a): Total gross margin in Germany vs USA over the years.

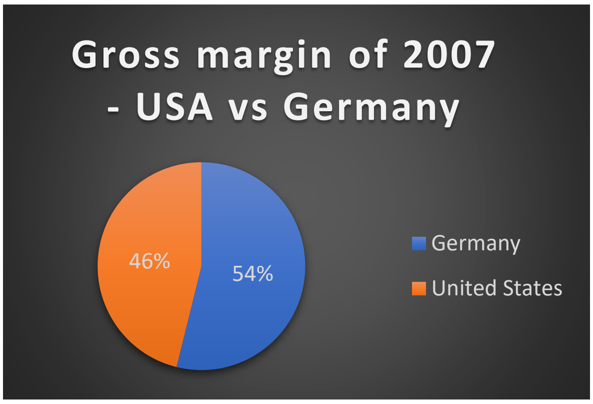


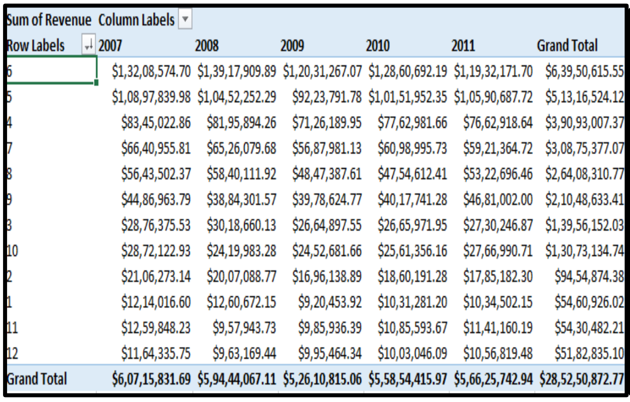
Fig 2.4.4(b) Gross Margin of USA vs Germany for the year 2007

A pivot table is inserted for all the data in a new worksheet. A calculated filed called "gross margin" (Profit acquired) is created using the formula **= *'Net Sales' – 'Cost of Goods M USD'***. in the PivotTable Builder, Gross Margin is added to Values and Calendar Year to Rows. This resulted in a crosstab in which the Rows were sorted by descending sum of gross margin to determine the year with the highest gross margin value as seen in Fig 2.4.4(a). To make effective profitability comparisons between US and Germany for the year with the highest gross margin, the Calendar Year is moved to Filters and Country Desc is added to Rows. Filtered the year with the highest gross margin and inserted a pivot pie chart to illustrate the profit comparison (Fig 2.4.4(b). Used advanced chart styles to show the percentage composition of the two countries in the total gross margin.

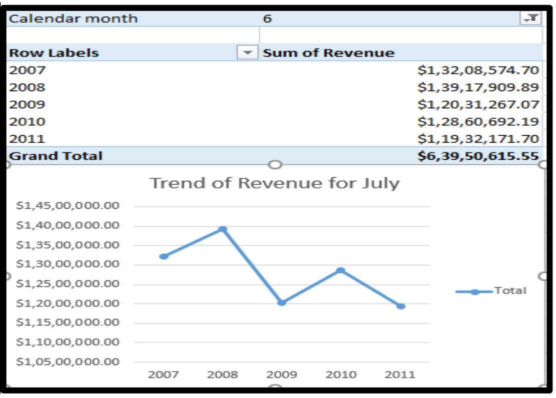
Key Findings: The year 2007 had the highest gross margin. Germany had highly contributed to GBI's profitability (8 percent more) in comparison to USA for that year.

2.4.5 Is there a seasonality in revenue during the year? If so, what month has the highest revenue? Is the seasonality similar from year to year?

In the PivotTable Builder, Calendar Month was added to Rows, Calendar Year to Columns and Revenue in Values. To determine the month with the highest revenue, the rows were sorted by descending sum of revenue. This resulted into a crosstab as shown in Fig 2.4.5(a)



Sum of Revenue	Column Labels				
Row Labels	2007	2008	2009	2010	2011	Grand Total
6	$1,32,08,574.70	$1,39,17,909.89	$1,20,31,267.07	$1,28,60,692.19	$1,19,32,171.70	$6,39,50,615.55
5	$1,08,97,839.98	$1,04,52,252.29	$92,23,791.78	$1,01,51,952.35	$1,05,90,687.72	$5,13,16,524.12
4	$83,45,022.86	$81,95,894.26	$71,26,189.95	$77,62,981.66	$76,62,918.64	$3,90,93,007.37
7	$66,40,955.81	$65,26,079.68	$56,87,981.13	$60,98,995.73	$59,21,364.72	$3,08,75,377.07
8	$56,43,502.37	$58,40,111.92	$48,47,387.61	$47,54,612.41	$53,22,696.46	$2,64,08,310.77
9	$44,86,963.79	$38,84,301.57	$39,78,624.77	$40,17,741.28	$46,81,002.00	$2,10,48,633.41
3	$28,76,375.53	$30,18,660.13	$26,64,897.55	$26,65,971.95	$27,30,246.87	$1,39,56,152.03
10	$28,72,122.93	$24,19,983.28	$24,52,681.66	$25,61,356.16	$27,66,990.71	$1,30,73,134.74
2	$21,06,273.14	$20,07,088.77	$16,96,138.89	$18,60,191.28	$17,85,182.30	$94,54,874.38
1	$12,14,016.60	$12,60,672.15	$9,20,453.92	$10,31,281.20	$10,34,502.15	$54,60,926.02
11	$12,59,848.23	$9,57,943.73	$9,85,936.39	$10,85,593.67	$11,41,160.19	$54,30,482.21
12	$11,64,335.75	$9,63,169.44	$9,95,464.34	$10,03,046.09	$10,56,819.48	$51,82,835.10
Grand Total	$6,07,15,831.69	$5,94,44,067.11	$5,26,10,815.06	$5,58,54,415.97	$5,66,25,742.94	$28,52,50,872.77



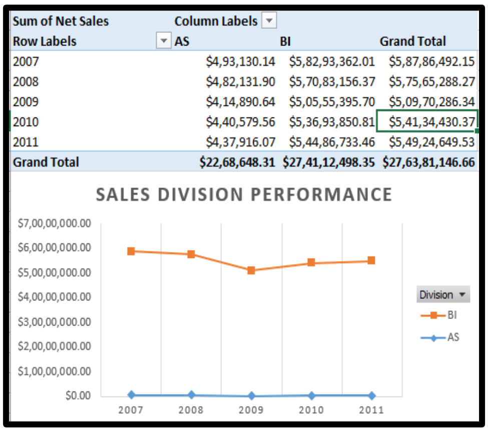
Calendar month	6
Row Labels	Sum of Revenue
2007	$1,32,08,574.70
2008	$1,39,17,909.89
2009	$1,20,31,267.07
2010	$1,28,60,692.19
2011	$1,19,32,171.70
Grand Total	$6,39,50,615.55

Fig 2.4.5(a): Month with the highest revenue

Fig 2.4.5(b) Revenue trends for July over the years

Key Findings: June (6) is the month with the highest revenue. This has been a continuous trend over all the years (2007-2011). It can be concluded that during May and June (the beginning of summers) GBI does maximum sales and generates highest revenue. Fig 2.4.5(b) demonstrates how the revenue trend for July has been changing over the years.

2.4.6 Are all the sales divisions equally performing towards selling products during the years 2007- 2011?



Sum of Net Sales	Column Labels		
Row Labels	AS	BI	Grand Total
2007	$4,93,130.14	$5,82,93,362.01	$5,87,86,492.15
2008	$4,82,131.90	$5,70,83,156.37	$5,75,65,288.27
2009	$4,14,890.64	$5,05,55,395.70	$5,09,70,286.34
2010	$4,40,579.56	$5,36,93,850.81	$5,41,34,430.37
2011	$4,37,916.07	$5,44,86,733.46	$5,49,24,649.53
Grand Total	**$22,68,648.31**	**$27,41,12,498.35**	**$27,63,81,146.66**

Fig 2.4.6(a): Performance of the two sales division over the years

In the PivotTable builder, Division is added to Columns, Calendar Year to Rows and Net Sales to Values which is summarized as Sum of Net Sales. The resulted crosstab is visualized using a pivot line chart as seen in Fig 2.4.6(a)

Key Findings: Over the years, the Sales Division AS has done very poor sales for GBI in comparison to BI. On the other hand, the sales performance of the BI division has been constant with a slight decline in the year 2009. GBI can focus on the AS division to analyse the reasons of poor performance and streamline the sales operation in an optimised way.

SECTION 5: ANALYSIS AND CRITIQUE OF THE TOOL

Having intermediate proficiency of Excel, I could easily create crosstabs and pivot charts from this dataset. The pivot table feature of Excel is easy to follow and can help beginners perform an in-depth analysis of the dataset. Pivot tables have a simple interface, and a user can easily get accustomed to its basic functioning. I could easily create reports at a fast speed having accurate results. There is much flexibility of rearranging the table to suit the business case needs. One can consistently format the data in real-time. For complex analytics project, I usually refer to pivot tables first to speculate and explore the different data categories and its relationships. This helps me to gain a good understanding of the data and later use advanced analytics tools to build cognitive BI solutions. Although, pivot tables are efficient for data analysis, I cannot state the same for its supported visualization constituents — Pivot charts. Having explored advanced automated data visualization tools, I feel that there are very limited types of pivot charts with very less variation. Also, pivot charts require manual effort for reformatting values, axis titles, chart titles, etc. This feature can be upgraded to visualize data in an optimized way.

SECTION 6: CONCLUSION

The functional purpose of pivot tables can help analysts understand their business problems and gain comprehensive knowledge of the data values but there are several advanced tools that support building of impactful BI solutions with much ease and optimization.

CHAPTER 3 : SAP PREDICTIVE ANALYTICS FOR VISUALIZATION

SECTION 3.1 - INTRODUCTION

SAP Predictive Analytics can be described as a business technology platform designed for advanced data preparation, developing insightful visualizations, and building predictive models (SAP Predictive Analytics, 2016). SAP PA have two versions for target users-desktop (two layered architecture) and enterprise (three-tier client server architecture) (SAP PA, n.d.). For this analysis, desktop version and Expert Analytics view was used. The features of Expert Analytics include the following:

- Automate data prep, visualize, story compositions, predictive modeling and deployment. Harness in-database predictive scoring for a wide range of target systems. Integrate multiple data sources together which include unstructured data.
- Automated Predictive Modelling helpful in understanding data patterns and relationships and support decision-making.

SECTION 3.2 - DATASET DESCRIPTION

SAP Predictive Analytics will be used to analyse and visualize the sales and customers data for the years 2015-2018 of a US organization called GBI Inc. Both the datasets were provided in class as two separate Excel Workbooks (.xlsx) which included attributes that were further examined, evaluated, and manipulated to understand the past performance of GBI and support fact-based decision-making.

GB_Data_GM.csv – Contains information about the sales data for GBI in two different currencies (Euros and USD) for its several products and customers in two countries— USA and Germany over the years (2015-2018). It includes 17 columns (Year, Month, Day, Sales Quantity, Revenue, Currency, discount, COGS, Gross Margin, etc) of examinable data attributes which would be utilized in the tool for advanced data manipulation, processing, analyzing and visualizing.

Customers.csv — Contains information about the customers of GBI Inc belonging to different sales organizations and locations in USA and Germany. It includes 6 columns of examinable data values like Customer, Customer description, Location, Country, Sales Organization and Language Key. No data cleaning was required for this dataset and was merged with the GB_Data_GM dataset to analyze the geographic demographics of GBI's customers.

SECTION 3.3 BUSINESS PROBLEM AND RESEARCH QUESTIONS

3.3.1 – Business Use Case

Profitability Analysis helps organizations to assess the growth of their business, keep track of business performance, and identify the areas where profits can be maximised. This business scenario is focussed on the profitability analysis of a company called Global Bike Inc and learn about the changing market trends and consumer demands. Therefore, to speculate the data patterns, establish data relationships and visually analyse the company's profits for different products, customers, regions, and years the following research questions have been formed.

3.3.2- Research questions

1. Is there seasonality in revenue during the year? If so, what month has the highest revenue and what is the day of that month? Is the seasonality similar from year to year?

2. Which customer has the highest gross margin ratio on average? What has been the trend of that customer's profit contribution over the years?

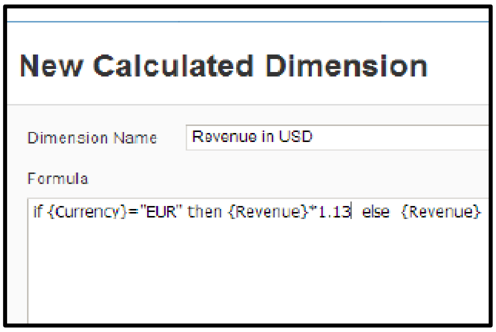
3. Which product has the highest sales quantities? Does it appear that the ratio of sales by product changes over time?

4. How does the profitability differ between US and Germany for the year with the highest gross margin? Are US and Germany buying and selling products at the same price?

SECTION 3.4 ANALYTICAL PROCESS AND KEY FINDINGS

3.4.1 Is there is any seasonality for sales at GB? What is the day of the month with the highest sales? What are the revenues on this day?

For this question, the dataset — GB_Data_GM.xlsx — was loaded and visualized using the Expert Analytics PA tool. In the prepare view, it was observed that the dataset had the measurable attributes like Revenue, COGS, etc in Euros and USD currency. To develop a uniform currency model, ***new calculated dimensions which configure certain measurable quantities like Revenue, Discount and COGS in USD were created using the formula = if {Currency} ="EUR" then {Revenue}*1.13 else Revenue*** (as seen in Fig 3.4.1(a)). ***Also, another calculated field called Net Sales = {Revenue in USD} – {Discount in USD} was created.*** Furthermore, the new calculated dimensions were converted to measures by clicking on the cog for each dimension (e.g., Revenue in USD) and "Create a Measure" option was chosen from the dropdown list. For this question as monthly data is expected, a time hierarchy was created from the dimension – Year by converting it to a number format, renaming the new dimension to Years and later choosing "Create a Time Hierarchy" from the cogged drop-down menu as seen in Fig 3.4.1(b)



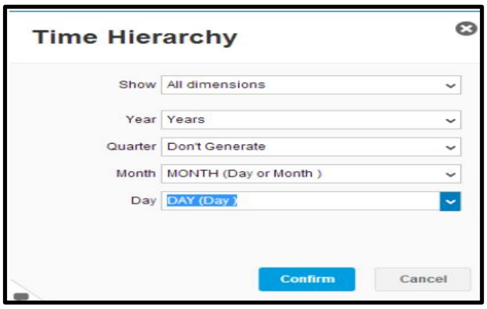
New Calculated Dimension

Dimension Name: Revenue in USD

Formula:

```
if {Currency}="EUR" then {Revenue}*1.13 else {Revenue}
```

Fig 3.4.1(a): Formula for Revenue in USD



Time Hierarchy

Show: All dimensions

Year: Years

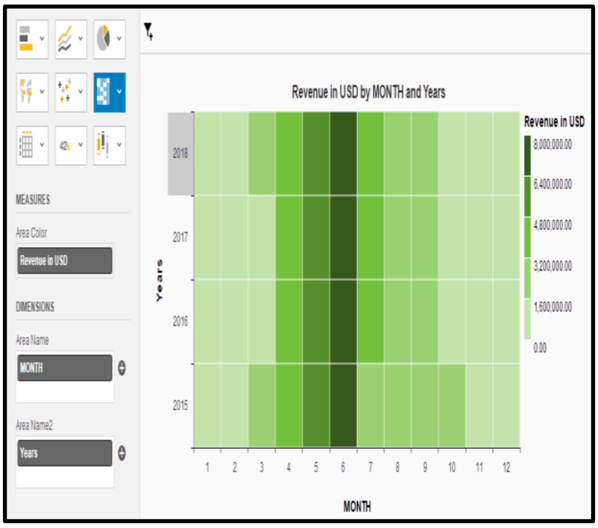
Quarter: Don't Generate

Month: MONTH (Day or Month)

Day: DAY (Day)

Fig3.4.1(b): Time Hierarchy Settings

To create visualization, in the Visualize view, a heat map was chosen. Revenue in USD was added to the Area Color, month to Area Name and Years to Area Name 2. This resulted into the visualization as seen in Fig 3.4.1(c) To further find the day of the month with the highest sales, the month with the highest revenue was focussed and filtered, also Day was added to Area Name 2. This resulted into the visualization in Fig 3.4.1(d)



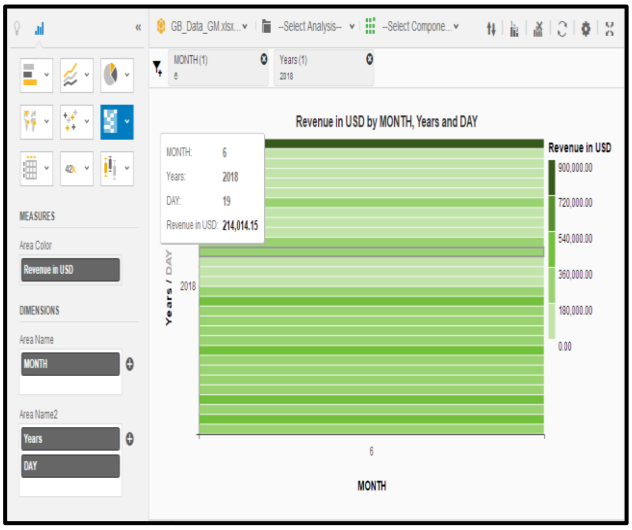
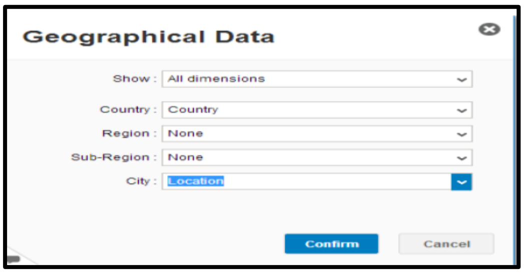


Fig 3.4.1(c)Revenue per Month and Years **Fig 3.4.1(d) Revenue per day for the Month of May**

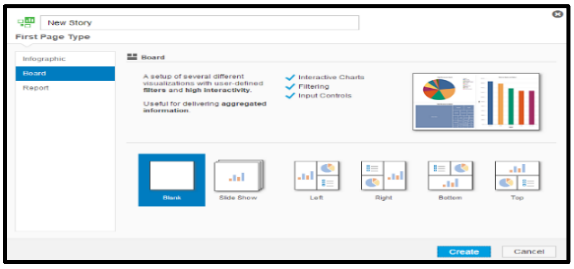
***Key Findings**:* GBI produced the highest revenue — 7,160,869.99 USD — during the month of May of the year 2018. This has been a continuous trend over all the years (2015 -2018). Also, 30th May 2018 was the day of highest revenue and sales for GBI. It can be concluded that during May (the beginning of summers) GBI does maximum sales and generates highest revenue.

