Analysis Project

Course Technion Certified ‘Big Data Analyst’

Lecturer Amir Ben Haim

Submitted by Hedva Meiri and Einat Farkash

Table of Contents

Project Overview.............................. 01

Star Schema... 02

The Project Phases.. 03

Challenges & Solutions.................................. 04

Dashboards... 05

Reports.. 06

Conclusions..................................... 07

Project Overview

Goal: Utilize our knowledge of programs such as Power Query/Pivot, Power BI, and DAX language to establish and present data-backed business insights.

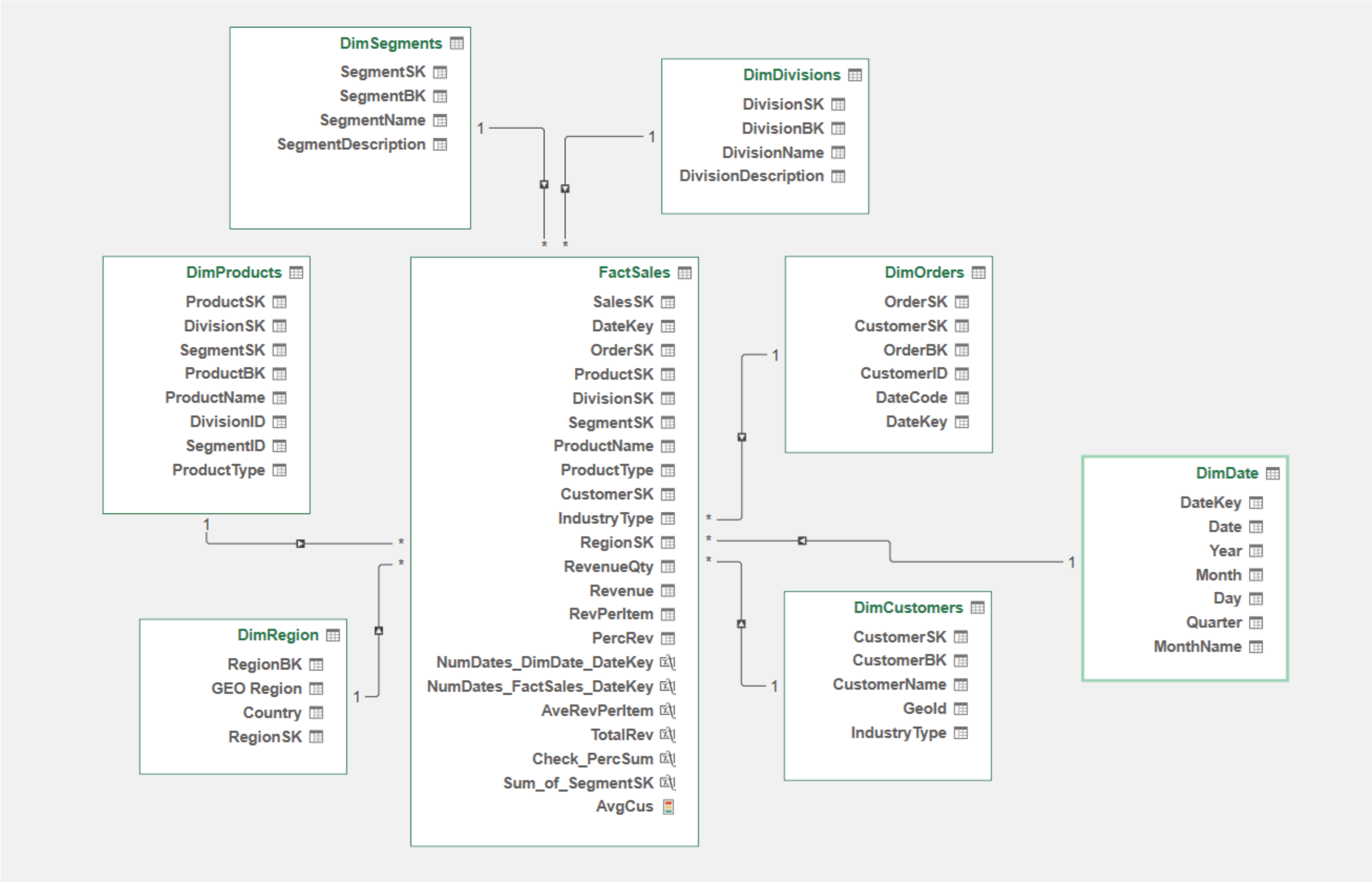
Methods:

1. Create Happy Insurance DB from provided .sql file and import data from Region.csv.
2. Query the DB in order to become familiar with the data and extract basic information regarding tables, columns, and relationships.
3. Upload DB to excel workbook.
4. Create Model with Power Query.
5. Produce measures and calculated columns; hierarchies and functions. Then, using Power Pivot, produce visualizations and reports based on the Model.
6. Generate visualizations and reports in Power BI, assisted by DAX language.

Assumption(s):

1. The DB is complete. There is no missing information that could impact the veracity of business insights.

Star Schema – Power Pivot



DimSegments

SegmentSK
SegmentBK
SegmentName
SegmentDescription

DimDivisions

DivisionSK
DivisionBK
DivisionName
DivisionDescription

DimProducts

ProductSK
DivisionSK
SegmentSK
ProductBK
ProductName
DivisionID
SegmentID
ProductType

FactSales

SalesSK
DateKey
OrderSK
ProductSK
DivisionSK
SegmentSK
ProductName
ProductType
CustomerSK
IndustryType
RegionSK
RevenueQty
Revenue
RevPerItem
PercRev
NumDates_DimDate_DateKey
NumDates_FactSales_DateKey
AveRevPerItem
TotalRev
Check_PercSum
Sum_of_SegmentSK
AvgCus

DimOrders

OrderSK
CustomerSK
OrderBK
CustomerID
DateCode
DateKey

DimDate

DateKey
Date
Year
Month
Day
Quarter
MonthName

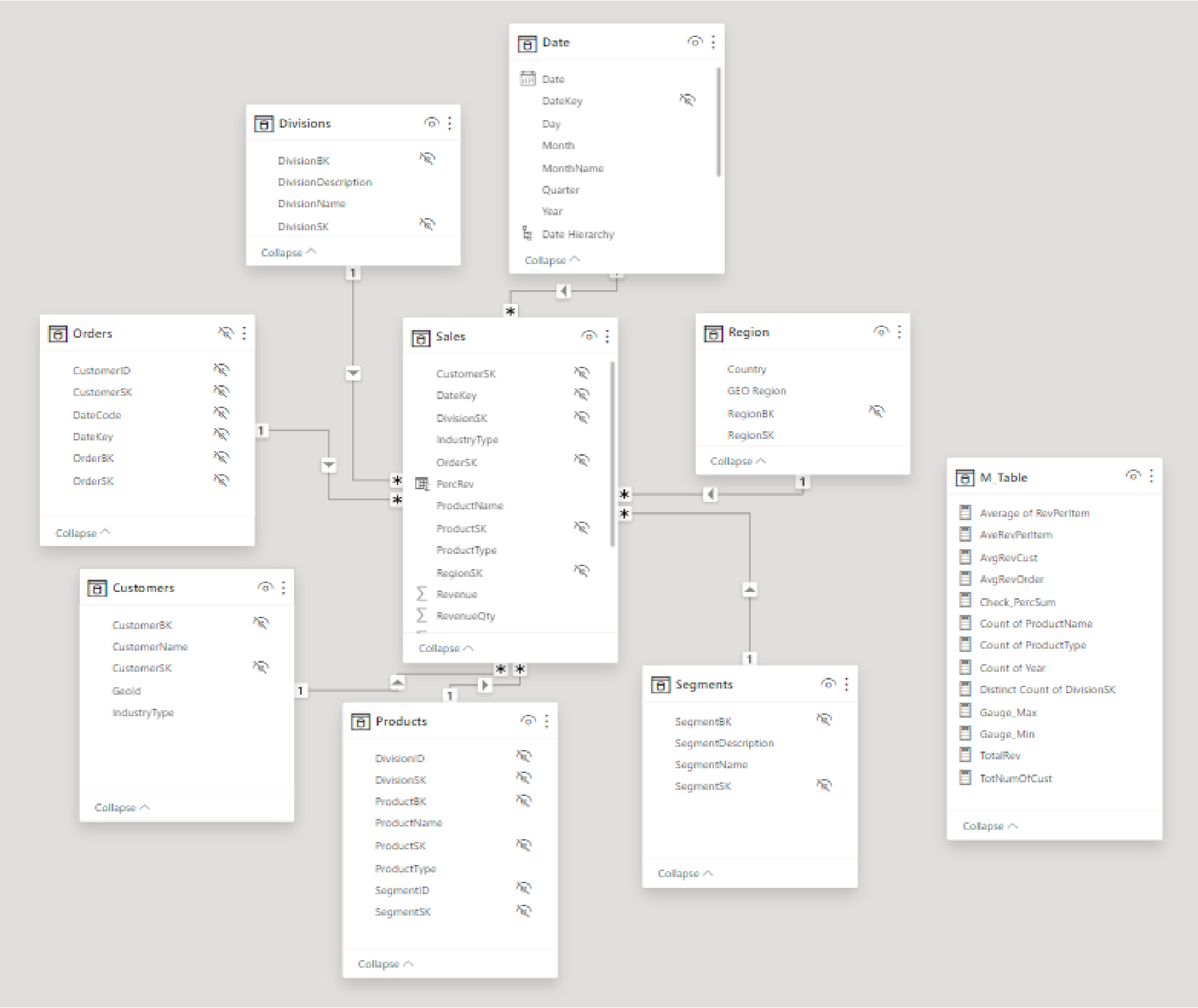
DimRegion

RegionBK
GEO Region
Country
RegionSK

DimCustomers

CustomerSK
CustomerBK
CustomerName
GeoId
IndustryType

Star Schema – Power BI

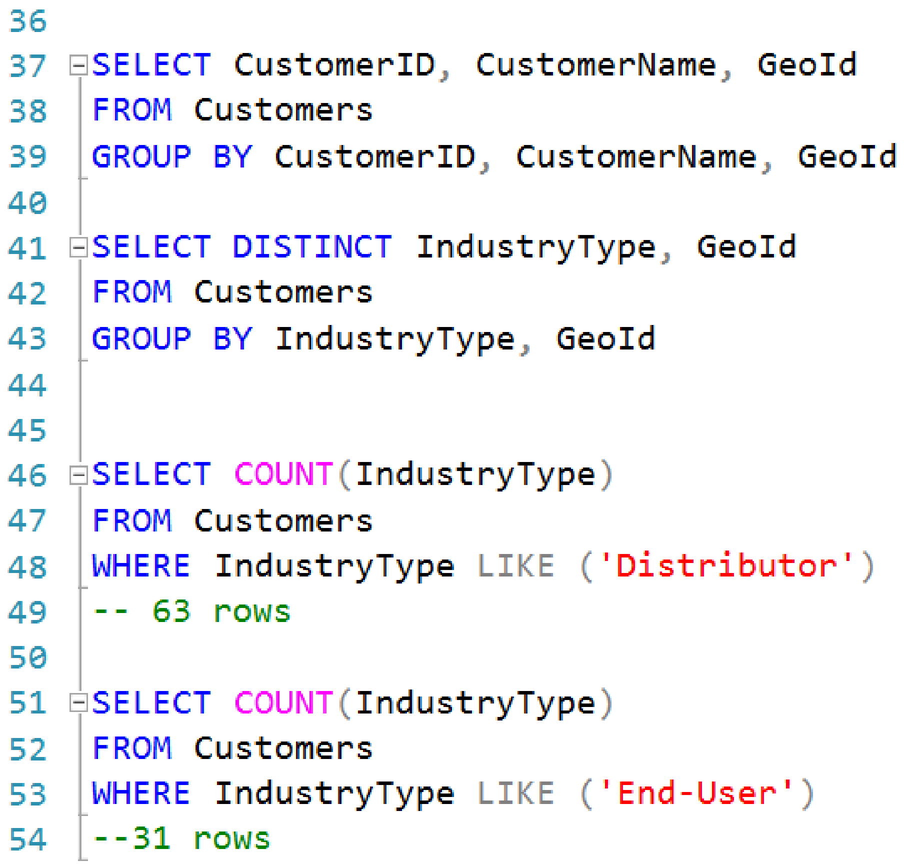


The Project Phases

Phase 1 - ETL:

The first phase of the project was dedicated to the ETL process. The Happy Insurance Database (DB) was created using the .sql and .csv files provided. Data was initially looked at using SQL queries to understand quantities, groupings, and relationships (partial example on right).

Afterwards, data was exported to an excel workbook and analyzed with Power Query. Replacements for NULL values in Products[ProductType]and Segments[SegmentDescription] were provided.



```sql

SELECT CustomerID, CustomerName, GeoId
FROM Customers
GROUP BY CustomerID, CustomerName, GeoId

SELECT DISTINCT IndustryType, GeoId
FROM Customers
GROUP BY IndustryType, GeoId


SELECT COUNT(IndustryType)
FROM Customers
WHERE IndustryType LIKE ('Distributor')
-- 63 rows

SELECT COUNT(IndustryType)
FROM Customers
WHERE IndustryType LIKE ('End-User')
--31 rows
```

The Project Phases

Phase 2 - Star Schema:

The second phase of the project was dedicated to creating the Star Schema in Power Query. Each table from the DB was examined. The tables were transformed into dimension tables with the addition of Surrogate Keys (SKs) and the removal of unnecessary columns.

In addition, a DImDate table was created. The DimDate table was produced by creating a list of consecutive dates from 1.1.2011 - 31.12.2013. Each date was then transformed into a DateKey of datatype text.

Following that, Merge Queries was used to join between relevant columns among the dimension tables, merge the SKs from the dimension tables and DateKey from DimDate with the OrderLines table, and convert OrderLines to FactSales.

On occasion we utilized M Language to simplify these procedures.

The final model can be viewed on slide 4.

The Project Phases

Phase 3 - Power Pivot:

Once the Star Schema was finalized we began working on the data in Power Pivot. Here we created measures in the calculation area of the FactSales tab and the calculated columns in the tables.

At this point we discussed which business insights we wanted to focus on and based our choice of visualizations on them.

Examples of Measures:

Check percent sum
Average Revenue Per Item
Total Revenue
Average Customer Revenue

.