3.4.2 What has been the customer trend of GBI over the years? Which customer has the highest gross margin, revenue, and gross margin ratio on average?

For this analysis, a geographic hierarchy and a new calculated field called ***Gross Margin in USD = {Net Sales in USD} – {COGS in USD}*** was created in the Data preparation mode. Another dataset called Customers.xlsx file was loaded and merged with the existing Sales data. The integrated dataset included a dimension- Country from which a geographic hierarchy was established as seen in Fig 3.4.2(a). In the Visualize mode, three visualizations were created to compose a story demonstrating the customer trend of GBI over the years and across countries. First chart was a Geo Pie chart, having Gross Margin in USD in Value, Country in Geography and Customer in Overlay Data. The second chart was a 3D column chart, having Revenue in USD and Net Sales in Measures, Years in X- Axis and Customer in Legend Color. The chart was optimized using the Rank option to select the top 5 customers. The third being a Bubble Chart, Sales Quantity was placed on the X-axis, Discount in USD on the Y-axis*, **Gross Margin Ratio (a calculated field created using formula – Gross Margin in USD/ Net Sales)*** as the bubble width *and* Customer was added to Legend Color. Later in the compose mode, New Story -> Board (Fig 3.4.2(b)) was chosen, and all three charts were inputted to create a visual as seen in Fig 3.4.2(c)


Geographical Data

Show : All dimensions
Country : Country
Region : None
Sub-Region : None
City : Location


New Story
First Page Type

Board

A setup of several different visualizations with user-defined **filters and high interactivity.**

Useful for delivering **aggregated information.**

✓ Interactive Charts
✓ Filtering
✓ Input Controls

Fig 3.4.2(a) Geographical Hierarchy Setting **Fig 3.4.2(b) Selecting dashboard from compose mode**

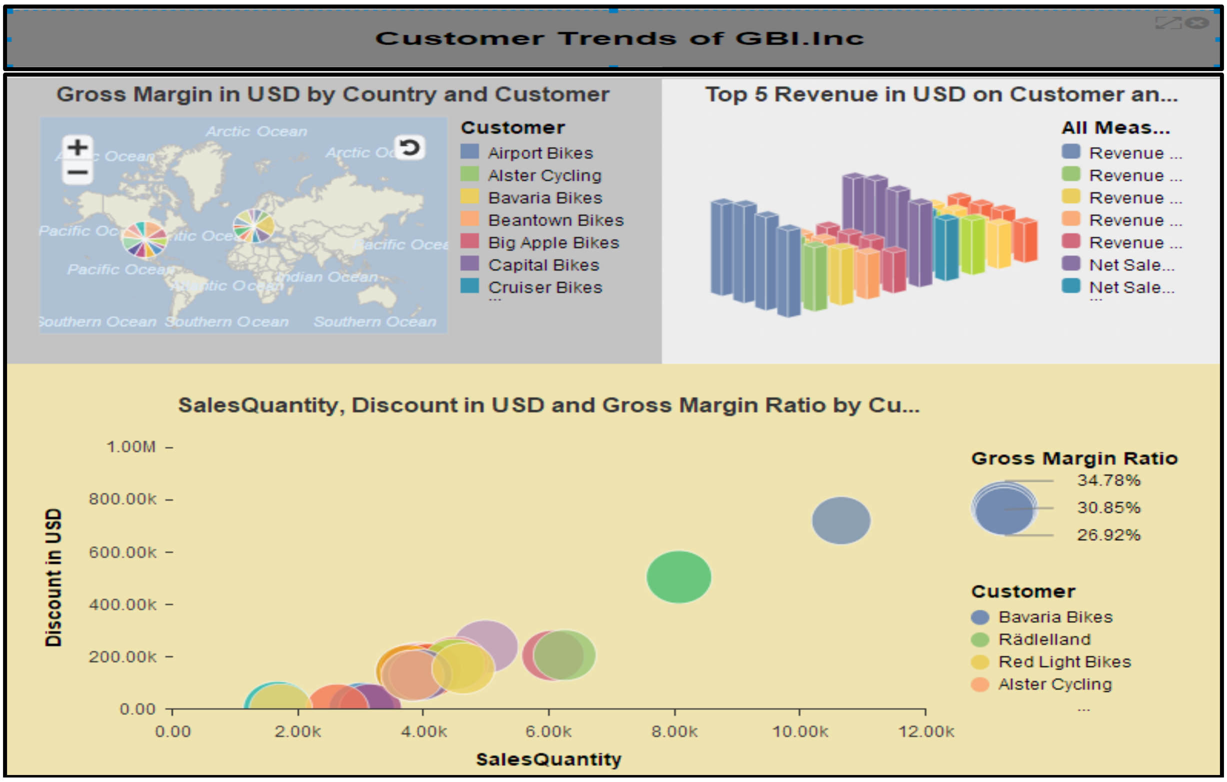


Fig 3.4.2(c)Customer Trends Dashboard of Global Inc.

Key Findings: Bavaria Bikes customer from Germany has been responsible for the highest Sales Revenue, Net Sales and Gross Margin of GBI over the years followed by Beantown Bikes customer from USA. Both these customers have gained maximum discount on the GBI products too. However, GBI has gained only 26.92% profit over their net sales to Bavaria Bikes, while gaining 31.41% profit over the net sales done for Beantown Bikes. Furniture City Bikes being the least performing customer of GBI has however gained 34.78 % profit over their net sales having a zero-discount benefit from GBI.

3.4.4 Which product has the highest sales quantities? Does it appear that the ratio of sales by product changes over time?

To understand the product trends of GBI, all the profitability measures – Revenue, Sales Quantity, Net Sales, Discount and Gross Margin ratio — had to be evaluated against the Dimension Product, Years and Country. Therefore, a trellised pie chart was constructed by selecting pie chart and adding Sales Quantity as the pie Sectors and Product as Legend Color also adding the Years dimension in Columns. Thereafter, using the Rank component only the top 5 products having high sales quantity were filtered for the last two years — 2017 and 2018. To understand the trend of gross margin ratio for products, a Tag Cloud chart was selected, having Gross Margin Ratio as Word Weight, Revenue in USD as Word Color and Customer in Dimensions. Thirdly, to analyze the correlation between all the measures, a Marimekko chart was selected, in the Measures section, Revenue in USD was added to Y-axis and Discount in USD in Column Width but in the Dimensions section, X Axis had Years and Legend Color had Product. Using the compose mode, all the three charts were then infused in a story dashboard as seen in Fig 3.4.4(a)

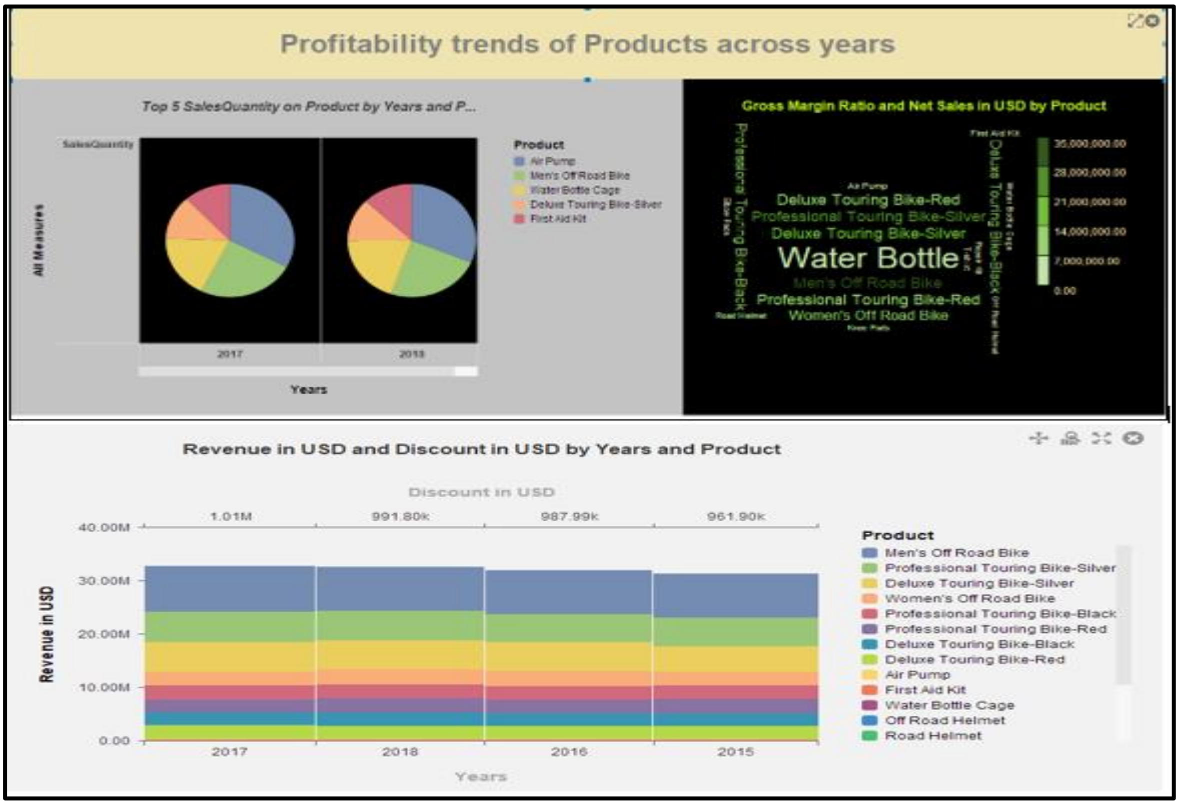


Fig 3.4.4(a) Product trends Dashboard of GBI

Key Findings: The product — **Men's Off Road Bike**— has produced the highest Revenue for GBI over the years with maximum Net Sales. It makes sense that as there was quite a good amount of discount on this product which accelerated its capability to sell out in huge quantity over the years. However, the gross margin percentage for this product is not that high whereas the product **Water Bottle** with very minimum Revenue, Discount and Net Sales has the highest gross margin percentage of 40.32 %. If GBI decides to invest more in this product, it has the potential to increase all its profitable measures.

3.4.5 How does the profitability differ between US and Germany for the year with the highest gross margin? Are US and Germany buying and selling products at the same price?

It is important to analyse how the two countries — USA and Germany— are contributing towards revenue generation for GBI. To achieve this, in the Visualize mode, a pie chart was created to find Revenue of the two different countries. Later, a Tree Map was created to visualize the Gross Margin and Sales Quantity distribution by different regions in each Country. In the Prepare Mode, two calculated fields — Cost Price = {COGS in USD}/{Sales Quantity} and Selling Price = {Net Sales}/{Sales Quantity} — were created to demonstrate the average cost and selling price of all the products sold in USA and Germany. A Bar chart with Cost Price and Selling Price in Y-Axis and Country in X-Axis was constructed. All the three illustrations were used to create a Dashboard as seen in Fig 1.4.5 (a).

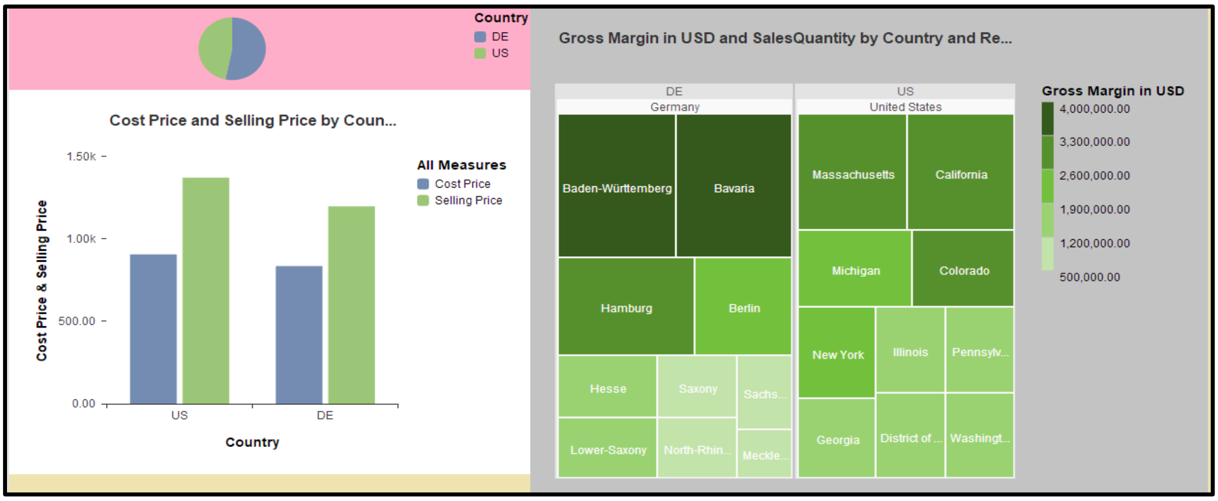


Fig 3.4.5(a) Profitability contribution of GBI as per Country

Key Findings: 53.19% of GBI's Revenue is coming from Germany with Baden-Wurttemberg and Bavaria regions generating high rates of Gross Margin and Sales Quantity. 46.81% of the Revenue is coming from USA where the regions Massachusetts and California remain the highest contributors towards GBI's gross margin. Although USA is buying products at an average rate of 906.15 and selling them at 1371.63 average rate which is much higher than Germany, additional metrics like customers, discount rates, product popularity, etc as seen in the above questions are supporting Germany to become the highest profit contributor of GBI.

SECTION 5: ANALYSIS AND CRITIQUE OF THE TOOL

As a first-time user of SAP Predictive Analytics, I did not find the working and functionality of this tool hard to follow. The tool facilitates advanced data preparation and visualizing which made it easier to implement data manipulation techniques (merging, formulas, formatting, etc) smoothly. I found the tool user-friendly in developing illustrations of several types also enjoyed creating different types of charts and story dashboards. However, few features like Rank and Undo shortcut key (Ctrl +Z) do not always yield accurate results. Also, unlike Tableau, any changes made on the individual charts do not reflect in real-time in the story dashboard and vice-- versa. Although, I could explore most of the chart styles for the visualization process, I am looking forward to exploring the predictive modelling component of the tool for accurate and automated prediction of future insights (SAP PA, n.d.). I am also looking forward to exploring SAP Predictive Analytics interoperability with R scripts and live data connectivity with SAP Lumira Cloud.

SECTION 6: CONCLUSION

SAP Predictive Analytics is no doubt efficient in merging, manipulating, analysing, and visualizing the underlying datasets. With real-time insights offered by this tool, organizations can effectively examine customer behavior for profitable results, gain better understanding of the business and support reliable decision-making (SAP PA, n.d.), making it one of the effective BI- tools. The automated machine learning algorithms feature of this tool fascinate me, and I am determined to use SAP Predictive Analytics in the future.

CHAPTER 4 TABLEAU DESKTOP

SECTION 4.1 INTRODUCTION TO THE TOOL

Tableau is one of the most preferred data-visualization oriented analytics platforms due to its user-friendly interface, fast analytics, and organization-wide knowledge engagement capabilities (Backaitis,2018). Tableau Desktop includes interactive workbooks and dashboards to uncover valuable insights from data and share knowledge discovery easy via visual patterns. The main features of Tableau include toggle-view, drag-and-drop, data cleaning, merging, translate queries into visuals, commenting and sharing dashboards, creating stories, etc. (Anurag,2018). Although Tableau has several versions for different user types, for this analysis, Tableau Desktop personal is used.

SECTION 4.2 DATASET DESCRIPTION

Tableau Desktop was used to analyze the following datasets.

UK-Bank-Customers.csv: This dataset contains information of the customers of a UK Bank which has branches in several locations. It includes 9 data columns out of which 4 categorical variables (Gender, Age, Region and Job Classification) and discrete data like Bank Balance were analyzed to create a Customer Segmentation Dashboard for the Bank which would help any financial institution to understand customer demographics and tailor their services as per different customer groups.

Startup-Expansion.csv: This dataset contains information of the sales revenue of a small laundry-pickup services start-up company — WeWashYouSleep — in USA and the funds used in marketing their different stores. It includes 7 examinable data attributes (Store ID, City, Sate, Sales Region, New Expansion, Marketing Spend and Revenue) that were utilized to create custom territories and marketing clusters using Tableau to help the company build a vast network in the market.

World_Bank_CO2.csv: This dataset contains information of CO2 emission in Kilo Tons(KT) and CO2 emission per capita(metric tons)by different countries and regions across the globe from the year 1960. This dataset has both the cleaned and uncleaned versions of the data.

SECTION 4.3 BUSINESS PROBLEM AND RESEARCH QUESTIONS

4.3.1 – Business Use Case: Customer Segmentation helps organizations classify their customers characteristics into features that can help identify their ideal customers and employ diverse marketing strategies. This business scenario is focussed on the customer segmentation for a bank operating at UK.

Research questions

1. Describe the customer demographics of the Bank by balance, age, gender, region, and job classification.
2. How does the different regions of UK differ from one another?

4.3.2 – Business Use Case: This business problem is to analyze the sales and marketing trends of a small start-up business called WeWashYouSleep and develop valuable insights to support their expansion in smaller cities.

Research questions

1. Which of the two sales regions is performing better?
2. Which of the 10 new locations have the best potential for the company to invest more funds into marketing?