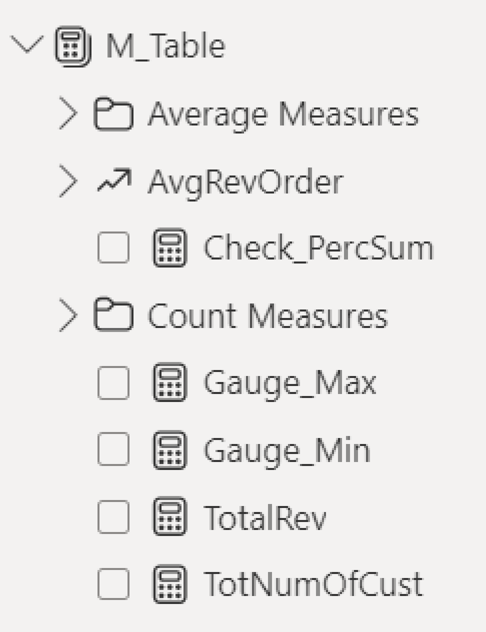
Examples of Calculated Columns:

Percent Revenue
Revenue Per Item

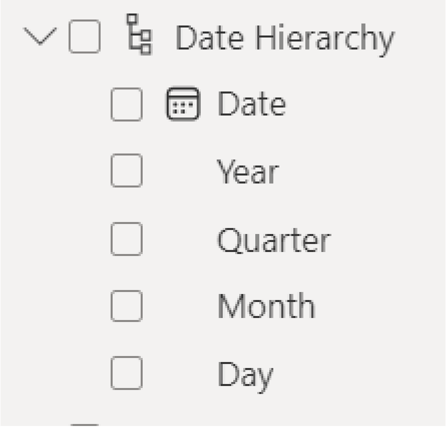
The Project Phases

Phase 4 - Power BI and DAX:

We continued our analyses of the data using Power BI and DAX language, leading to the final steps of the project and the creation of our business reports and dashboards. Improvements on the model created in excel include filtered rows / tables, kpis, and hierarchies. Also, a measures table was created in order to consolidate the measures that were created. The use of subfolders in the measures table allowed for further organization of the data.



M_Table
- Average Measures
- AvgRevOrder
- □ Check_PercSum
- Count Measures
- □ Gauge_Max
- □ Gauge_Min
- □ TotalRev
- □ TotNumOfCust



□ Date Hierarchy
- □ Date
- □ Year
- □ Quarter
- □ Month
- □ Day

Challenges and Solutions

Creation of DimDate table and merge with FactSales

One of the early challenges of the project was the formation of the DimDate table in Power Query. It was imperative that the table be formed with consecutive dates, even if no orders were placed on a particular day. Then the date parts needed to be extracted properly in order to form the DateKey.

First, we created a file with 3 years of consecutive dates from 1.1.11 - 31.12.13. This was accomplished by using M language to create two variables for the start and end dates. Then the following code created the list of dates by day:

```
dateList = List.Dates(
        StartDate,
        Duration.Days(EndDate - StartDate) + 1,
        #duration(1, 0, 0, 0)
                )
```

The next step was merging the DimDate table with the FactSales table. As only yearly and quarterly data were available, each quarter (per year) was paired with Jan 1, Apr 1, Jul 1 and Oct 1, respectively (per year).

Challenges and Solutions

Decision-making: Which measures should we create?
Which elements can be deleted from the model?

When the Star Schema was completed a significant challenge of this project became choosing which analyses would yield the most interesting and relevant results. Which measures does it make sense to create and what purpose would they serve if created?

After making and implementing those decisions, we were tasked with deleting the unnecessary fields from the Model. Though some fields were deleted immediately (e.g., Sum of DivisionSK), others required more careful consideration and discussion.

Challenges and Solutions

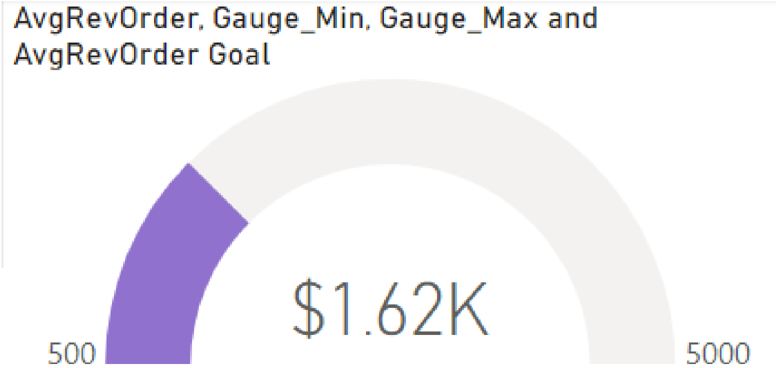
Creation of Gauges & KPIs

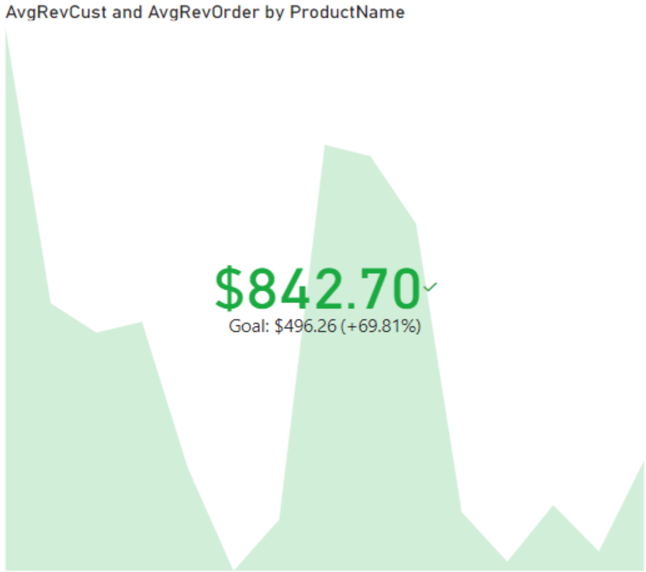
In the main dashboard, we created a gauge visualization of the Average Order Value by defining the minimum and maximum values of the gauge and the target value of the KPI.

The primary challenge lay in selecting the most appropriate measure for the KPI, and determining the values that would most accurately demonstrate the goal values.

The solution was developed through iterative experiments. This involved gradually building the KPI and examining the data at each stage.

For example, choosing values that provide data for comparison purposes, such as using "Segment name" instead of "SegmentSK".





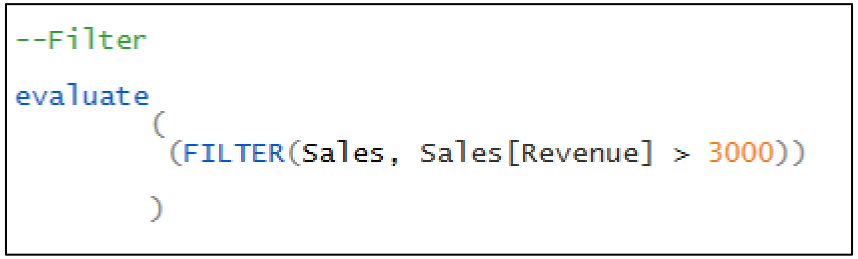
Challenges and Solutions

Creation of Queries in DAX

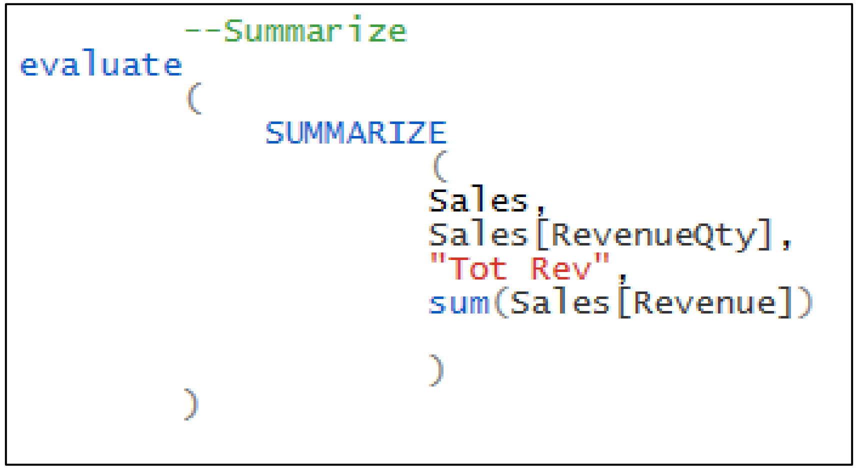
DAX is designed for data analysis in the Power BI tool and in order to explore the data, it was necessary to create queries, in a similar way to SQL.