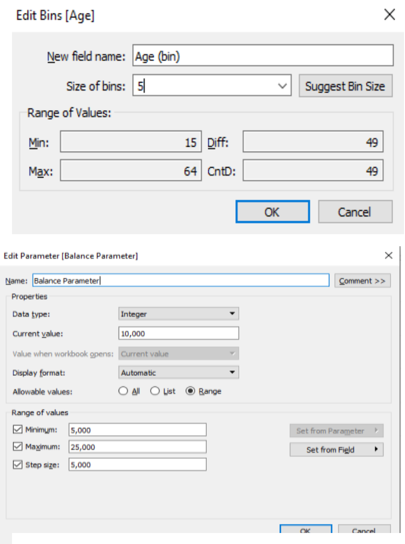
4.3.3 – Business Use Case: This business problem is to visualize the trend of CO2 emission levels by years for several countries across the globe.

Research questions

1. Show the trend of CO2 Emission by countries over the years? What is the trend of CO2 emission level for USA over the years?

SECTION 4.4: ANALYTICAL PROCESS AND KEY FINDINGS

4.4.1. Describe and compare customer demographics by balance, age, gender, region, and job classification?


Edit Bins [Age]

New field name: Age (bin)

Size of bins: 5

Range of Values:

Min: 15 Diff: 49

Max: 64 CntD: 49

Edit Parameter [Balance Parameter]

Name: Balance Parameter

Properties

Data type: Integer

Current value: 10,000

Value when workbook opens: Current value

Display format: Automatic

Allowable values: ○ All ○ List ● Range

Range of values

☑ Minimum: 5,000

☑ Maximum: 25,000

☑ Step size: 5,000

Fig 4.4.1(a): Age Bins settings; Balance Parameter settings

The dataset, UK-Bank-Customers, was added to Tableau and in Sheet1, a geographic map was created by dropping Region to the sheet and the default measure count object (indicating the number of records) in Detail. In sheet 2, using a pie chart, gender distribution was visualized. In Sheet 3, instead of visualizing Age as a single entity, it was categorized using bins as seen in Fig 4.4.1(a) and then Age (bins) was dropped to Columns and number of customer measure on rows. The Y-axis setting was changed to display number of customers as percentage of Total. Similarly, the Balance attribute was also categorized using bins, with Balance (bins) in Columns and % of Customers in Rows of a new Sheet 4. To leverage the power of parameters in Tableau, a balance parameter was established (lower section of Fig) to allow users adjust the balance chart as per different bank balance ranges by editing the measure and choose Balance parameter option in Size. This Balance parameter is then displayed as a slider in the worksheet. In Sheet 5, to visualize the number of customers of the bank as per their job types, a tree map was used. All these 5 sheets were then used to create a customer segmentation dashboard as seen in Fig4.4.1(b)

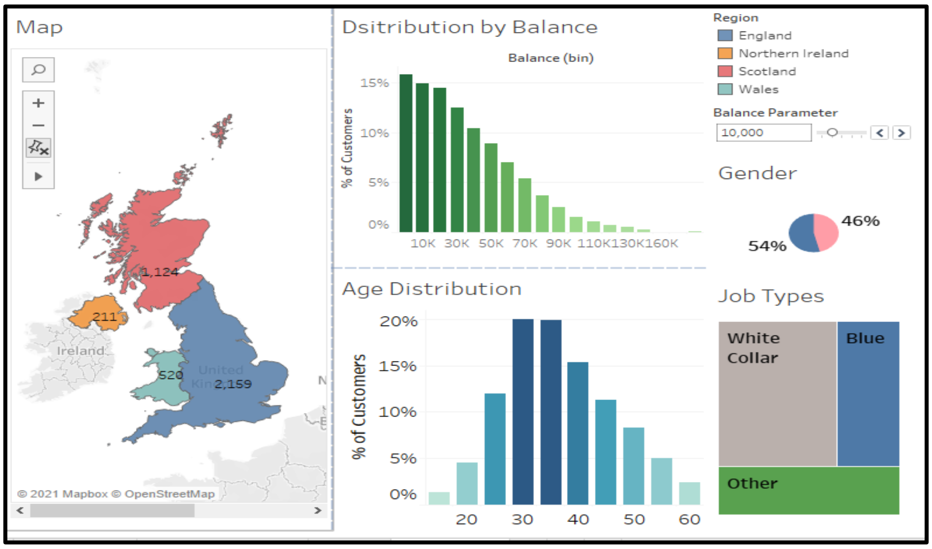


Fig 4.4.1(b)- Customer Segmentation Dashboard of UK-Bank

Key Findings: In this customer baseline dashboard, most of the customers are in the region England for the UK-Bank and the least in Wales. The female customers (illustrated as pink in the pie-chart to be intuitive) form 46 % of the overall customers while male population being the majority – 54%. In the breakdown of the Balance band, it can be observed that the lower is the balance band (0k to 30k Bank Balance), the more customers are found in that category aggregating to 45% of the total customers. From the Age sheet, it can be observed that almost 40% of the total customers are in their 30s. Also, almost half of the Bank customers have White Collar jobs. It will be interesting to see how each of these regions differ from the baseline customer observations.

4.4.2 How does the different regions of UK differ from one another?

To illustrate how each region of UK differ from one another and the baseline, the story feature of Tableau was used, and the customer segmentation dashboard was dropped in the story for further analysis. The various parts of the story can be seen in Fig 4.4.2 (a).

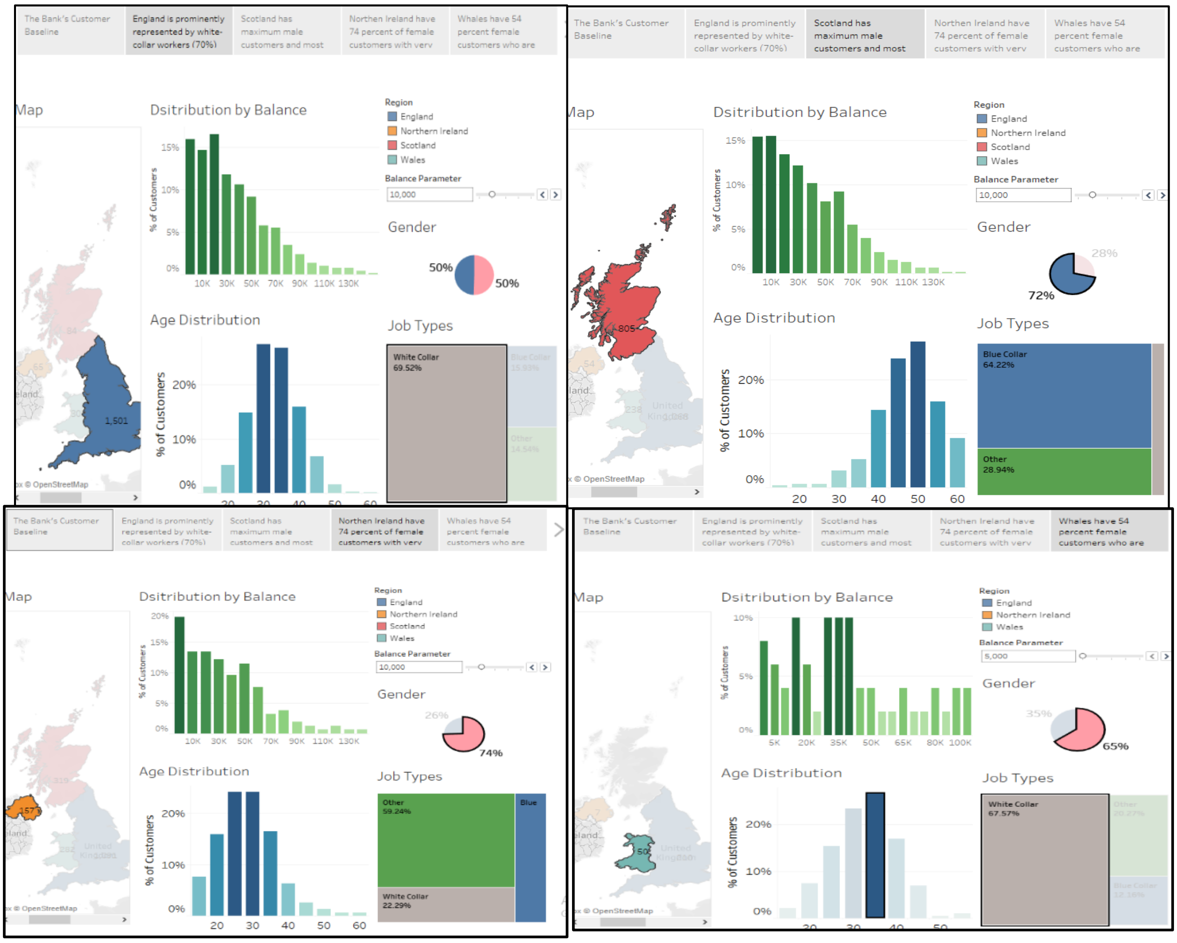


Fig 4.4.2(a) Different sections of the Customer Segmentation Story board

Key Findings : England has a maximum number of White-Collared customers (70 percent), may be because London is the economical hub of UK. Scotland has a vast majority of male customers (70 percent) with very less people in the White-Collar industry. Northern Island and Whales have majority of Female customers with most of them have White-Collar jobs. Therefore, we can conclude that the UK-Bank should market and tailor its services targeting

White-collar job customers and approach the female customers for Northern Ireland and Whales.

4.4.3 *Which of the two sales regions is performing better?*

For this question, the dataset, StartupExpansion.xlsx, was loaded into the Data view. As the sale regions are spread across different cities of United States, analyzing each sale region required grouping geographical territories. Also, to make performance comparisons between the two sales region, three metrics were considered — Average Revenue per city per sale region, average Marketing speed per city and average ROMI (Return on Marketing investment) per city. To visualize this, State was dropped into Sheet 1, Marks was changed to Maps. On doing this, Null values appeared which were rectified by changing the location settings to USA. Revenue dropped in Label and Sales Region in Color. This resulted into a visualization which showed revenue of each city in each Sales region. But, to compare average revenue between the two regions, grouping of the cities was required. So, each city in of Region 2 were selected and on right click, the Group option (By All dimensions) was chosen, Revenue was changed to Average, as a result Fig 4.4.3(a) was created.

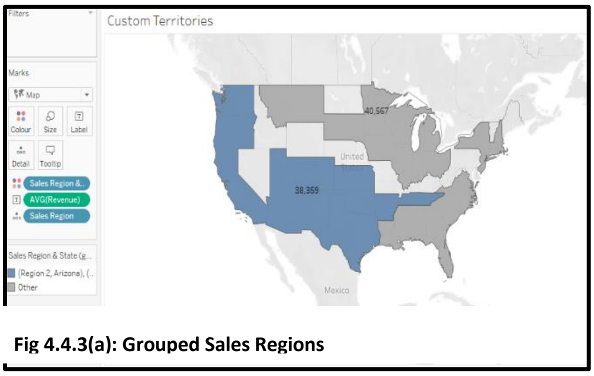


Fig 4.4.3(a): Grouped Sales Regions

The Color was edited and changed to orange for Region 1 and Blue for Region 2. Marketing spend was changed to Average and dropped into detail. A calculated field called ROMI was generated using formula = Revenue/Marketing spend, changed to Average and dropped to Label, which created the final illustration as seen in Fig 4.4.3(b)

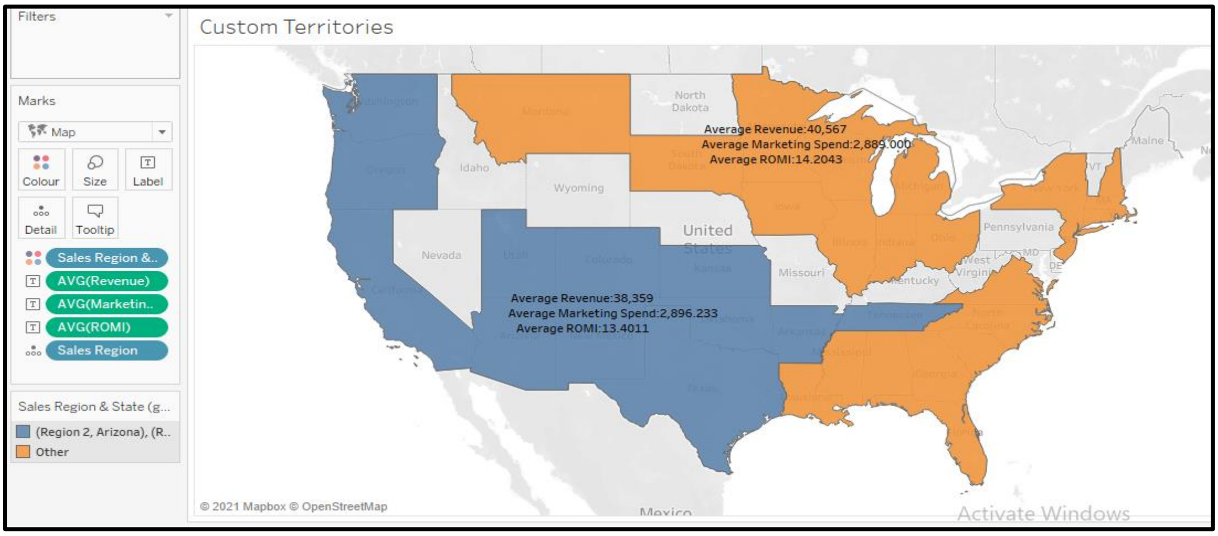
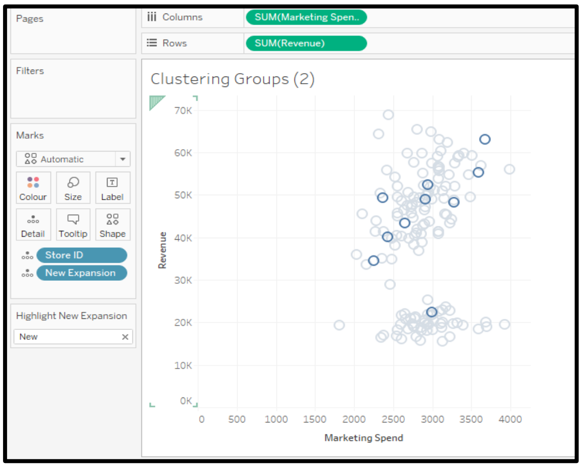


Fig 4.4.3(b) Displaying the two sales regions with their average Revenue, Marketing Spend and ROMI respectively

Key Findings: Sales Region 1 is producing more average Revenue per city in comparison to Region 2 also getting more returns on marketing investment while the later metric remains almost same for both the regions. Therefore, Region 1 is comparing better than Region 2.

4.4.4 Which of the 10 new locations have the best potential for the company to invest more funds into marketing?

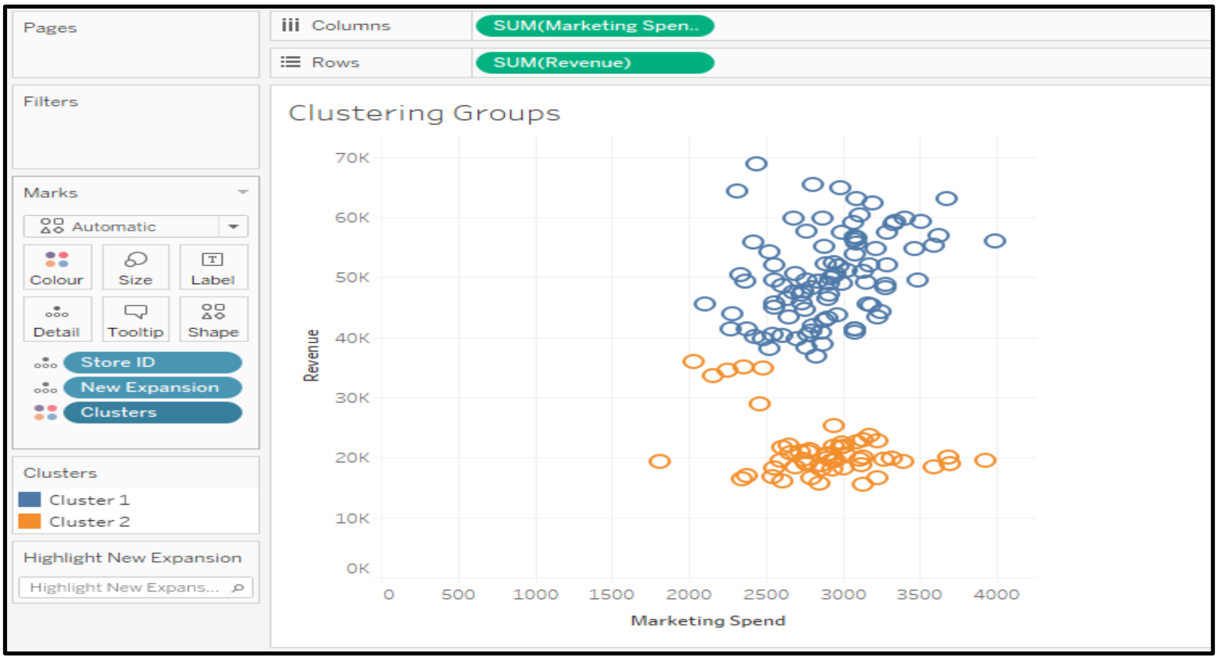
It has been observed that Sales Region 1 is clearly making more profits for the start-up than Sales Region 2. As, this a start-up company trying to expand their business in several locations, it is important for them to understand if they are gaining many returns on their marketing investments. To analyze this, Revenue was dragged to Rows and Marketing Spend to Columns. Store ID and New Expansion were added to Detail. New Expansion was highlighted to identify the new and old expansions across USA by selecting "Show Highlighter" option. This resulted into a visualization as seen in Fig 4.4.4(a)



Columns: SUM(Marketing Spen..)
Rows: SUM(Revenue)
Marks: Automatic
Detail: Store ID
Detail: New Expansion
Highlight New Expansion: New

Fig 4.4.4(a): Highlighting the new expansion stores of the Start-up

Clustering groups is an essential component of Analytics to help organizations identify the areas of potential improvement. To clearly demonstrate which stores (locations) are generating more revenue as the company spends on marketing, in the Analytics tab, Cluster option was chosen. This resulted into an automatic visualization as seen in Fig 4.4.4(b) where two cluster groups were formed. The cluster group 2 was coloured as orange and cluster group 1 as blue.