One of the main challenges in working with DAX was ensuring our queries correctly reflected the data so that our business insights were precise.

In addition, working with the DAX language requires a logical and technical understanding in order to perform calculations and use the measures correctly. Our understanding of the language was not immediate and required a large investment of time.

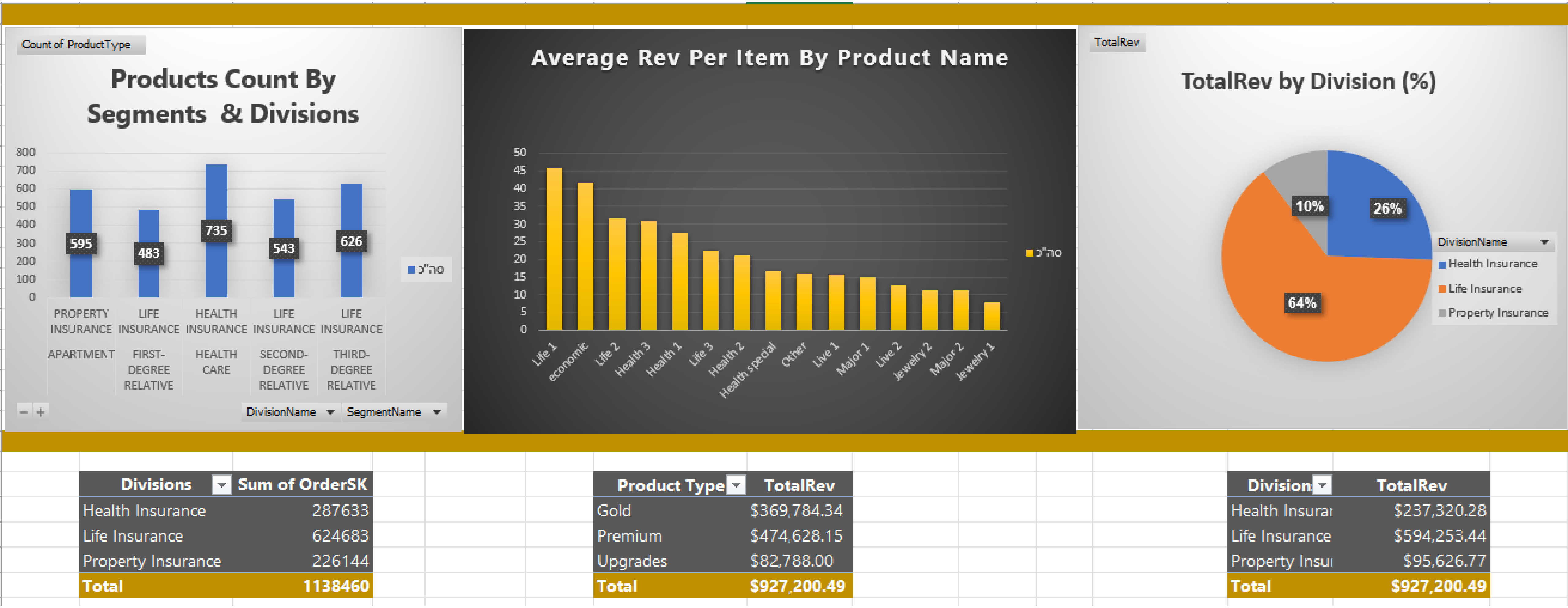


```
--Filter

evaluate
        (
         (FILTER(Sales, Sales[Revenue] > 3000))

        )
```



```
        --Summarize
evaluate
        (
            SUMMARIZE
                    (
                    Sales,
                    Sales[RevenueQty],
                    "Tot Rev",
                    sum(Sales[Revenue])

                    )
        )
```