Columns: SUM(Marketing Spen..)
Rows: SUM(Revenue)
Marks: Automatic
Detail: Store ID
Detail: New Expansion
Colour: Clusters

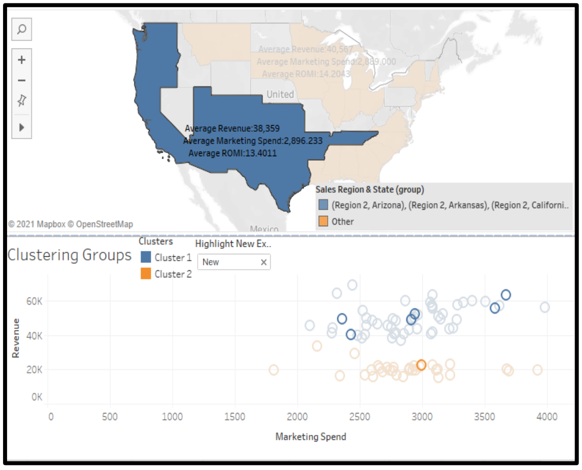
Clusters
- 🟦 Cluster 1
- 🟧 Cluster 2

Highlight New Expansion: Highlight New Expans...

Fig 4.4.4(b) Displaying the two clusters based on Marketing Spend and Revenue

Key findings: In Cluster 1, though the marketing spend remains between 2000-3500 range, the Revenue produced from each of these stores is also increasing whereas in the stores in Cluster 2, are producing less revenue with the marketing spend being the same as Cluster 1.

Therefore, it can be concluded that the company should invest more funds into the new stores in Cluster 1, as they are producing significantly high revenue. To further check which regions these clusters belong to, a dashboard was created (Fig 4.4.4(c)) and using the Regions as a filter, it was observed that Sales Region 2 has more number of new locations which produce high revenue utilizing less marketing funds in comparison to Sales Region 1. However, most of the old stores in Sales Region 1 belong to cluster 1 (the cluster comprised of high revenue stores with less marketing spend) as seen in Fig. 4.4.4(d)



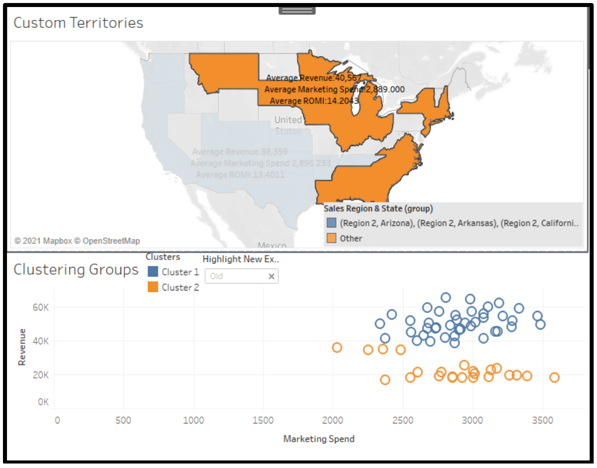
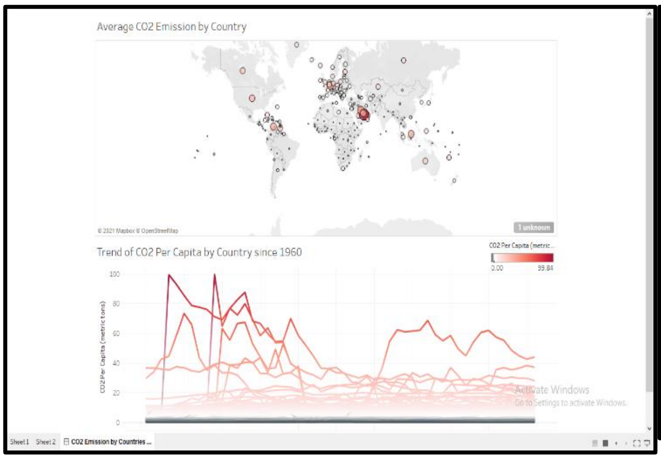


Fig 4.4.4(c) Custom Territories Dashboard showing the new stores in Sales Region 2

Fig 4.4.4(d) Custom Territories Dashboard showing the old stores in Sales Region 1

4.4.5 Show the trend of CO2 Emission by countries over the years? What is the trend of CO2 per capita emission level for USA over the years?

For this analysis, the dataset World_Bank_C02.xslx was loaded in Tableau and CO2 Data Cleaned worksheet was dropped in the data box. In the new sheet, 'Longitude' was added into columns and 'Latitude' into rows, Country Name' into Detail (within Marks box) and 'CO2 Per Capita' into Size, CO2 Per Capita' (from Measures) into Color. The measure of 'CO2 Per Capita' was changed from Sum to Average. Using Color formatting, Red-Black-White Diverging' was chosen from Palette Drop-down menu and was given 'Opacity' as 70% with border as 'Black' color.


Average CO2 Emission by Country
© 2021 Mapbox © OpenStreetMap
1 unknown
Trend of CO2 Per Capita by Country since 1960
CO2 Per Capita (metric..
0.00
99.84
100
80
60
40
20
0
CO2 Per Capita (metric tons)
Activate Windows
Go to Settings to activate Windows.
Sheet 1
Sheet 2
CO2 Emission by Countries ...

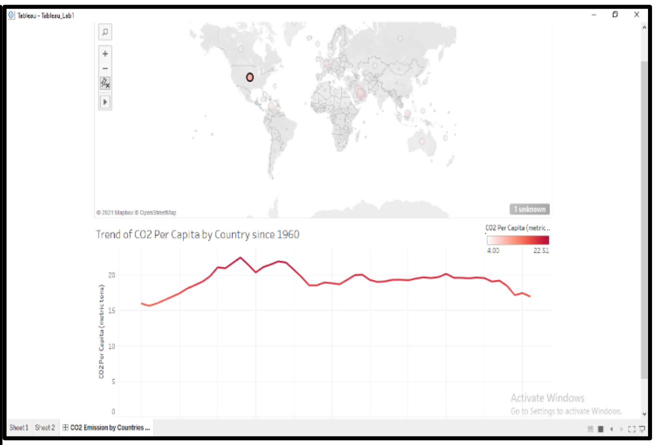


Tableau - Tableau_Lab1
© 2021 Mapbox © OpenStreetMap
1 unknown
Trend of CO2 Per Capita by Country since 1960
CO2 Per Capita (metric..
4.00
22.51
20
15
10
5
0
CO2 Per Capita (metric tons)
Activate Windows
Go to Settings to activate Windows.
Sheet 1
Sheet 2
CO2 Emission by Countries ...

Fig 4.4.5(a) Trend of CO2 Emission by countries over the years

Fig 4.4.5(b) Trend of CO2 emission level for USA from 1960

In Sheet 2, Year' was dropped into columns and 'CO2 Per Capita' into rows,'Country Name' into Detail (within Marks box) and 'CO2 Per Capita' into Color (within Marks box). CO2 Per Capita' was dropped into Filters as SUM. The color formatting was done same way as in Sheet 1. The two sheets were dropped in a dashboard to visualize the CO2 Emission by Countries since 1960 as seen in Fig 4.4.5(a). To see the trend of CO2 Emission for Canada, a filter was selected in Worksheet 1 of the dashboard to focus only on USA which resulted in Fig 4.4.5(b).

Key Findings: For most of the highly populated countries like India and China, the CO2 per capita emission levels has been increasing ever since from 1960. However, for USA the trend of CO2 per capita emission levels has been somewhat constant since 2000.

SECTION 4.5 ANALYSIS AND CRITIQUE OF THE TOOL

Having intermediate proficiency of Tableau, I could easily analyse the data, utilize advance features, and create dashboards from the two datasets. Tableau is very user-friendly and according to me the best data visualization tool I have worked with so far. Any changes made on an individual Sheet gets reflected in real-time on the story or dashboard in Tableau. Also, the speed of this tool can be highly appreciated. Datasets from multiple sources cab be integrated, cleaned, pivoted, and exported using automated features like joins, Data Interpreter, etc. The Analytics mode of Tableau can help users easily customise, summarise, and model the data. I am looking forward to work with the advanced/ professional version of Tableau which is inclusive of several advanced features and machine-learning capabilities. Also, I will be exploring the ability of this Tool to load and run Python scripts using TabPy API (Beran, 2016). Overall, Tableau is a powered solution designed to solve data problems in a fast and reliable manner.

SECTION 4.6 CONCLUSION

The functional purpose of Tableau can help analysts understand their business problems, and gain comprehensive knowledge of customer and product demographics, market trends also past-performance. This tool can be widely used within a corporation to share insights using story presentations and Tableau Server. Although, there are several BI self-service tools available in the market, I believe Tableau is the best one and can be utilized to build optimized BI solutions and would always prefer using it above SAP technologies.

CHAPTER 5 SAP ANALYTICS CLOUD

SECTION 5.1 INTRODUCTION TO THE TOOL

SAP Analytics Cloud supports advanced analytics and combines BI, augmented, and predictive analytics with machine-learning capabilities, along with useful planning features in one cloud environment (SAP Analytics Cloud, n.d.). This SAP technology can help improve business outcomes as it provides instant data insights, end-to-end user experience, and predictive features that enable data-driven decisions. As all the data is in Cloud, the security and access control features are very effective of this tool.

SECTION 5.2 DATASET DESCRIPTION

SAP Analytics Cloud will be used to analyze and visualize the dataset of a business simulation game called ERP Simulation Game (ERPSim). This ERPSIM.xlsx dataset was provided in class and contained information of the different types of Muesli products sold by the participant teams of the game. The data attributes Round and Day indicated the number of the round and the day of that round where using a specific Distribution channel and Sales Order number, the products were distributed. The measurable metrics of the dataset included Price, Quantity and Revenue. This tool used this dataset to measure each team's performance during and at the end of the game.

SECTION 5.3 BUSINESS PROBLEM AND RESEARCH QUESTIONS

5.3.1 – Business Use Case

Analysing business data and understanding the key performance indicators of an integrated business process is essential for an organization to keep track of their contribution margin, and market trends. To understand the market-demand of the Muesli product and how each team built their procuring and selling strategies during the game, the below research questions were used.

5.3.2 - Research questions

1. Which team had the overall highest and lowest revenue? Did both these teams sell the same products?
2. What is the trend of Revenue vs Quantity by Team and Product? What is the trend of Revenue per Round per team? How did the teams with the highest and lowest revenue manage their prices of 1 kg and 500 gm Strawberry Muesli throughout the rounds?
3. What is the trend of Revenue vs Product per team? What could have been different for the team with the overall lowest revenue?
4. Show the market share (in terms of quantity) of distribution channels per each product (with percentage %)? Are there any products that do not sell in specific distribution channels?
5. What are the key influencers of Revenue in this game?

SECTION 5.4 ANALYTICAL PROCESS AND KEY FINDINGS

5.4.1 Which team had the overall highest and lowest revenue? Did both these teams sell the same products?

In the “Create Story” mode of the tool, using the Access and Explore Data, the dataset, P1-ERPSIM.xlsx, was added. In the Data view of the tool, certain data enhancements were performed to refine the data for visualization purposes. Here, the price and Revenue measure units were changed to Euros. For the first part of the question, in the Story mode, a comparison Bar/Column chart structure was selected. Revenue was added to Measures and Team to Dimensions. This resulted in a chart as seen in Fig 5.4.1(a)

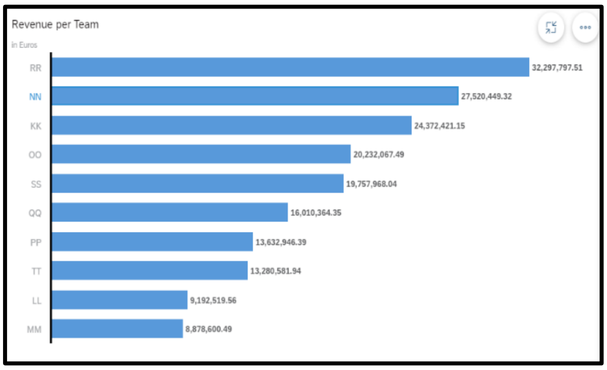
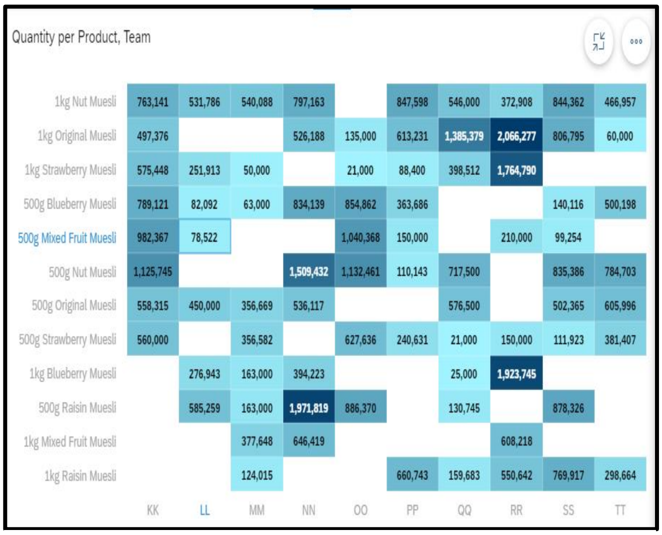


Fig 5.4.1(a): Bar Chart showing the overall Revenue per Team

For the second part of the question, a distribution Heat map was selected. Team was added to the X-Axis Dimensions and Product to the Y-Axis. Color was dropped in quantity which resulted in the quantity per product and Team illustration as seen in Fig. 3.4.1(b).

Key Findings: Team RR produced the highest revenue in the game whereas Team MM had the lowest revenue. Also, Team RR gained maximum revenue on selling 1 kg Muesli products and sold very less of 500g products. On the other hand, Team MM sold very less quantity of products in comparison to RR and the quantity of 500g Nut Muesli remained negligible for both the teams. Also, RR sold the highest quantity of 1 kg Original Muesli whereas MM did not sell any portions of this product. This could be a vital reason of Team MM coming last in the game as they did not sell the product which had sales demand.


Quantity per Product, Team

	KK	LL	MM	NN	OO	PP	QQ	RR	SS	TT
1kg Nut Muesli	763,141	531,786	540,088	797,163		847,598	546,000	372,908	844,362	466,957
1kg Original Muesli	497,376			526,188	135,000	613,231	1,385,379	2,066,277	806,795	60,000
1kg Strawberry Muesli	575,448	251,913	50,000		21,000	88,400	398,512	1,764,790		
500g Blueberry Muesli	789,121	82,092	63,000	834,139	854,862	363,686			140,116	500,198
500g Mixed Fruit Muesli	982,367	78,522			1,040,368	150,000		210,000	99,254	
500g Nut Muesli	1,125,745			1,509,432	1,132,461	110,143	717,500		835,386	784,703
500g Original Muesli	558,315	450,000	356,669	536,117			576,500		502,365	605,996
500g Strawberry Muesli	560,000		356,582		627,636	240,631	21,000	150,000	111,923	381,407
1kg Blueberry Muesli		276,943	163,000	394,223			25,000	1,923,745		
500g Raisin Muesli		585,259	163,000	1,971,819	886,370		130,745		878,326	
1kg Mixed Fruit Muesli			377,648	646,419				608,218		
1kg Raisin Muesli			124,015			660,743	159,683	550,642	769,917	298,664

Fig 5.4.1(b): Heat Map showing Sales quantity per product per Team.

5.4.2 What is the trend of Revenue vs Quantity by Team and Product? What is the trend of Revenue per Round per team? How did the teams with the highest and lowest revenue manage their prices of 1 kg and 500 gm Strawberry Muesli throughout the rounds?

In the story mode, a Bubble chart was selected. In the Design builder, Revenue was added to X-axis, Price to Y-axis, Quantity in Size, Team in Dimensions and Product in Color. Smart Grouping (groups of 3) feature was selected to form clusters which resulted in a visualization as seen in Fig. Also, another line chart was inserted in a new sheet to track the trend of Revenue of selected teams (top 2 and lowest 2 teams) per round. In the Design builder, Revenue was added to the left Y-axis, Round in Dimensions, Team in Color and filter was applied to select only Teams RR, LL, MM and NN which resulted in a visualization as seen in Fig 5.4.2(a). For the third part of the question, a Trellis was added to the second chart and product was added to it. Further, Revenue was replaced by price in the left Y-axis, Quantity was added to the Right Y-axis, the teams filter only included Team RR and MM, and also only the products which were commonly sold by both these teams – 1 kg Strawberry Muesli and 500 gm Strawberry Muesli — resulting into Fig 5.4.2(b)

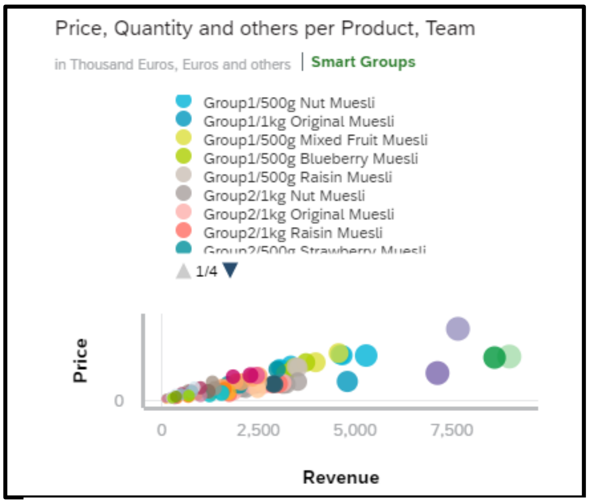


Fig 5.4.2(a): Bubble Chart showing correlation between price, quantity, and Revenue per Product per Team

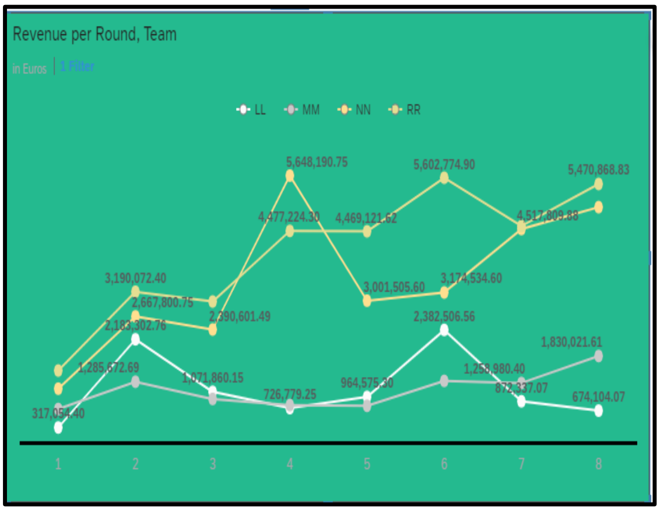


Fig 5.4.2(b): Line Chart showing the Revenue per round per Team for RR, LL, NN and MM

Key Findings: Team RR and NN made profits every round by selling products which were likely in demand (1 kg Muesli products) in huge quantities. On the other hand, there was not a significant rise in the revenue trend of Team MM and LL throughout all rounds as they mostly sold 500 gm products in quite less quantity. Also, on comparing the prices trend of Team RR and MM(Fig 5.4.2(c)), it is clearly visible that Team MM is increasing its price for the 1-6 rounds of 500g Strawberry Muesli although the quantity of sales keeps reducing whereas Team RR simultaneously manages the prices of the products as per the market demand.