Excel Dashboard

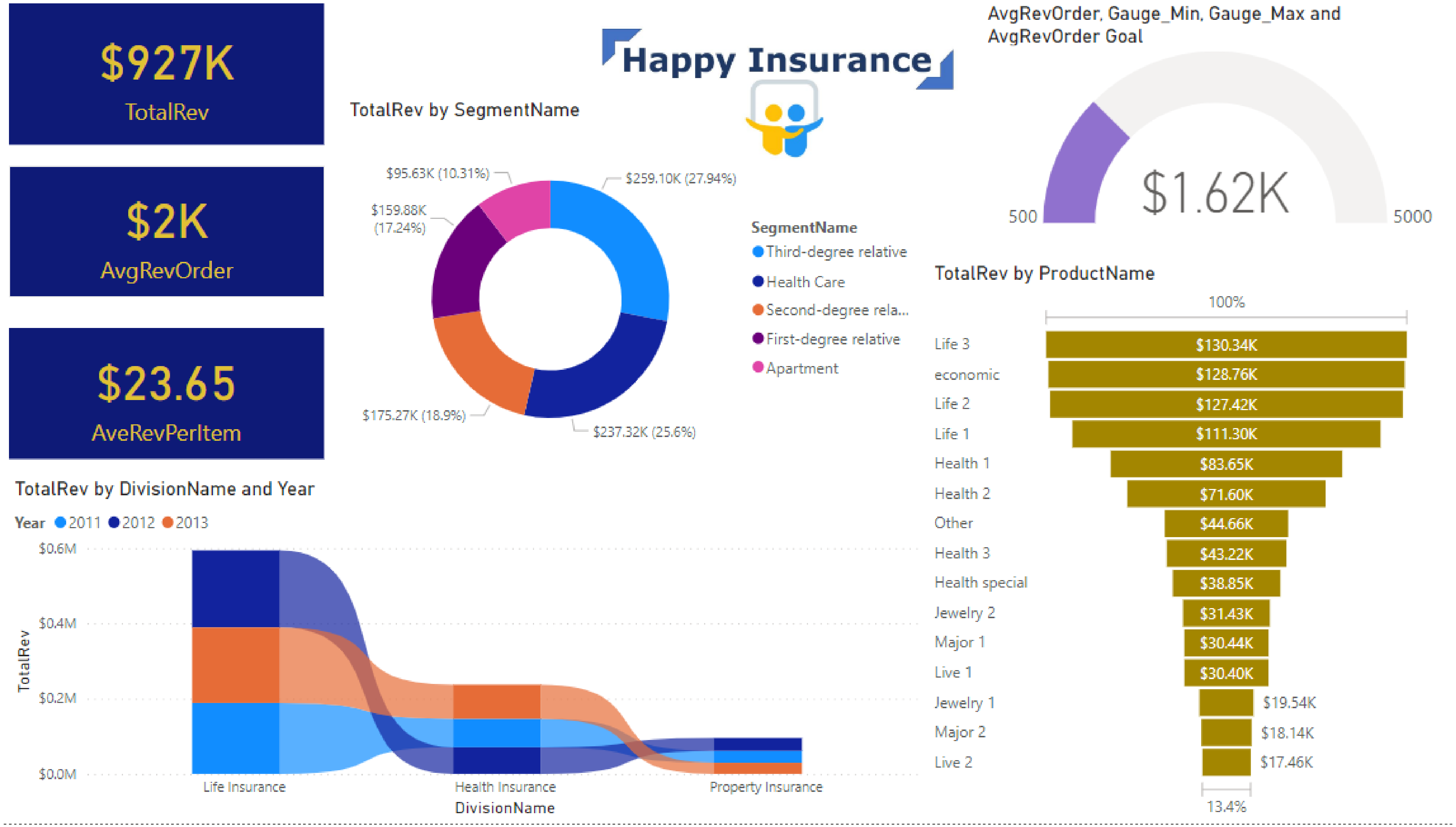


Divisions	Sum of OrderSK
Health Insurance	287633
Life Insurance	624683
Property Insurance	226144
Total	**1138460**

Product Type	TotalRev
Gold	$369,784.34
Premium	$474,628.15
Upgrades	$82,788.00
Total	**$927,200.49**

Division	TotalRev
Health Insurai	$237,320.28
Life Insurance	$594,253.44
Property Insui	$95,626.77
Total	**$927,200.49**