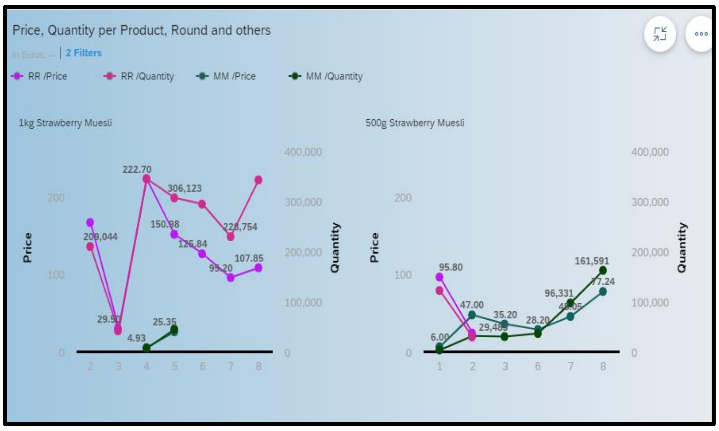
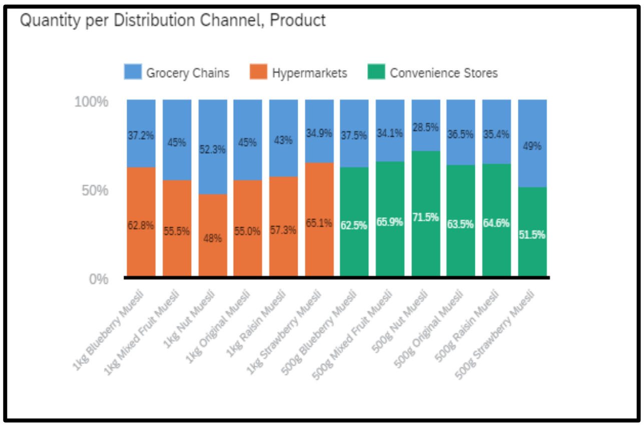


Fig 5.4.2(c) Price vs Quantity of 1kg and 500 gm Strawberry Muesli for team RR AND MM

5.4.3 Show the market share (in terms of quantity) of distribution channels per each product (with percentage %)? Are there any products that do not sell in specific distribution channels?

In the Design builder, a Stacked Bar/Column chart under 'Comparison' was chosen. MEASURES – Quantity, DIMENSIONS – Product, COLOR: Distribution channel, Under Chart Orientation, "Show Chart as 100%" was selected. This resulted into the visualization as seen in Fig 5.4.3(a). For the second part of the questions, a Heat map under 'Distribution' was chosen. DIMENSIONS (X) – Distribution Channel, DIMENSIONS (Y) – Product, COLOR – Revenue were added resulting into Fig 5.4.3(b).



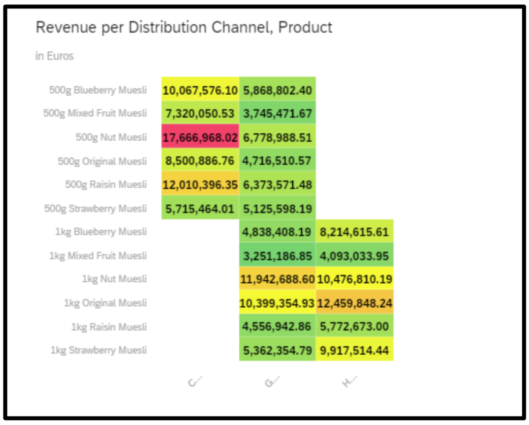


Fig 5.4.3(a): Stacked Bar Chart showing the quantity of each product per Distribution Channel.

Fig 5.4.3(b): Heat Map showing the Revenue per product on each Distribution Channel.

Key Findings: The distribution channel Convenience Stores is a major distributor for the 500 gm Muesli products whereas Hypermarkets form more than 50 percent of the distribution market for the 1kg Muesli products while Grocery chains remain a common distribution channel for all the products. All the 1kg products are not sold via Convenience Stores whereas all the 500 gm Muesli products do not get distributed via Hypermarkets.

5.4.4 What is the trend of Revenue vs Product per team? For the team with the highest and lowest revenue, what is the quantity of the highest Revenue product sold?

For this analysis, in the Design Builder, a Marimekko map was chosen. The measures, Revenue and Quantity were respectively dropped in Height and Width. Product was added to dimensions, and Team in Color. This resulted into a visualization as seen in Fig 5.4.4(a)

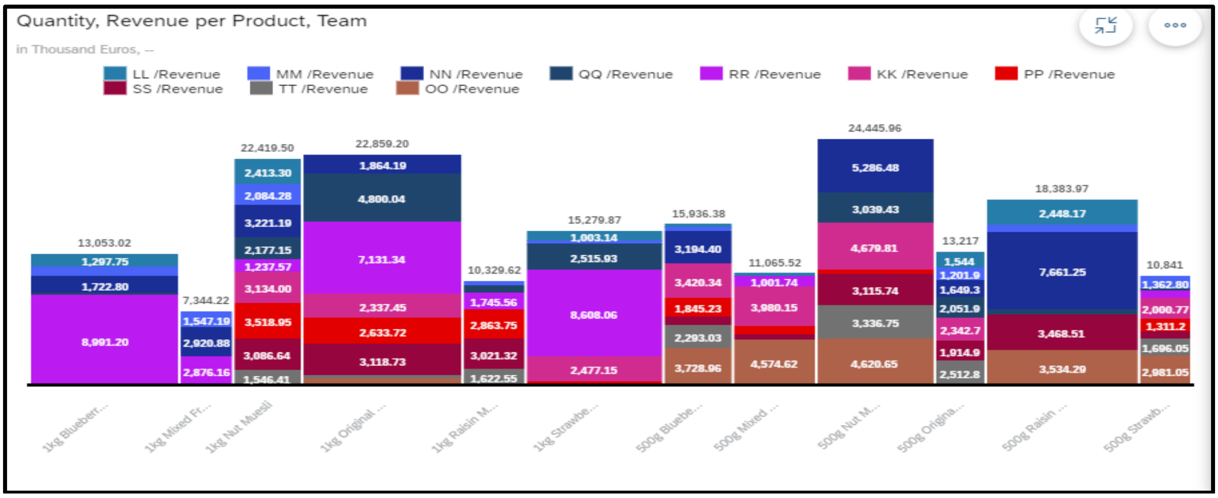


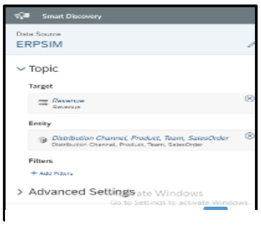
Fig 5.4.4(a) Marimekko map showing revenue of each product per Team.

Key Findings: It has been observed in 5.4.1 that Team RR (team with highest Revenue) does not sell any portions of 500 gm Nut Muesli which is the product that generates the highest revenue in this game. However, Team NN (team with second highest Revenue) sells this 500 gm product in huge quantities. Team MM (team with lowest Revenue) did not sell any quantities of 500 gm product or 1 kg Original Muesli (the products having highest Revenue). This indicates that Team MM could have changed the prices of these products accordingly to maximise profits as the game progressed. After solving all the above questions, it has been observed that the following measure could have been adopted by TEAM MM during the game to maximise profits.

- Team MM should have analyzed the sales report of their past rounds and understand the market demand of the products they were procuring and selling. As, they did not procure and sell 500 gm Nut Muesli or 1 kg Original Muesli – highest revenue products.
- Team MM did not study the quantity of the products being sold by them during the game and did not make price changes accordingly. If they would have strategically planned their selling strategy and predicted outcomes based on past performance, their overall revenue could have been improved.
- Team MM could have analysed the distribution channels as per product type.

5.4.5 What are the key influencers of Revenue in this game?

For this analysis, the ***Smart Discovery*** feature under the section "More" section of the tool was selected. In the settings, Revenue was selected as Measure, the advanced options were kept as default and Entity included – Product, SalesOrder, Distribution Channel and Team as seen in Fig5.4.5(a). The Run button was clicked — Overview of Revenue, Key Influencers of Revenue (Fig 5.4.5(b)), Unexpected Values and Simulation reports.


Smart Discovery

Data Source: ERPSIM

Topic

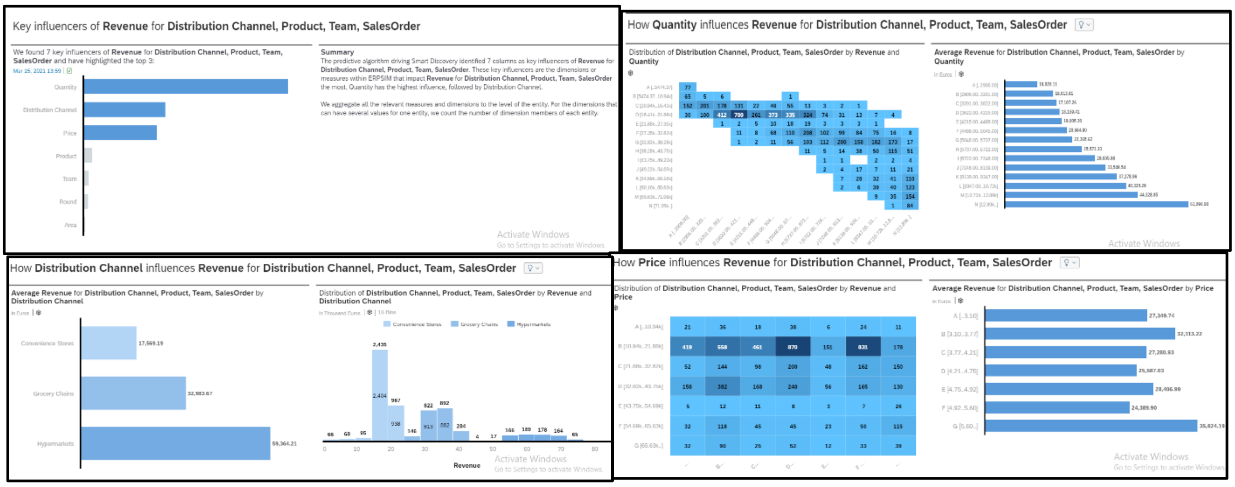
Target: Revenue (Revenue)

Entity: Distribution Channel, Product, Team, SalesOrder (Distribution Channel, Product, Team, SalesOrder)

Filters: + Add Filters

Advanced Settings

Fig 5.4.5(a)Smart Discovery Settings


Key influencers of **Revenue** for **Distribution Channel, Product, Team, SalesOrder**

We found 7 key influencers of **Revenue** for **Distribution Channel, Product, Team, SalesOrder** and have highlighted the top 3:

Mar 15, 2021 13:59

Summary

The predictive algorithm driving Smart Discovery identified 7 columns as key influencers of **Revenue** for **Distribution Channel, Product, Team, SalesOrder**. These key influencers are the dimensions or measures within ERPSIM that impact **Revenue** for **Distribution Channel, Product, Team, SalesOrder** the most. Quantity has the highest influence, followed by Distribution Channel.

We aggregate all the relevant measures and dimensions to the level of the entity. For the dimensions that can have several values for one entity, we count the number of dimension members of each entity.

How **Quantity** influences **Revenue** for **Distribution Channel, Product, Team, SalesOrder**

Distribution of **Distribution Channel, Product, Team, SalesOrder** by **Revenue** and **Quantity**

Average **Revenue** for **Distribution Channel, Product, Team, SalesOrder** by **Quantity**

How **Distribution Channel** influences **Revenue** for **Distribution Channel, Product, Team, SalesOrder**

Average **Revenue** for **Distribution Channel, Product, Team, SalesOrder** by **Distribution Channel**

Distribution of **Distribution Channel, Product, Team, SalesOrder** by **Revenue** and **Distribution Channel**

How **Price** influences **Revenue** for **Distribution Channel, Product, Team, SalesOrder**

Distribution of **Distribution Channel, Product, Team, SalesOrder** by **Revenue** and **Price**

A [..10.94k]	21	36	18	38	6	24	11
B [10.94k..21.88k]	419	558	461	870	151	831	178
C [21.88k..32.82k]	52	144	98	200	48	162	150
D [32.82k..43.75k]	158	382	168	240	56	165	130
E [43.75k..54.69k]	5	12	11	8	3	7	28
F [54.69k..65.63k]	32	118	45	45	23	50	115
G [65.63k..]	32	90	25	52	12	33	39

Average **Revenue** for **Distribution Channel, Product, Team, SalesOrder** by **Price**

A [..3.10]	27,349.74
B [3.10..3.77]	32,113.22
C [3.77..4.21]	27,280.93
D [4.21..4.75]	25,587.03
E [4.75..4.92]	28,406.99
F [4.92..5.60]	24,389.90
G [5.60..]	35,824.19

Figure 5.4.5(b) Key Influencers of Revenue

Key Findings: As seen, in the analysis of the 5.4.1 to 5.4.4 questions, Sales quantity and Price plays an important role for a Team aspiring high Revenue in this game. The predictive analysis report in Fig generated by SAC using automated machine-learning algorithms also shows that the attributes Quantity, Distribution Channel and Price are the key influencers of Revenue for this dataset.

SECTION 5.5 ANALYSIS AND CRITIQUE OF THE TOOL

As a first-time user of SAP Analytics Cloud, I found this tool very convenient. Firstly, it did not cause the additional trouble of downloading this software unlike the other SAP products. Having the required access to the trial version of SAP Analytics Cloud, I could explore the story creating feature of this tool which was quite amazing. I liked that using the tool I could leverage machine-learning insights on the dataset. This can be effective for business managers of any enterprise who do not have significant technical expertise. Also, Cloud Analytics has several advantages over on-premises BI tools as cloud computing allows users to easily consolidate information from all sources, re-establish online connections and deliver reports at a much faster speed (Idexcel Technologies, 2017).

The prime reasons why cloud analytics is widely adopted nowadays are flexible cloud storage capacity, high performance and scalability, real-time analytics, cross-organizational analysis, robust security features and low maintenance costs (Weaver, 2017).

Although the tool is quite user-friendly, I found the design layout of the story mode somewhat confusing. For example, the Insert and Designer button are related, and it would take some time for first time users to understand that on clicking the Designer mode, a chart needs to be inserted using the Insert option. Unlike Tableau and SAP Predictive Analytics, the predictive analysis feature is somewhat hidden in the More section of this tool, which should be prominently displayed on the UI for faster-access. Also, as the tool is cloud-based, sometimes its performance is affected by poor internet connections. However, I personally liked this SAP product the most and looking forward to exploring its supported advanced features like creating analytic application and integration with SAP HANA for easy data amalgamation.

SECTION 5.6 CONCLUSION

SAP Analytics Cloud has given SAP technologies a new uplift by being cloud-driven as it increases the scope of working with large amounts of data in a secured manner, allowing co-works share a collaborative data environment and businesses to leverage AI and ML capabilities.

CHAPTER 6 : SAP PREDICTIVE ANALYSIS FOR DATA MINING

SECTION 6.1 - INTRODUCTION

Data Mining is the process of knowledge discovery by analyzing large portions of data to obtain insights like data patterns, relationships, and trends (Kale & Jones, 2020). SAP Predictive Analytics is designed to support statistical analysis, data mining and build predictive models (SAP Predictive Analytics, 2016). SAP PA when configured using R is capacitated to perform both descriptive and predictive analysis. With its automated data mining competency, this tool leverages the usefulness of supervised and unsupervised ML algorithms, making it suitable for managers and data analysts to make informed decisions. For this analysis, the desktop version (2-tired architecture), Expert Analytics view of SAP PA was used.

SECTION 6.2 - DATASET DESCRIPTION

SAP Predictive Analytics will be used to perform data mining techniques, Clustering, Regression and Forecasting, on the below datasets which were provided in class.

Stores.csv – Contains information of a retail chain having 150 stores located in several regions of North America. The dataset includes four measurable attributes, sales turnover, staff size, store size and profit margin that were utilized to segment stores into clusters for developing promotion strategies.

GB_AnalyticsData.csv – Contains information about the sales data for GBI in two different currencies (Euros and USD) for its several products, product categories, multiple sales organizations, and customers in two countries— USA and Germany over the years (2007-2019). Several dimensions were used as independent variables to understand their relationship with the target measure and predict sales for GBI. Regression analysis was used for this estimation process.

GBSales_transactions.xlsx — This dataset contained information about of the revenue of GBI demonstrated in two currencies (EUR and Dollars) for different quarters and years (2007 to 2018). This dataset was used to forecast sales revenue for GBI based on past performance.

SECTION 6.3 BUSINESS PROBLEM AND RESEARCH QUESTIONS

6.3.1 – Business Use Case- For a store manager, having stores in different locations, it is important to group stores based on profitability and sales performance metrics to identify the unique behaviour of each group and develop relevant promotion strategies. Therefore, using K-Means Clustering algorithm of Predictive Analytics on the dataset, Stores, the following question was tackled.