Power BI Dashboard 1

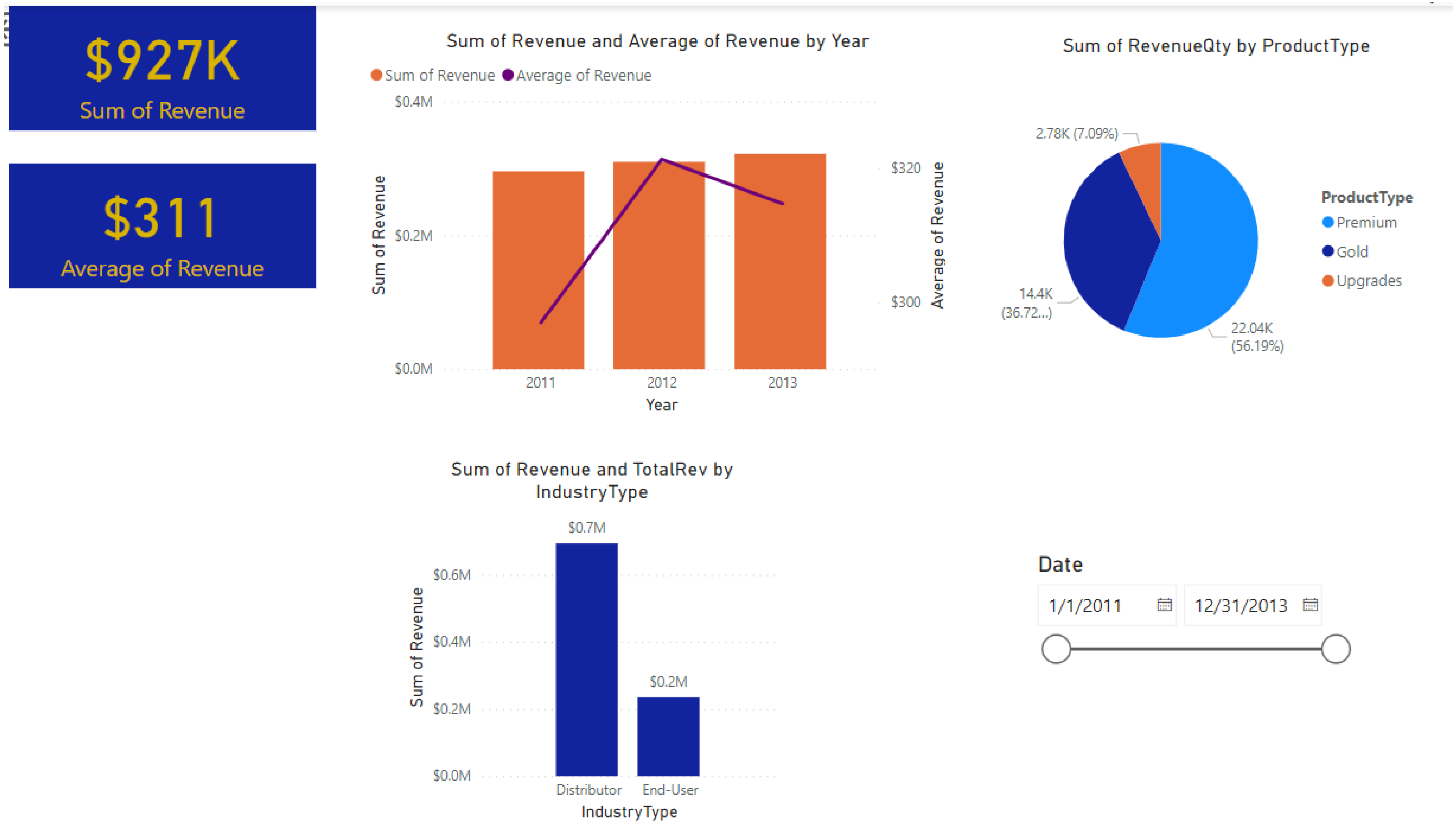


$927K
TotalRev

$2K
AvgRevOrder

$23.65
AveRevPerItem

Power BI Dashboard 2



$927K

Sum of Revenue

$311

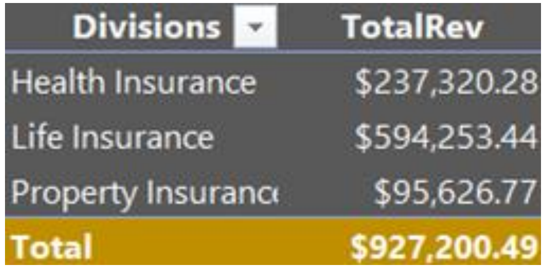
Average of Revenue

Date

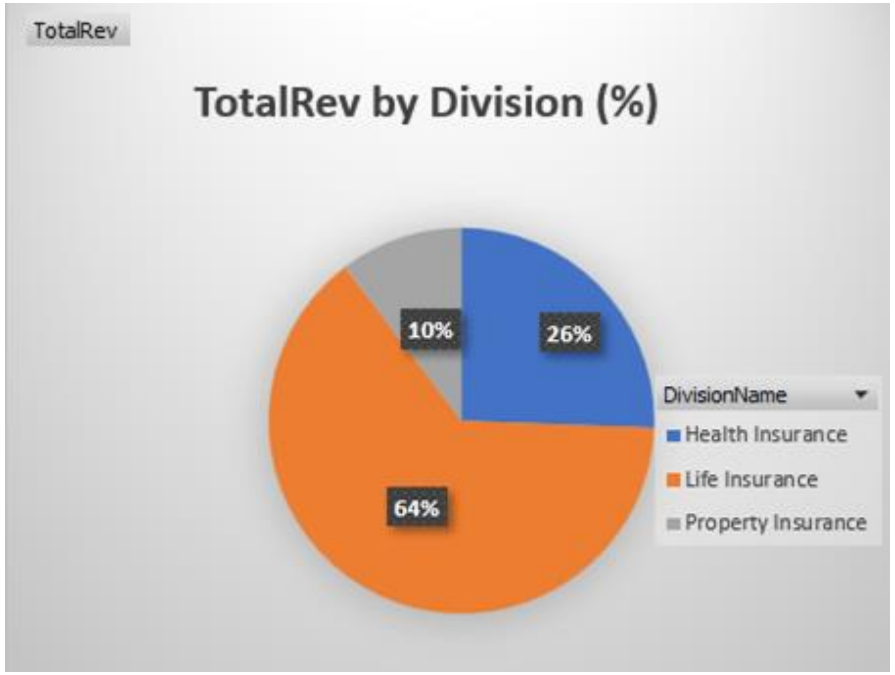
1/1/2011 12/31/2013

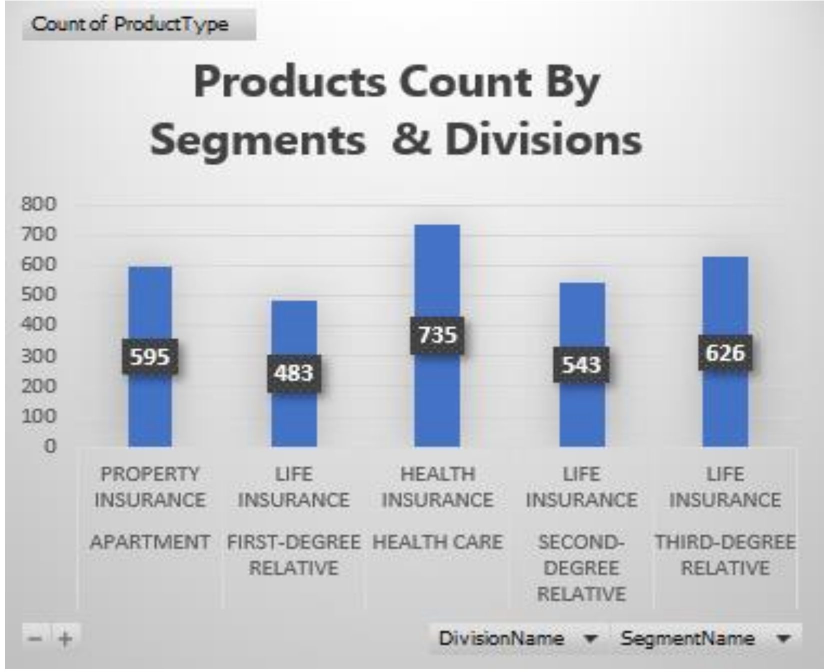
Report 1

★ The Life Insurance division is the strongest revenue producer over time, even accounting for the fact that it is made up of three segments.



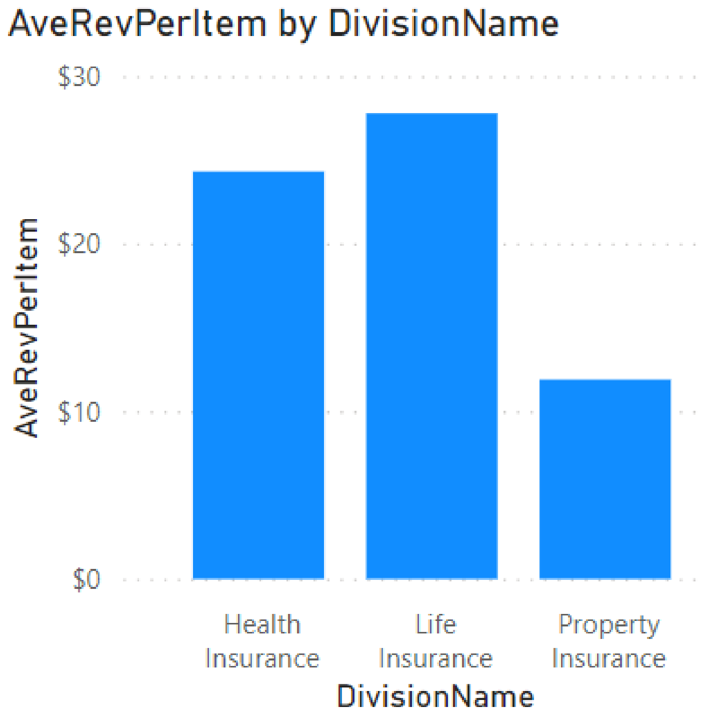
Divisions	TotalRev
Health Insurance	$237,320.28
Life Insurance	$594,253.44
Property Insuranc	$95,626.77
Total	**$927,200.49**

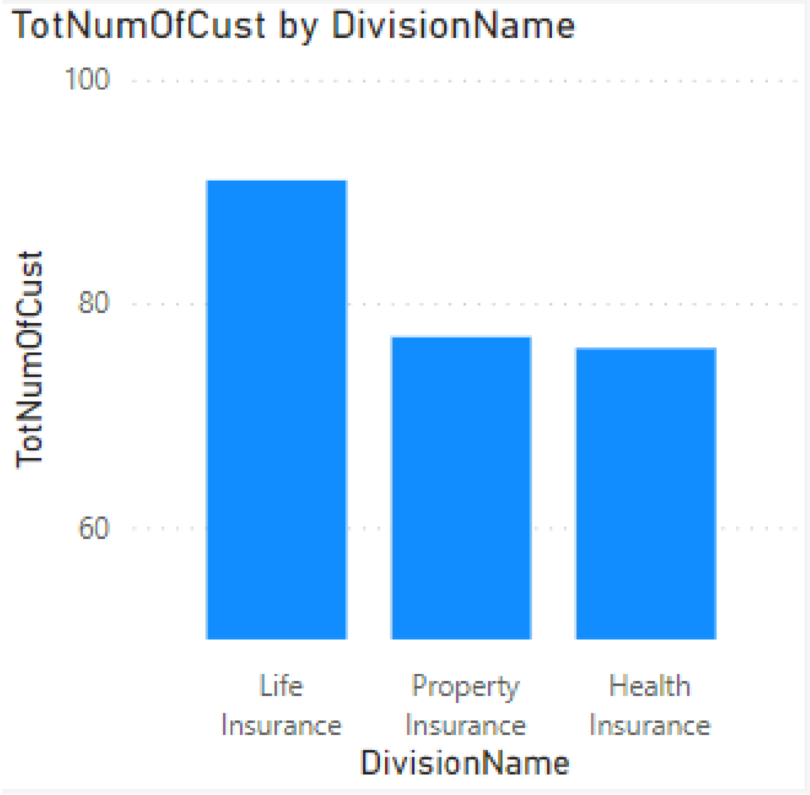




Report 2

★ The Life Insurance division also had the highest number of customers and highest revenue per item.