Research questions:

1. What kind of the marketing strategy can be used to improve the sales for the stores in a specific cluster?

6.3.2 – Business Use Case - Estimation models can help organizations to determine outcomes based on multiple variables. Global_Bike intends to use a budgeting model to predict sales of a particular product to be always stocked with the in-demand products and maximise the financial indicators. This estimation is carried out using regression analysis to answer the below question.

<u>Research question</u>*:*

1. ***1.*** What is the predicted value of sales quantity of a particular product essential to plan production and purchases for GBI?

<u>6.3.3 – Business Use Case</u> – Forecasting is used to estimate the value of a variable in the future using time-series analysis. Global Bike intends to forecast revenue in the next year based on past performance. This will be done using Triple exponential smoothing and derive valuable insight for the following question.

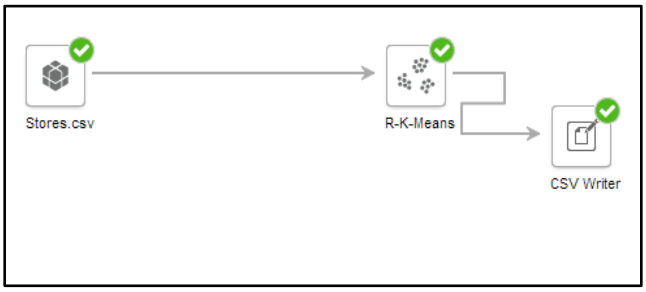
<u>Research question</u>*:*

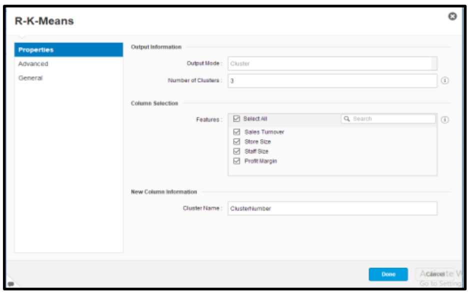
1. What is the trend of the forecasted sales revenue of GBI for the year 2020?

SECTION 6.4 ANALYTICAL PROCESS AND KEY FINDINGS

6.4.1 What kind of the marketing strategy can be used to improve the sales for the stores in a specific cluster?

For this question, the dataset — **Stores.csv** — was loaded and visualized using the Expert Analytics PA tool. To capacitate the tool with machine learning abilities, R was installed and configured in the tool. In the prepare view, it was observed that the dataset had the measurable attributes like Profit Margin, Sales Turnover, Staff Size and Store size. In the Predict view where the loaded dataset was already added as the data source, within the component panel R-K means algorithm was dragged and dropped as well as the CSV Writer component from the Data Writers tab as seen in Fig 6.4.1(a). In the configuration settings of the R-K means algorithm, the number of clusters were provided 3, all the four measures were chosen for cluster analysis and the default values of the advanced properties were retained (Fig 6.4.1(b)).




R-K-Means

Properties
Advanced
General

Output Information

Output Mode : Cluster

Number of Clusters : 3

Column Selection

Features : ☑ Select All
☑ Sales Turnover
☑ Store Size
☑ Staff Size
☑ Profit Margin

New Column Information

Cluster Name : ClusterNumber

Fig 6.4.1(a): Predict panel including the R-K Means algorithm

Fig6.4.1(b): R-K Means Configuration settings

In the configuration settings of the CSV Writer, a CSV file name was used to store the clustering results. The algorithm was then run which yielded summary results and cluster representations as seen in Fig 6.4.1(c)

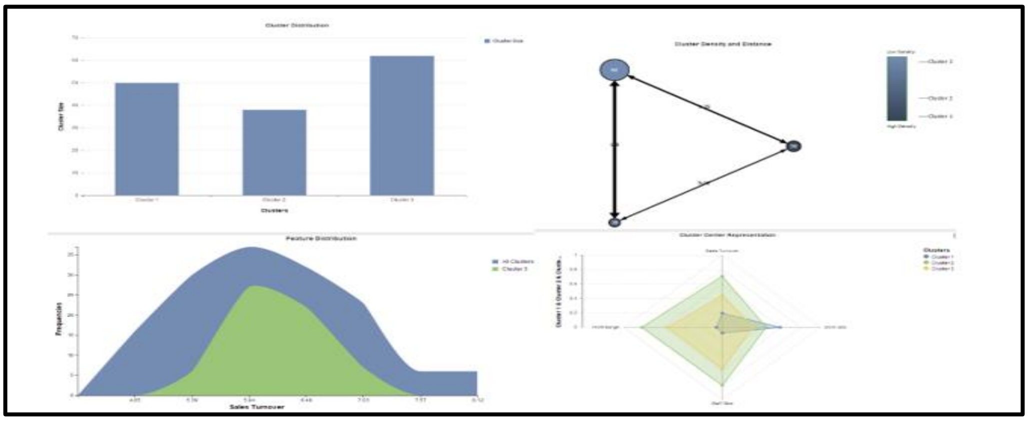
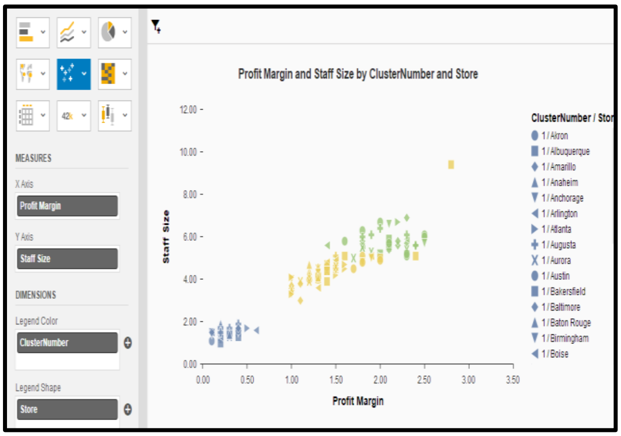


Fig 6.4.1(c)Cluster Representation of Stores



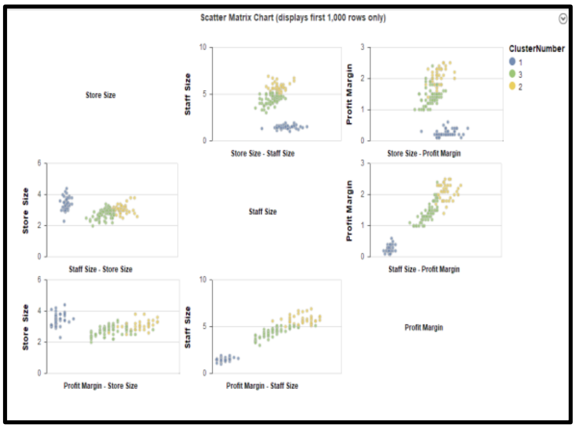
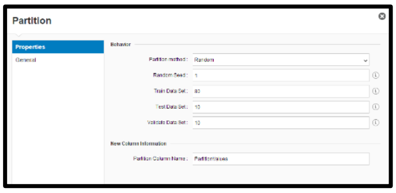


Fig6.4.1(d)Profit Margin and Staff Size by Clustering Visualization **Fig6.4.1(e) Scatter Matrix chart**

Key Findings: Cluster1 which has very similar components (highest density) has limited number of staff in the respective stores and a low profit margin – ***Low profit+ less staff***. Most of the stores in this cluster show similar trend with profit values lying with a specific range. It is evident that cluster 3 has the maximum number (62) of stores with the lowest density – loosely associated data elements as compared to the other two clusters. In this cluster, the profit margin characteristics of the store widely differ from each other, lying between (1 to 3.50 scale of measure), while the staff size predominantly lying between 2 to 5- ***Medium profit + medium range of staff***. Furthermore, it is observed that as the staff size of the stores increases, the sales turnover and profit margin also increase. Therefore, in Cluster 2, the profit margin is high as the staff size is more- ***High Profit + More staff***. Most of the components show similar traits while few stores fall out of the range and thereby this cluster has medium density. The store manager can use the results of this clustering analysis to increase the staff members in cluster 1 stores, to see if the sales turnover is becoming better on. Also, the store manager can increase the staff size of the stores in cluster 2 and 3 with **high revenue & less staff** to maximise store profits.

6.4.2 How does the predicted value of sales quantity differ from the actual values? production and purchases for GBI?

For this predictive analysis, the dataset – GB_AnalyticsData.xslx is loaded on PA. In order to predict the sales quantity for all the products of GBI, a regression analysis model was built in the Predict mode of the tool. Firstly, from the Preprocessors panel, the partition method was dropped to the canvas and configured to split the dataset into training and test data (Fig 6.4.2(a)). AutoRegression algorithm was dropped on the canvas and configured to predict the target variable – sales quantity against the multiple independent variables (Month, CustDescr, City, SalesOrg, ProdDescr, CatDescr). The model was then run to produce regression analysis results(Fig 6.4.2(b)). In the grid view of the results section, there was an additional column for the Predicted Sales quantity for each transaction. The model representation displayed the various contributions of the dependent variables on the target quantity. (Fig 6.4.2(c)).


Partition

Properties | General

Behavior

Partition method : Random

Random Seed : 1

Train Data Set : 80

Test Data Set : 10

Validate Data Set : 10

New Column Information

Partition Column Name : PartitionValues

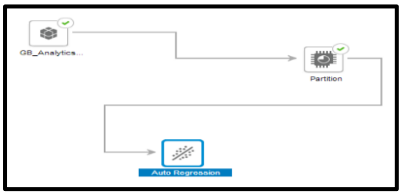


Fig6.4.2(a) Partition Configuration Setting **Fig 6.4.2(b) Performing AutoRegression on the dataset**

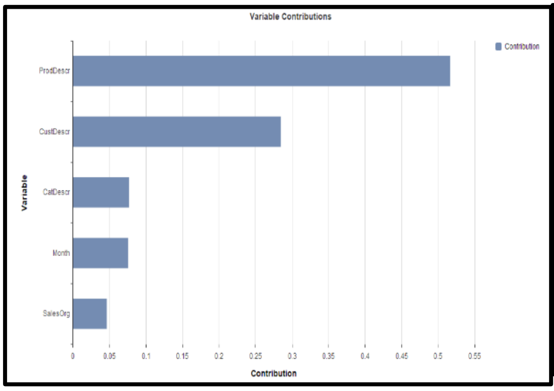


Fig 6.4.2(c)Variable distributions to predicted sales quantity.

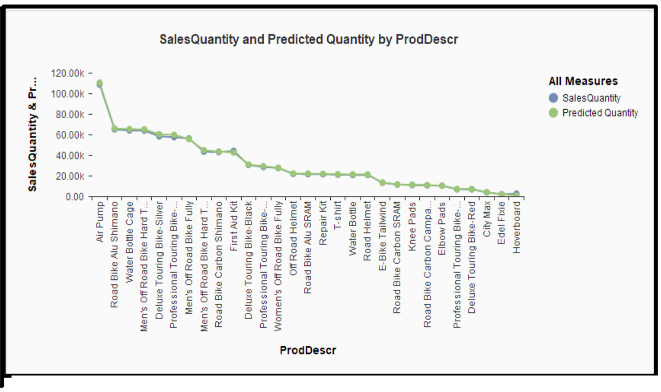


Fig 6.4.2(d): Actual Sales quantity vs predict sales quantity per product

The predicted sales quantity was plotted against the actual sales quantity for the different products of GBI using a line chart in the Visualize mode of PA which resulted in Fig 6.4.2(e).

Key Findings: As, we know the closer the predicted sales quantities are to the actual sale quantities, the more accurate the regression model is. It is clearly observed in Fig 6.4.2(e), there is very less difference between the actual and predicted values of sales quantity making this analysis quite effective. As per distribution analysis, Product highly effects the sales quantity of GBI followed by Customer. Therefore, now GBI managers can utilize the predicted sales value of each product and plan the production and purchases of products accordingly.

6.4.3 What is the trend of the forecasted sales revenue of GBI for the years 2020 and 2021?

Forecasting is used to unravel variations, trends, and seasonality to help predict future values of a variable. Time series analysis is a technique using which data can be modelled to make forecasts. To further decompose a time series into its components – trends, seasonality, cycles and randomness, mathematical methods called ***Exponential smoothing*** was used for this question. For this, the dataset- **GBSales_transactions.xlsx**- was loaded in the predictive analytics tool. In the Prepare mode, a new calculated measure called Revenue in USD was created to convert all the Revenue data values in Dollars from Euros using the formula ***Revenue in USD = if{CURRENCY} = "EUR" then {REVENUE}/1.2 else {REVENUE}.*** In order to evaluate the seasonality of GBI sales revenue, a time hierarchy was created out of YEAR dimension having MONTH(Day or Month) and then the updated dataset was illustrated in the Visualize mode using a time-series line chart to visualize the revenue trend. (Fig 6.4.3(a)). As observed in the Fig 6.4.3(a), the revenue trend of GBI has both trends and seasonality, therefore triple exponential

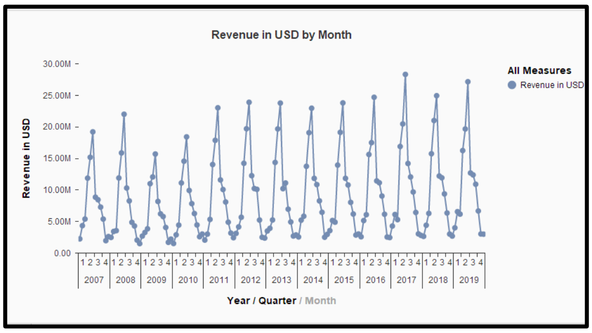


Fig 6.4.3(a): Revenue in USD per Month

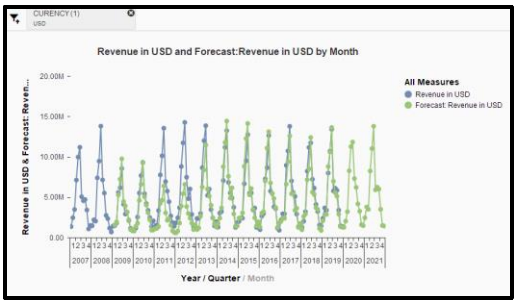
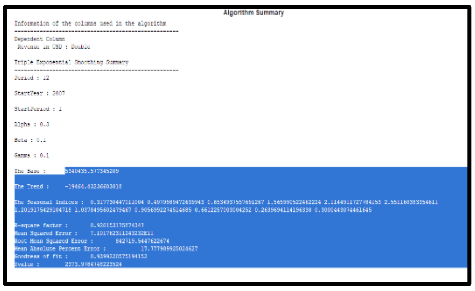


Fig 6.4.3(b): Forecasted Sales Revenue per month for the years 2020 and 2021.

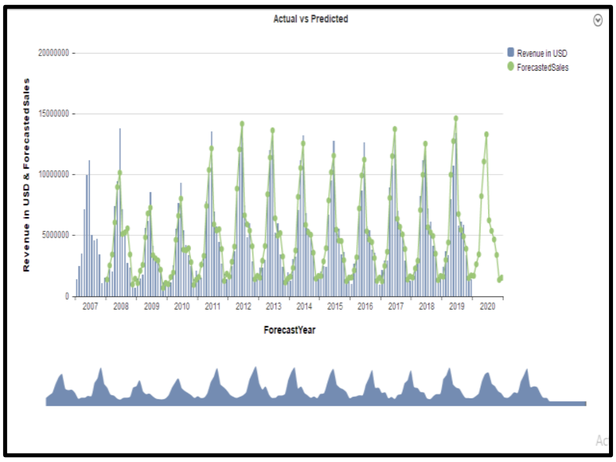
smoothing was used for forecasting sales revenue. So the forecast was added to the visualization by selecting Pre dictive Calculation -> Forecast with SAP Predictive Analytics for 24 months resulting into the forecasted visualization (Fig 6.4.3(b)) of GBI sales revenue for the year 2020 &2021. The results were filtered only for US and the results were compared with the analysis done in Predict mode using Triple Exponential smoothing algorithm. The algorithm parameters' (alpha, beta and gamma) values were retained and on running the algorithm, summary (Fig 6.4.3(c)) and Trend Chart(Fig 6.4.3(d)) were produced. To improve the fitting capabilities of the forecast, the values alpha =0.5, Beta = 0.5 and Gamma = 0.5 values were changed and the year period was extended to 24 months which resulted in the trend chart as seen in Fig 6.4.3(e).



Algorithm Summary

```
Information of the columns used in the algorithm
--------------------------------------------------
Dependent Column
 Revenue in USD : Double

Triple Exponential Smoothing Summary
--------------------------------------------------
Period : 12

StartYear : 2007

StartPeriod : 1

Alpha : 0.3

Beta : 0.1

Gamma : 0.1

The Base :      5340435.577345209

The Trend :     -19464.43236603018

The Seasonal Indices :  0.317730447011004 0.4979989472639943 0.6534937557691267 1.565990522462224 2.114491172778415 2.551180383354611
1.2019175429104718 1.0378495602479467 0.9056992274514685 0.6612257008004252 0.2639694114196338 0.3000443074461645

R-square factor :       0.920153175874347
Mean Squared Error :    7.1017623112432321E11
Root Mean Squared Error :       842719.5447622674
Mean Absolute Percent Error :           17.777909925024627
Goodness of fit :       0.9399220575194152
Fvalue :        2373.9784748223524
```

Fig 6.4.2(c): Summary of Regression Analysis



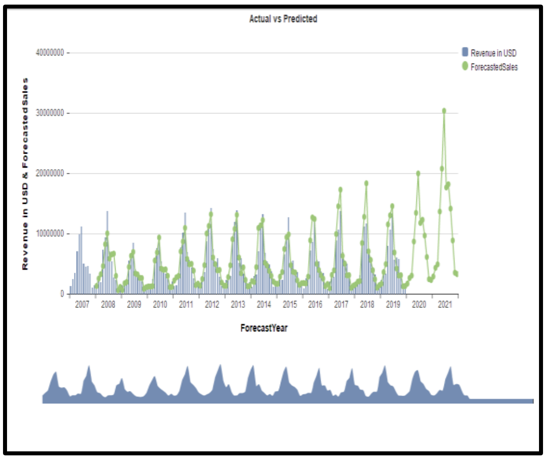


Fig 6.4.3(d) Forecasted Sales Revenue of GBI 12 months

Fig 6.4.3(e)Forecasted Sales Revenue for 24 months

Key Findings: The forecast done in the visualize mode, does not capture the trend and seasonality forecasts quite well for the years 2011 and 2012. However, it shows the forecasted sales revenue around 11.8 and 13.8 Million for the years 2020 and 2021 during the seasonal month of June. This diagnosis is done internally by PA which makes it difficult to understand the default values of alpha, beta, and gamma for this analysis. On the other hand, on analyzing the predict mode Triple Exponential Smoothing algorithm with default data, trend and seasonal smoothing parameters, result into a column and line chart showing the historical sales revenue, fitted curve of revenue, and forecast for revenue for 12 months (Fig 6.4.3(d)). The forecast captures the trend and seasonality quite well with a R-square value of 0.92 indicating a well-fitting model (R-square close to 1). However, to reduce the mean-squared error from 7.81, and to create a forecast based on recent historical data points, the smoothing parameter values were increased to 0.5 and the forecast period was extended to 24 months. As observed in Fig 6.4.3(e), the forecasted sales revenue are higher for the years 2020 and 2021 as they are forecasted mainly based on the last years'(2019 &2018) sales revenue trends. Also, the R-square factor for this setting is 1.18 with MSE lying at 1.81. Therefore, based on the requirements of forecast, early or later past performance relative, the GBI managers can forecast the sales revenue for the years 2020 and 2021 and accordingly plan their inventory control and purchases. June is the constant seasonal month for sales and

GBI management can utilize the sales revenue forecast results to stock the products in accordance to seasonality and trends forecast of 2020 & 2021.