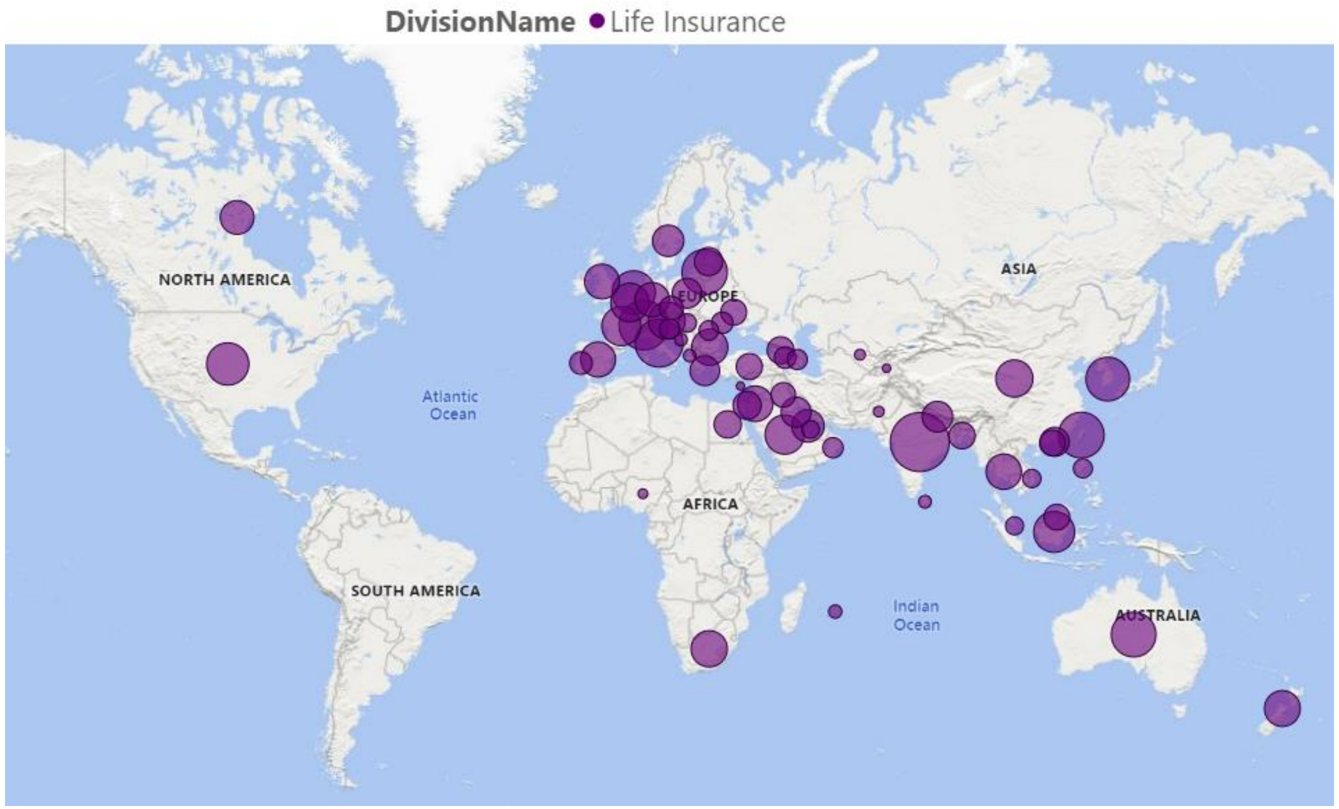




Report 3

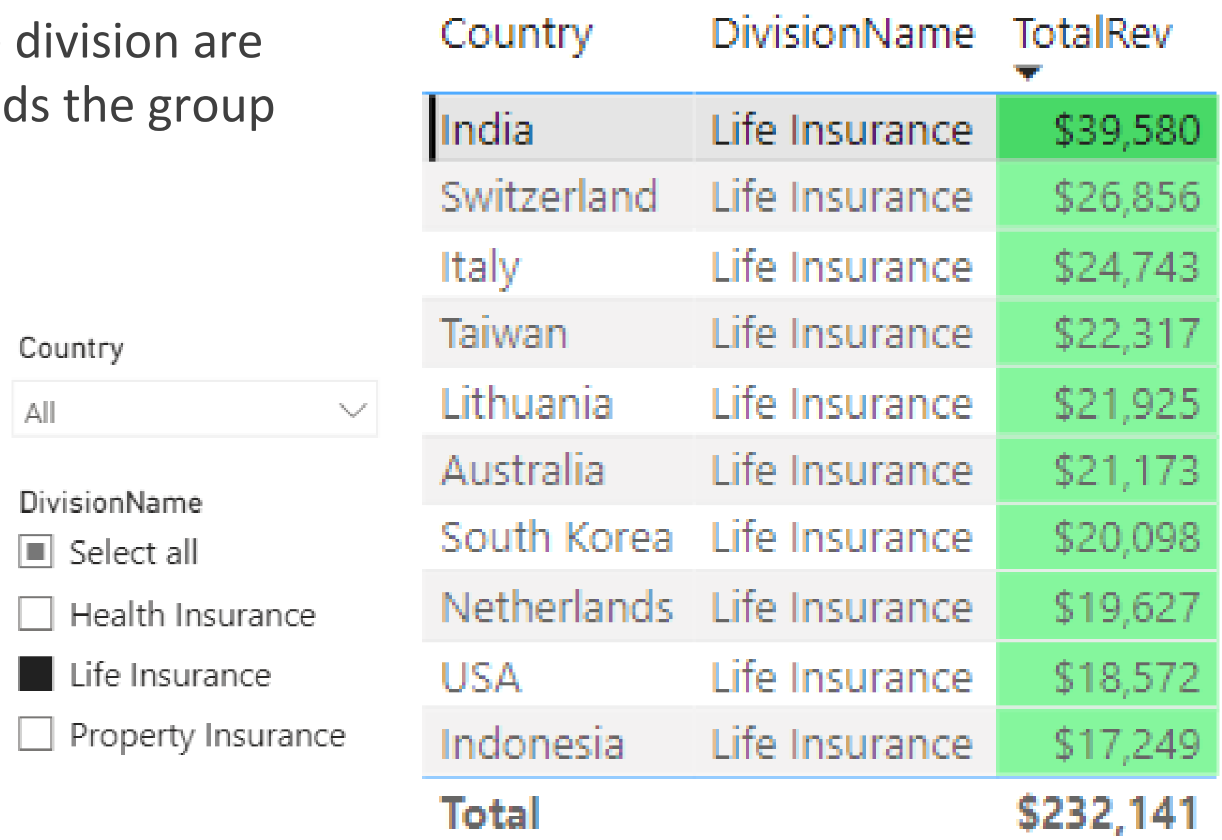
★ Sales of Life Insurance are lacking in South America and can be increased in areas of Asia. Increasing global sales will lead to even greater revenue.

SUM OF PERCREV BY COUNTRY AND DIVISIONNAME



Report 4

★ The top 10 countries which contributed to the total revenue in the Life Insurance division are represented in this table. India leads the group by a large margin.