SECTION 6.5 ANALYSIS AND CRITIQUE OF THE TOOL

As I had previously used SAP Predictive Analytics for data visualizations (Chapter 3), I was quite comfortable working with the tool this time. I had to install and configure R on SAP PA to leverage its machine learning capabilities. Learning how to configure and run the descriptive and predictive algorithms in PA is not much of a challenge. However, one needs to know the data mining concepts to select the algorithms based on the business requirements and derive valuable insights from the trend charts. But a data specialist, with intermediate knowledge of data mining can easily choose the suitable model in the Predict mode of PA based on the requirements, validate its working, monitor its functionality, and use it for business improvement.

This tool can be used as a competitive advantage for analysts because of its predictive modeling environment that automates the development of robust data-mining functions operating on multiple input attributes (Hart,2015). In this data-driven economy, where data must be integrated from various sources and processed through several algorithms for knowledge discovery, this tool can make the life of analysts simpler. Being a data scientist myself, I have used mostly python programming tool and coded ML algorithms from the scratch to explore enormous data corpus. Working on SAP PA helps accomplish most data pre-processing tasks without recourse to code and thereby increasing productivity (Hart,2015). This tool is suitable for advanced customer-focused analytics and can even be used by managers with less-technical proficiency. Expert Analytics enables you produce deep analysis of the data using different visualization techniques, such as scatter matrix charts, parallel coordinates, cluster charts, and decision trees. It also enables you to perform various analyses and build models on the data, including time series forecasting, outlier detection, trend analysis, classification analysis, segmentation analysis, and affinity analysis. Use a range of predictive algorithms, the R open-source statistical analysis language, and in-memory data mining capabilities for handling large volume data analysis efficiently.

However, after working with the desktop version of PA, I am keen to explore the three-tiered client architecture version of PA designed for enterprise level of analysis (SAP Predictive Analytics Enterprise Edition, n.d.)

SECTION 6.6 CONCLUSION

SAP Predictive Analytics is no doubt efficient in data preparation, prediction, and visualization. With real-time insights offered by this tool, organizations can effectively examine customer behavior for profitable results, gain better understanding of the business and support reliable decision-making (SAP PA, n.d.), making it one of the effective BI- tools. The automated machine learning algorithms feature of this tool fascinate me, and I am determined to use SAP Predictive Analytics in the future.

CHAPTER 7 SAP HANA

SECTION 7.1 INTRODUCTION TO THE TOOL

SAP HANA is an in-memory database (IMDB) with database administration, management, security, multi-model processing, application development and data virtualization capabilities (SAP HANA, n.d.). Being a cloud-based data foundation of SAP, it integrates data from across the enterprise, enabling faster decision-making on live data. SAP HANA provides columnar data storage, simpler indexing, improved data compression, data partitioning, parallel data processing and supports both transactional and analytical data processing – OLTP+OLAP (Kale & Jones, 2020). SAP HANA permits horizontal scaling of the database and provides true real-time analytics at unprecedented speed by merging all transaction processing and reporting into one database. The two perspectives of SAP HANA Database Server were used for this analysis – HANA CATALOG (includes database schema, physical and virtual tables, flat files & data provisioning menu), HANA EDITOR where the database models, information models(views) database flowgraphs, etc were created.

SECTION 7.2 DATASET DESCRIPTION

SAP HANA was used to create the data model for the following datasets. These datasets called as flat files were already present in the SAP HANA server. SAP Predictive Analytics was used to connect to SAP HANA for creating visualizations from the information model - SALES_CUBE view which was created and populated in SAP HANA.

Customers.csv: This dataset contains information of the customers of GBI and includes 7 data columns– Customer_Number, Customer_Name, City, Valid_to, Valid_from, Sales_Organization and Country.

Product.csv: This dataset contains information of the different products, product categories and groups of GBI and their corresponding prices.

SalesTransaction.csv: This dataset contains information of the sales data of GBI including 3 important measurable attributes (Revenue, Discount and Sales quantity). This file contained the customer number and product number using which the sales data was merged with the product and customers data in SAP HANA.

SECTION 7.3 BUSINESS PROBLEM AND RESEARCH QUESTIONS

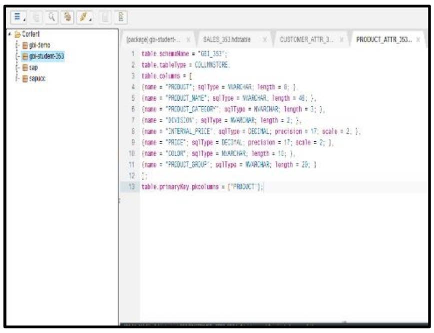
7.3.1 – Business Use Case: GBI intends to build a real-time computing business data platform using an in-memory data warehouse solution – SAP HANA. It is important to build an underlying data model for a small subset of data to evaluate and test the reporting capabilities of the created model. Therefore, this business scenario would be dealing with the sales data of GBI to help the organization build a suitable data model and information data models (views) implementing columnar data storage for fast retrieval and improved response time. SAP PA would be later used to demonstrate the data analysis, exploration and visualizing capabilities provided by the powered solution.

7.3.2 – Research questions

1. What is the trend of net revenue and sales quantity country wise for GBI over the years?
2. Which product category is the highest revenue generator of GBI?
3. Who is the star customer of GBI?

SECTION 7.4: ANALYTICAL PROCESS AND KEY FINDINGS

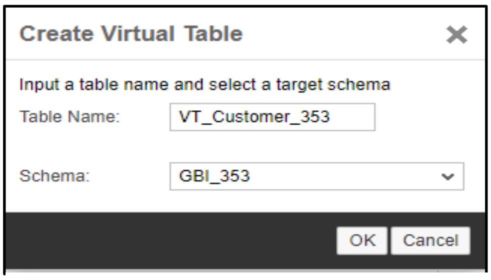
7.4.1. What is the trend of net revenue and sales quantity country wise for GBI over the years?



```
table.schemaName = "GBI_353";
table.tableType = COLUMNSTORE;
table.columns = [
{name = "PRODUCT"; sqlType = NVARCHAR; length = 8; },
{name = "PRODUCT_NAME"; sqlType = NVARCHAR; length = 40; },
{name = "PRODUCT_CATEGORY"; sqlType = NVARCHAR; length = 3; },
{name = "DIVISION"; sqlType = NVARCHAR; length = 2; },
{name = "INTERNAL_PRICE"; sqlType = DECIMAL; precision = 17; scale = 2; },
{name = "PRICE"; sqlType = DECIMAL; precision = 17; scale = 2; },
{name = "COLOR"; sqlType = NVARCHAR; length = 10; },
{name = "PRODUCT_GROUP"; sqlType = NVARCHAR; length = 20; }
];
table.primaryKey.pkcolumns = ["PRODUCT"];
```

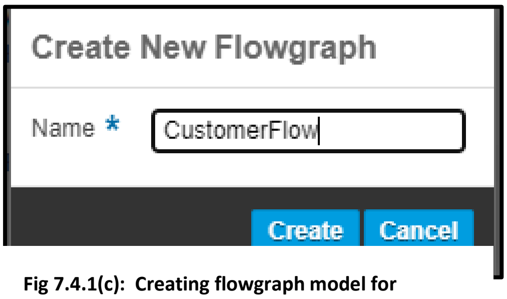
Fig 7.4.1(a): Creating database definition for product master data- *PRODUCT_ATTR_353.hdbtable*

The first step was to create database tables in the SAP HANA Web-based Development Workbench, which laid out the data foundation for the information models (views) to be created at a later stage. As indicated in the business scenario, three tables were created, one for sales transaction data, one for the customer master data, and one table for product master data. In SAP HANA Editor, there was already a student package created. In the package, a new table definition for product- ***PRODUCT_ATTR_353.hdbtable*** was created using the context menu. Once the file was created, certain codes were used to build the database schemas and columns as seen in Fig 7.4.1(a). The same was repeated for the customer and sales transaction data. On viewing the SAP HANA CATALOG, under the database schema, GBI_353, all the three tables were found. The next step was data provisioning – to populate the created data models using Smart Data Integration Tool. For each table, a data flow was created that extracts data from the flat files and transfers it to the established database tables. To do this, in SAP HANA Catalog, under, provisioning, new remote source was created. This source had all the flat data files. However, the data is not directly migrated from these files to the created data definitions. There is an intermediate layer of connectivity called virtual tables. Therefore, virtual tables were created from each of the csv files (Fig 7.4.1(b)). As seen in Fig 7.4.1(c), in the Editor perspective of SAP HANA Web Workbench, under the gbi_student_353 package, a new flowgraph model was created for Customer – ***CustomerFlow***. Within the flowgraph model, establish a data connection between the virtual table *VT_Customer_353.csv* and the data table gbi-student-353_*Customer_Attr_353* by defining the virtual table as *data source* and the data table as *data sink*. Afterwards, field mapping between the virtual table and data table was performed as seen in Fig 7.4.1(d).



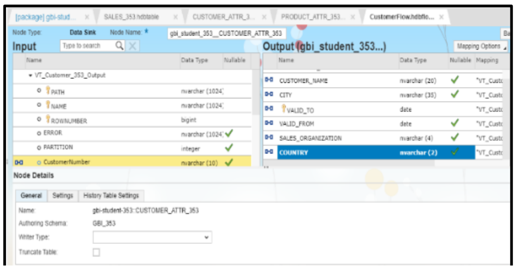
Input a table name and select a target schema
Table Name: VT_Customer_353
Schema: GBI_353

Fig 7.4.1(b): Creating virtual table from Customer.csv master data



Name * CustomerFlow

Fig 7.4.1(c): Creating flowgraph model for Customer table



Node Type: Data Sink Node Name: * gbi_student_353__CUSTOMER_ATTR_353

Name	Data Type	Nullable
VT_Customer_353_Output		
PATH	nvarchar (1024)	
NAME	nvarchar (1024)	
ROWNUMBER	bigint	
ERROR	nvarchar (1024)	✓
PARTITION	integer	✓
CustomerNumber	nvarchar (10)	✓

Name	Data Type	Nullable	Mapping
CUSTOMER_NAME	nvarchar (20)	✓	"VT_Custo
CITY	nvarchar (35)	✓	"VT_Custo
VALID_TO	date		"VT_Custo
VALID_FROM	date	✓	"VT_Custo
SALES_ORGANIZATION	nvarchar (4)	✓	"VT_Custo
COUNTRY	nvarchar (2)	✓	"VT_Custo

Name: gbi-student-353::CUSTOMER_ATTR_353
Authoring Schema: GBI_353
Writer Type:
Truncate Table: □

Fig 7.4.1(d): Field mapping between the customer virtual table and data table.

The flowgraph model was executed, and after checking the status of the data load in the monitoring console, the content of the table was seen in the Catalog perspective.The same process was repeated to create flowgraph models for the product and sales data and then field mapping of the data attributes was carried out.Since, the goal of our database model was to analyze the sales data by customer and product, a simple star schema (Fig 7.4.1(e) with the two information data tables (views) – Customer data and Product data views was built. The two views are dimensions in the schema. Therefore, a customer calculation view name - CUSTOMER_DIM_CV_353 was created for customer data. In the view workspace, a new projection was added which had the earlier created data table- CUSTOMER_ATTR_353 as its data source, and the fields were mapped as output columns as seen in Fig 7.4.1(f). As the view should only hold the information of the current customers of GBI, a filter was applied to the view ("VALID_TO"= 99991231). Since, the customer data table did not contain information of the country description and the sales organization description, the respective data tables - **GBI_DEMO_COUNTRY & GBI_DEMO_SALESORG** were joined with the customer data using a text join method (Fig 7.4.1(g). As seen in the Fig 7.4.1(g), the results of Join 1(merging customer data with country description with country being the common field between the two data tables) are then additionally text jointed with Join_2(which has GBI_DEMO_SALESORG as data source) using the field SALES_ORGANIZATION.

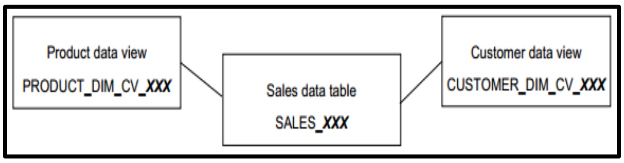
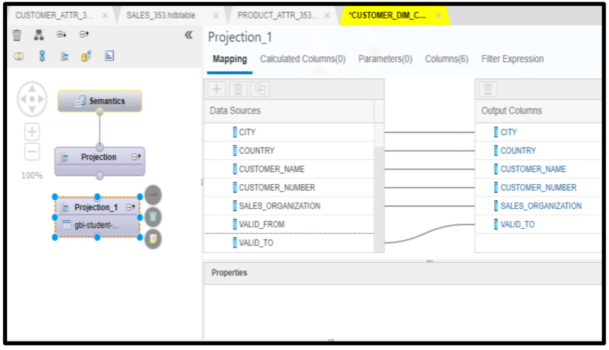


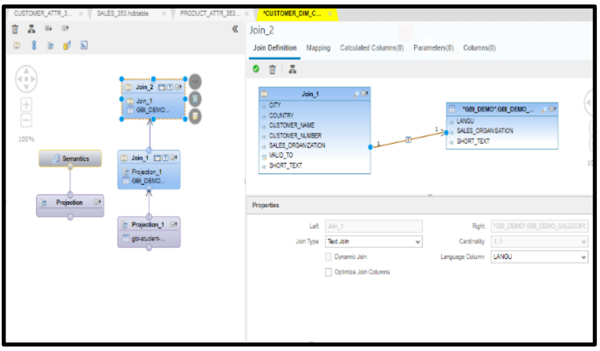
Fig 7.4.1(e): Star schema prototype



Projection_1

Mapping | Calculated Columns(0) | Parameters(0) | Columns(6) | Filter Expression

Data Sources	Output Columns
CITY	CITY
COUNTRY	COUNTRY
CUSTOMER_NAME	CUSTOMER_NAME
CUSTOMER_NUMBER	CUSTOMER_NUMBER
SALES_ORGANIZATION	SALES_ORGANIZATION
VALID_FROM	VALID_TO
VALID_TO	



Join_2

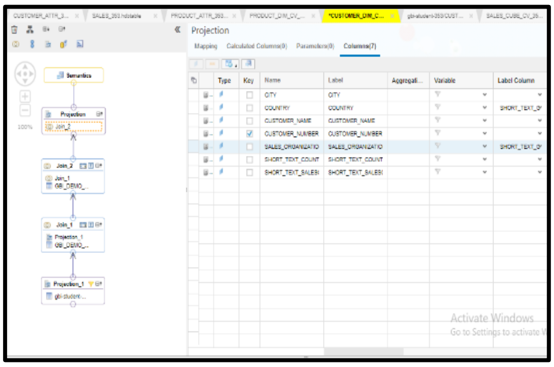
Join Definition | Mapping | Calculated Columns(0) | Parameters(0) | Columns(0)

Left: Join_1
Right: "GBI_DEMO"."GBI_DEMO_SALESORG
Join Type: Text Join
Cardinality: 1..1
□ Dynamic Join
Language Column: LANGU
□ Optimize Join Columns

Fig 7.4.1(f)- Mapping of Customer data for Customer Calculation view

Fig 7.4.1(g)- Joining customer data with salesorg data

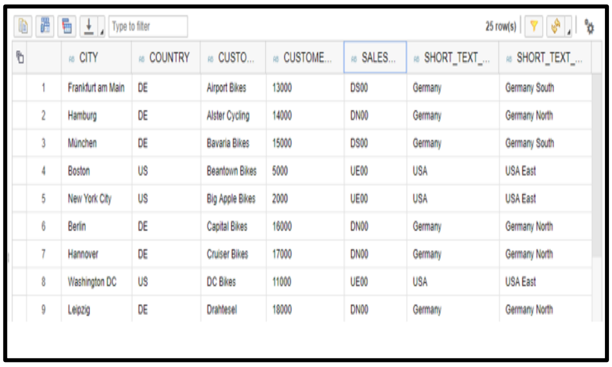
Description mapping for the fields COUNTRY and SALES_ORGANISATION was performed with the Objective being to link a key and text fields for an attribute. To perform the description mapping a value was set for the Label column in the Semantics node of the view. As seen in



Projection

Mapping | Calculated Columns(0) | Parameters(0) | Columns(7)

Type	Key	Name	Label	Aggregati...	Variable	Label Column
	□	CITY	CITY			
	□	COUNTRY	COUNTRY			SHORT_TEXT_C
	□	CUSTOMER_NAME	CUSTOMER_NAME			
	☑	CUSTOMER_NUMBER	CUSTOMER_NUMBER			
	□	SALES_ORGANIZATIO	SALES_ORGANIZATIO			SHORT_TEXT_O
	□	SHORT_TEXT_COUNT	SHORT_TEXT_COUNT			
	□	SHORT_TEXT_SALES	SHORT_TEXT_SALES			



	CITY	COUNTRY	CUSTO...	CUSTOME...	SALES...	SHORT_TEXT_...	SHORT_TEXT_...
1	Frankfurt am Main	DE	Airport Bikes	13000	DS00	Germany	Germany South
2	Hamburg	DE	Alster Cycling	14000	DN00	Germany	Germany North
3	München	DE	Bavaria Bikes	15000	DS00	Germany	Germany South
4	Boston	US	Beantown Bikes	5000	UE00	USA	USA East
5	New York City	US	Big Apple Bikes	2000	UE00	USA	USA East
6	Berlin	DE	Capital Bikes	16000	DN00	Germany	Germany North
7	Hannover	DE	Cruiser Bikes	17000	DN00	Germany	Germany North
8	Washington DC	US	DC Bikes	11000	UE00	USA	USA East
9	Leipzig	DE	Drahtesel	18000	DN00	Germany	Germany North

Fig 7.4.1(h): Customer Calculation view

Fig 7.4.1(i): Contents of Customer data view

Fig 7.4.1(h), the node JOIN_2 was linked with the node PROJECTION and all fields were added as output columns. By clicking the node Semantics, for field COUNTRY, label column was set to SHORT_TEXT_COUNTRY. For field SALES_ORGANISATION, label column was set to SHORT_TEXT_SALESORG. Finally, the fields SHORT_TEXT_COUNTRY and SHORT_TEXT_SALESORG were hidden. The customer number was defined as key field for the customer calculation view. On saving the view, it was then run to check the contents as seen in Fig 7.4.1(i). A calculation view for the product data was created (Fig 7.4.1(j)) in a similar way and the semantics was configured to make the field 'PRODUCT' as key field.