Country

All

DivisionName

☒ Select all

☐ Health Insurance

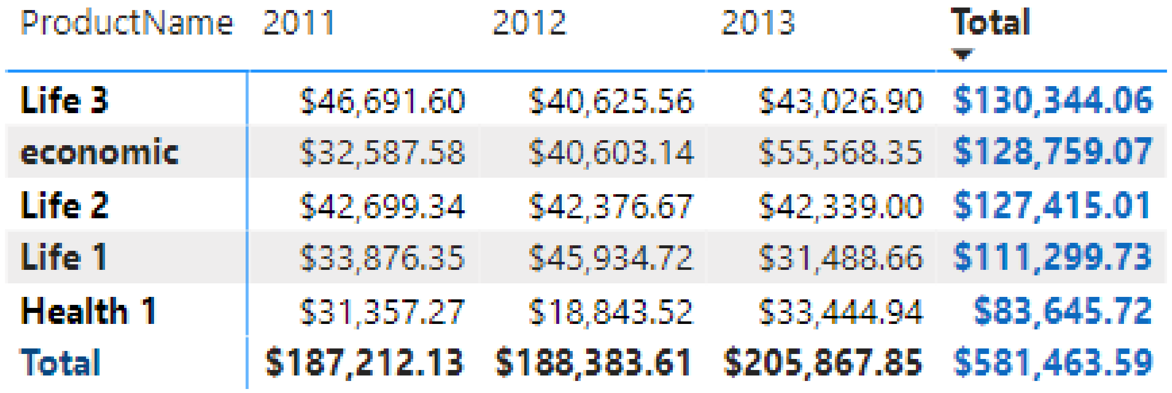
■ Life Insurance

☐ Property Insurance

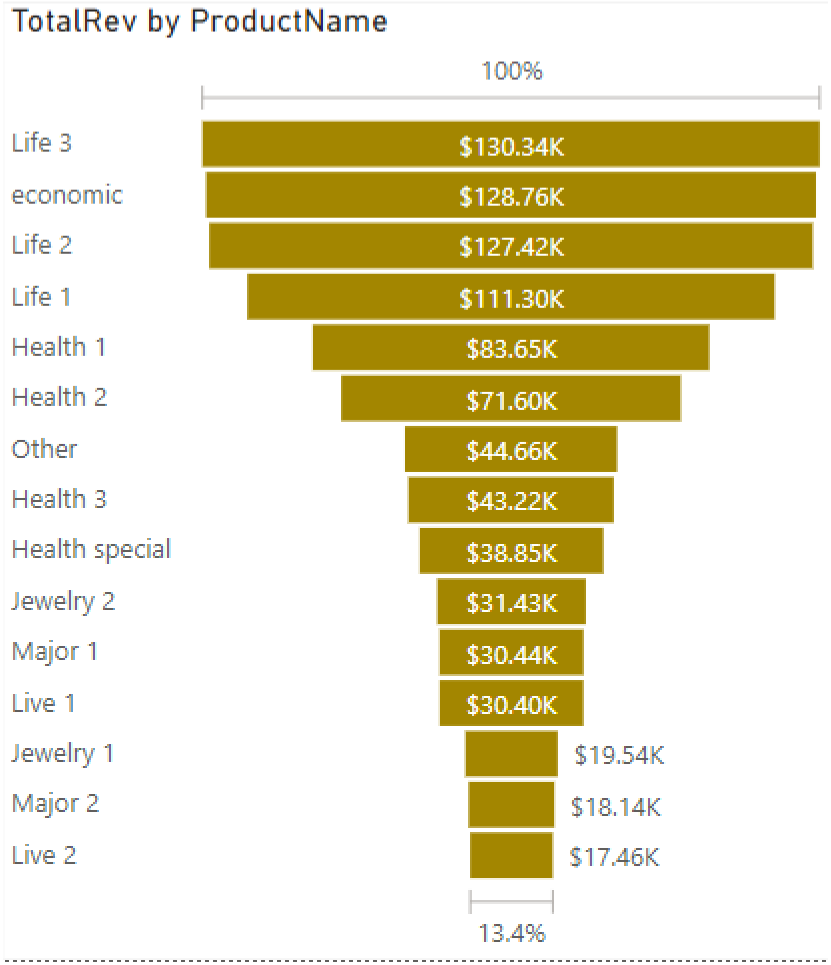
Country	DivisionName	TotalRev
India	Life Insurance	$39,580
Switzerland	Life Insurance	$26,856
Italy	Life Insurance	$24,743
Taiwan	Life Insurance	$22,317
Lithuania	Life Insurance	$21,925
Australia	Life Insurance	$21,173
South Korea	Life Insurance	$20,098
Netherlands	Life Insurance	$19,627
USA	Life Insurance	$18,572
Indonesia	Life Insurance	$17,249
Total		**$232,141**

Report 5

★ The table below shows the top 5 earning products, with the top 4 passing the $100,000 revenue mark. We recommend further statistical analyses to examine the gap in revenue between the Life 1 and Health 1 products and determine whether it is significant.



ProductName	2011	2012	2013	Total
Life 3	$46,691.60	$40,625.56	$43,026.90	**$130,344.06**
economic	$32,587.58	$40,603.14	$55,568.35	**$128,759.07**
Life 2	$42,699.34	$42,376.67	$42,339.00	**$127,415.01**
Life 1	$33,876.35	$45,934.72	$31,488.66	**$111,299.73**
Health 1	$31,357.27	$18,843.52	$33,444.94	**$83,645.72**
Total	**$187,212.13**	**$188,383.61**	**$205,867.85**	**$581,463.59**

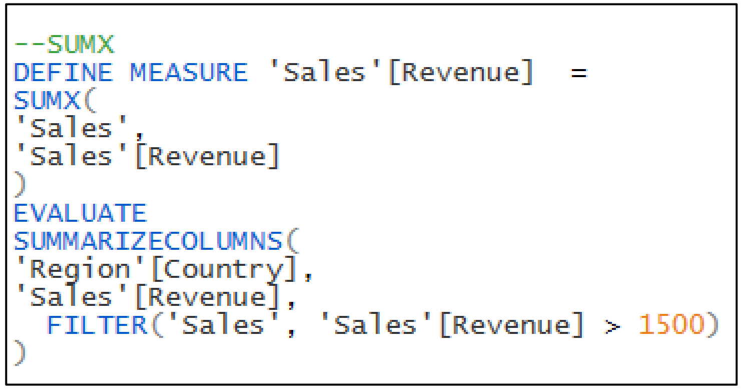


TotalRev by ProductName

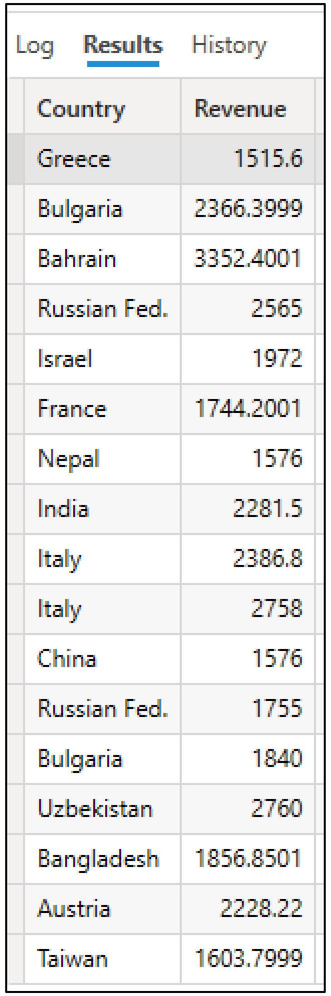
Report 6

DAX filtered table

In ‘DAX studio’ we used the ‘SUMX’ function to create a filter that shows the revenue from each country. The purpose of this filter is to return the sum of revenue evaluated for each row in the table that is greater than $1500.



```
--SUMX
DEFINE MEASURE 'Sales'[Revenue]  =
SUMX(
'Sales',
'Sales'[Revenue]
)
EVALUATE
SUMMARIZECOLUMNS(
'Region'[Country],
'Sales'[Revenue],
  FILTER('Sales', 'Sales'[Revenue] > 1500)
)
```



Country	Revenue
Greece	1515.6
Bulgaria	2366.3999
Bahrain	3352.4001
Russian Fed.	2565
Israel	1972
France	1744.2001
Nepal	1576
India	2281.5
Italy	2386.8
Italy	2758
China	1576
Russian Fed.	1755
Bulgaria	1840
Uzbekistan	2760
Bangladesh	1856.8501
Austria	2228.22
Taiwan	1603.7999

Thank you for your attention!