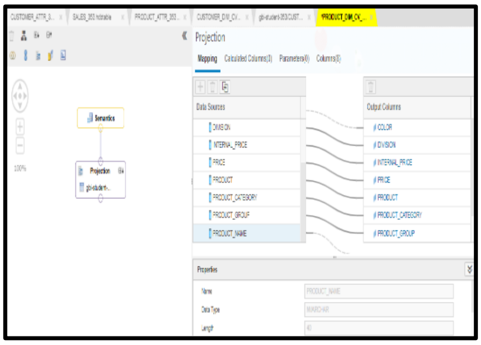
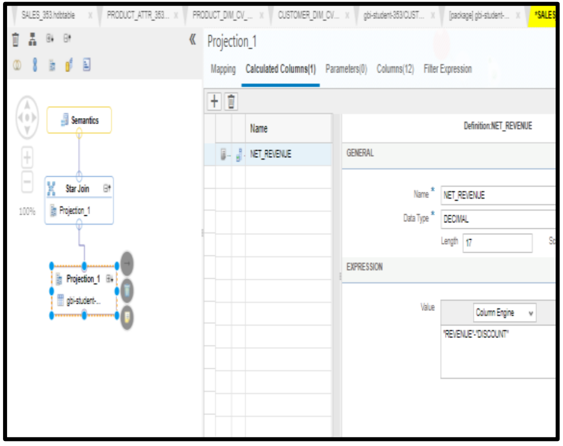
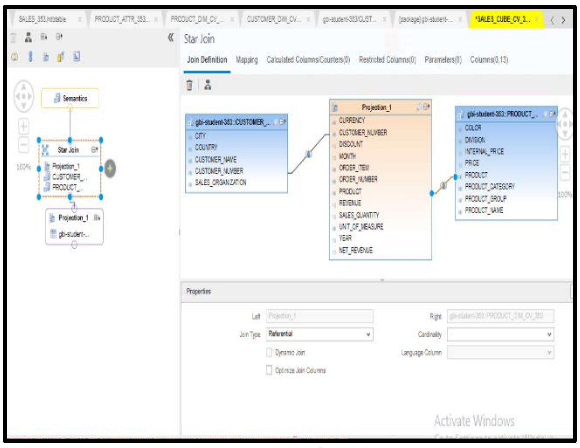


Fig 7.4.1(j) – Product data calculation view

To implement the star join as seen in Fig 7.4.1(e), a calculation view name- ***SALES_CUBE_CV_353*** – with data category as cube was created using star join for sales data. The sales table was added as input to the projection node, and all the fields except Day were added to the output structure of the projection. As seen in Fig 7.4.1(k) the projection node was linked to the star join node and NET REVENUE – a calculated column was created. The star join node was then configured to join the sales table with the customer view CUSTOMER_DIM_CV_353 using the field CUSTOMER_NUMBER and a referential join and with the product view PRODUCT_DIM_CV_353 using the field PRODUCT and a *referential* join. (Fig 7.4.1(l))



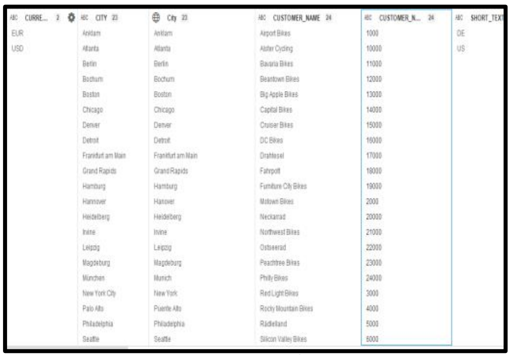
Name: NET_REVENUE
Data Type: DECIMAL
Length: 17
Value: "REVENUE"-"DISCOUNT"



Left: Projection_1
Join Type: Referential
Right: gbi-student-353::PRODUCT_DIM_CV_353
Cardinality
Language Column

Fig7.4.1(k)- Calculated column- Net Revenue in Sales cube view **Fig 7.4.1(l): Creating a star join of sales cube**

Under mapping, all fields were added as output columns except the fields CUSTOMER_NUMBER and PRODUCT as they are included in dimensions. The Semantics node was viewed to recheck the measures and dimensions of the dataset. The calculation view was then run to produce the sales data of GBI inclusive of customer and product details in the form of columnar storage (Fig 7.4.1(m)). SAP predictive expert analytics was used to analyze and visualize the sales dataset created in SAP HANA. For this, SAP HANA server was accessed from SAP PA using the same login credentials and the calculation view SALES_CUBE_CV_353 was selected for analysis. In the visualize mode, a combined column chart with line with 2 Y-Axes was selected with Net Revenue in Y-Axis 1 and Sales quantity in the Y-Axis 2, Year and Month in X-Axis and Short_TEXT_Country in Color which resulted in Fig 7.4.1(n)



CURRE... 2	CITY 23	City 23	CUSTOMER_NAME 24	CUSTOMER_N... 24	SHORT_TEXT
EUR	Anklam	Anklam	Airport Bikes	1000	DE
USD	Atlanta	Atlanta	Alster Cycling	10000	US
	Berlin	Berlin	Bavaria Bikes	11000	
	Bochum	Bochum	Beantown Bikes	12000	
	Boston	Boston	Big Apple Bikes	13000	
	Chicago	Chicago	Capital Bikes	14000	
	Denver	Denver	Cruiser Bikes	15000	
	Detroit	Detroit	DC Bikes	16000	
	Frankfurt am Main	Frankfurt am Main	Drahtesel	17000	
	Grand Rapids	Grand Rapids	Fahrpott	18000	
	Hamburg	Hamburg	Furniture City Bikes	19000	
	Hannover	Hanover	Motown Bikes	2000	
	Heidelberg	Heidelberg	Neckarad	20000	
	Irvine	Irvine	Northwest Bikes	21000	
	Leipzig	Leipzig	Ostseerad	22000	
	Magdeburg	Magdeburg	Peachtree Bikes	23000	
	München	Munich	Philly Bikes	24000	
	New York City	New York	Red Light Bikes	3000	
	Palo Alto	Puente Alto	Rocky Mountain Bikes	4000	
	Philadelphia	Philadelphia	Radleland	5000	
	Seattle	Seattle	Silicon Valley Bikes	6000	

Fig7.4.1(m) Columnar storage of SALES CUBE data

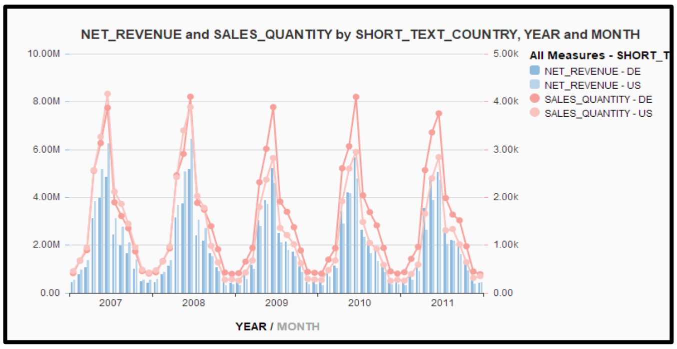


Fig 7.4.1(o) Net Revenue & Sales quantity trends over the years by country

Key Findings: The data table loaded from SAP HANA server is in the form of columnar storage and all values of a field are stored sequentially in memory. Also, it is easy to identify in which countries GBI sells its products and to which customers. The key finding of the visualization is that for the year 2007 and 2008, US has produced more revenue for GBI than Germany but the net revenue has decreased by almost 7 million in US for the remaining years along with a peak drop in sales from 2007-2008. On the other hand, Germany has been slowly increasing its revenue since 2007 and has performed better than US from 2009-2011. One more key aspect is that the seasonality of GBI sales has remained constant for all the years making June as the top-selling month.

7.4.2 *Which product category is the highest revenue generator of GBI?*

To illustrate which product category is the highest revenue generator of GBI, a Tree Map was selected in the Visualize mode of SAP PA. Net_Revenue was added to Area Weight, Sales_Quantity to Area Color and the dimension, Product_Category, Product_Name were added to Area Name. This resulted in a visualization as seen Fig 7.4.2 (a).

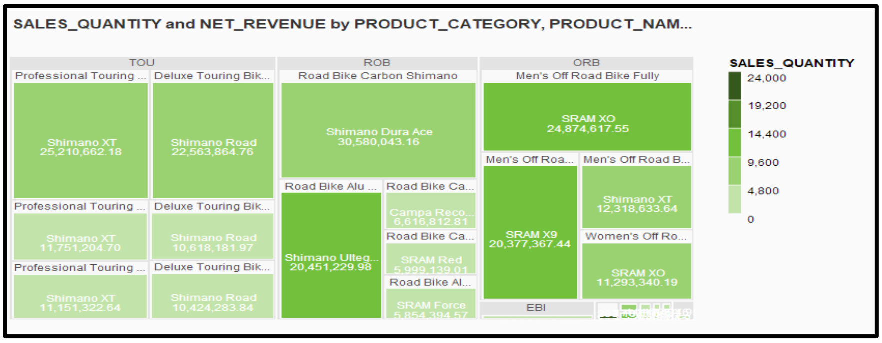


Fig 7.4.2(a) Net Revenue and Sales quantity by Product Category, Name and Group

Key Findings : As observed in the fig, Product Category – TOU (Touring Bikes) is the highest revenue generator for GBI following by ROB (Road Bikes) while TRE being the lowest. However, it is also observed that the products generating the highest revenue are not massively sold. On the other hand, Air pump (the product with the highest sales quantity) produces minimum revenue for GBI. Mostly there are other factors like Discount effecting the sales of these products. GBI can further analyze to understand why the top revenue products are being sold in less quantities and plan purchasing accordingly.

7.4.3 *Who is the star customer of GBI?*

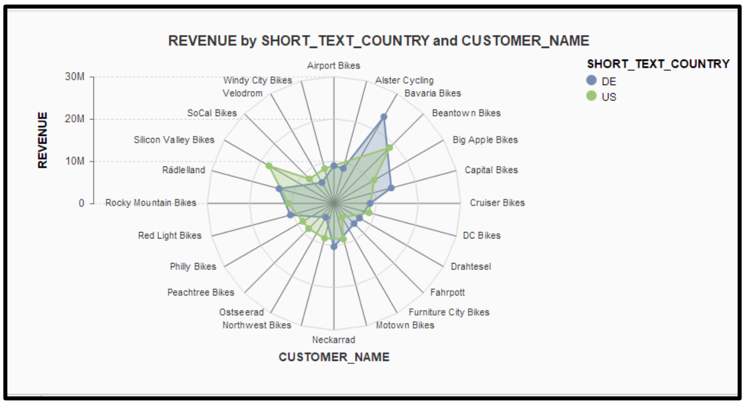


Fig 7.4.3(a) Revenue trends of GBI customers by country

For this question, the radar chart was used to plot the net revenue of all the customers of GBI based out in the two countries as seen in Fig 7.4.3(a).

Key Findings: Bavaria Bikes of Germany is the top customer of GBI followed by Bean Town Bikes of USA.

SECTION 7.5 ANALYSIS AND CRITIQUE OF THE TOOL

As a first-time user of SAP HANA, I enjoyed working in the development environment of an in-memory relational database management system. I have previously worked on other relational database servers like MySQL and Oracle that store data in hard-drive but building a database model using SAP HANA was exciting as it stored all relevant data in RAM. I found columnar data storage feature (finding all data values for a field at a single column) as the key highlight of SAP HANA. It has built-in, high-availability functions that keep the database running and ensure mission-critical applications are never down (Surya, n.d.).

Moreover, SAP HANA ensures significant reduction of data volume, fast processing and considerable reduction of backup and restore processes. Whenever, I executed a view the result could instantly turn up on the server, ensuring fast data performance. I also found the concept of flowgraph model very fascinating as data gets prepared and populated at run-time without worrying about the internal processing. On the other hand, when loading data from CSV files in MYSQL server, sql scripting is required. SAP HANA makes the whole data loading process procedural and easy. I also loved the way calculation views are created and merged in this tool. Although, the process is quite automated, I faced an issue when merging the customer data with country demo data using text join. Even though, the datasets were merged, the calculated view could not fetch the results from the country demo data table. It was difficult to drill down and download the CSV file as it was uploaded in the server.

Overall, my experience with SAP HANA has been great and I am looking forward to experiencing the development capabilities of SAP HANA.

SECTION 7.6 CONCLUSION

SAP HANA being an in-memory and cloud-oriented database becomes the forefront technology for processing Big Data. This technology can help businesses achieve real-time analytics which becomes very useful in providing reliable and fast services to clients and customers. SAP HANA is here to stay as speed is one of the essential components of Business Analytics.

REFERENCES

BI, V. (2017, August 21). *SAP Lumira Discovery: An Overview by Visual BI Solutions*. Visual BI Solutions. https://visualbi.com/blogs/sap/sap-businessobjects/sap-lumira-discovery/sap-lumira-discovery-overview/

SAP Lumira (n.d). *Self-Service Data Visualization & BI*. (n.d.). SAP. Retrieved April 8, 2021, from https://www.sap.com/products/lumira.html

Statistics Canada. (4 May 2015). *Value of sales of alcoholic beverages of liquor authorities and other retail outlets, by beverage type*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010001101

Statistics Canada. (29 September 2020). *Population estimates on July 1st, by age and sex*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501

Fransen, C. (2021, February 12). *Benefits of Using Excel Pivot Tables | Calgary Microsoft Tech Articles*. CTECH Consulting Group. https://www.ctechgroup.ca/benefits-of-using-excel-pivot-tables/

SAP Predictive Analytics (2016, November 2). *Reviews, Features, Pricing, Comparison.* PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices. https://www.predictiveanalyticstoday.com/sap-predictive-analytics/

SAP Predictive Analytics(n.d.). *An Overview*. STechies. http://www.stechies.com/predictive-analytics/

Anurag (2018, September 14). *What is Tableau, its uses and applications*. NewGenApps. https://www.newgenapps.com/blog/what-is-tableau-its-uses-and-applications/

Bora, B. (2016, November 4). *Leverage the power of Python in Tableau with TabPy*. https://www.tableau.com/about/blog/2016/11/leverage-power-python-tableau-tabpy-62077

Virginia, B. (2018, January 18). *An Introduction to Tableau: What It is and How It Can Provide Insight for Your Business*. CMSWire.Com. https://www.cmswire.com/analytics/an-introduction-to-tableau-what-it-is-and-how-it-can-provide-insight-for-your-business/

SAP Analytics Cloud(n.d.). *BI, Planning, and Predictive Analysis Tools*. SAP. https://www.sap.com/products/cloud-analytics.html

Idexel Technologies (2017, December 28). *Advantages of Cloud Analytics over On-Premise Analytics – Blog* | Idexcel. Retrieved April 8, 2021, from https://www.idexcel.com/blog/advantages-of-cloud-analytics-over-on-premise-analytics/

Weaver, L. (2017, January 16). *6 Ways Cloud Analytics Is Better, Faster, Cheaper*. Interconnections - The Equinix Blog. https://blog.equinix.com/blog/2017/01/16/6-ways-cloud-analytics-is-better-faster-cheaper/

SAP Predictive Analytics Enterprise Edition. (n.d*).SAP Help Portal*. Retrieved April 8, 2021, from https://help.sap.com/viewer/6887a1b4e5f348dc9d297b91701bda3d/3.3/en-US/cacdebd91a2641838bc1f7323dd4ef31.html

Hart, M. (2015, March 1). *SAP Predictive Analytics—Benefits and Features*. SapMe. https://sap.walkme.com/sap-predictive-analytics-benefits-and-features/

SAP HANA (n.d.). Why SAP HANA? Overview of Benefits, Features, and Capabilities. SAP. Retrieved April 8, 2021, from https://www.sap.com/products/hana/features.html

Surya (n.d.). *A Few benefits of SAP S/4HANA*.SAP Consultants | SAP Experts | On-Demand SAP Consulting & Help. Retrieved April 8, 2021, from https://www.erpfixers.com/blog/2020/6/2/a-few-benefits-of-sap-s4hana

Kale, N. and Jones, N. (2020). Chapter ten: Data mining. In *Practical analytics: Second edition* (pp. 409-420). Epistemy Press.

Kale, N. and Jones, N. (2020). Chapter thirteen: Big Data analytics. In *Practical analytics: Second edition* (pp. 499-520). Epistemy Press.