BME AUT Automatizálási és Alkalmazott Informatikai Tanszék

Documentation

Business Intelligence Laboratory
[Semester: Spring 2024]

Competitive Growth Analytics

Haseeb, Ahmed - (ILC918)
haseebravian756@gmail

MŰEGYETEM 1782

Introduction

I performed Analysis of internal sales, in contrast with market competitor sales of a fictional manufacturing company "**Sintec**". Following are the main features of analysis:

1. Revenue and Sales share by Top 5 Manufacturers
2. Sales Trend by Year and Month by all or any selected manufacturer
3. %Growth as of previous year vs Revenue
4. Revenue/Sales by operating countries
5. Revenue, Previous Year Revenue and %Growth Analysis by Product categories and operating countries

All the features support manufacturer and Time axis dynamic filters

Technologies Used and Solution files:

The source data file, SSIS project file, Power_bi file, and python code is placed in GitHub repository and can be accessed using the link:

https://github.com/haseebravian756/Business-Intelligence-Competitive-Growth-Analytics

For Exploratory Data Analytics, Google Co-lab code can be viewed at this link:
https://colab.research.google.com/drive/13yd-fc7r4eM5In_EMXPZMlUF1slR2pti?usp=sharing

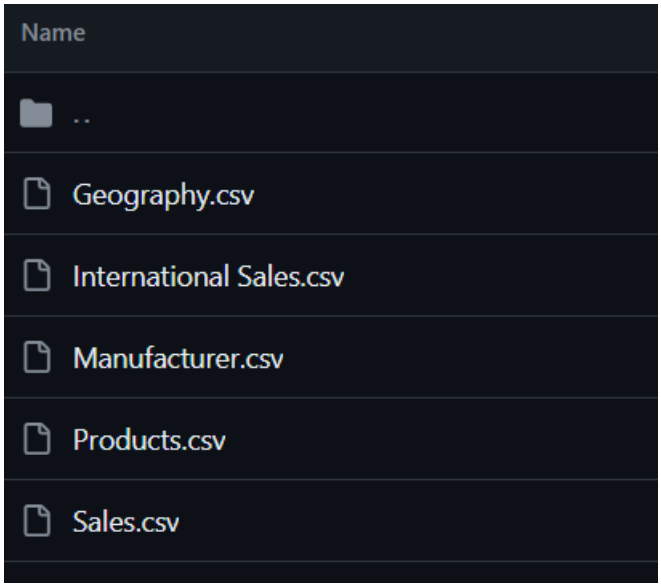
- Raw input data in .csv format
- MSSQL Database (Data Storage)
- MS SQL Server Management Studio
- SQL Server Integration Services (SSIS)
- Power Query
- Microsoft Power BI (Reporting and Visualization)
- Data-science applications:
    - Python
    - Pandas
    - Seaborn
    - Matplotlib

Source Raw Data:

- Revenue data sources for USA (Sales.csv)
- Revenue/Sales data for all other countries (international Sales.csv)

  A final fact table will be created for all operating countries of the company as a part of transformation.

- Geography Data: City, State, Zip Code related data (Geography.csv)
- Product Information Data (Products.csv)
- Manufacturing companies' information data (Manufacturer.csv)
- Calendar table (will be created in Power BI) for filtering and sorting the data using a central calendar.



ELT

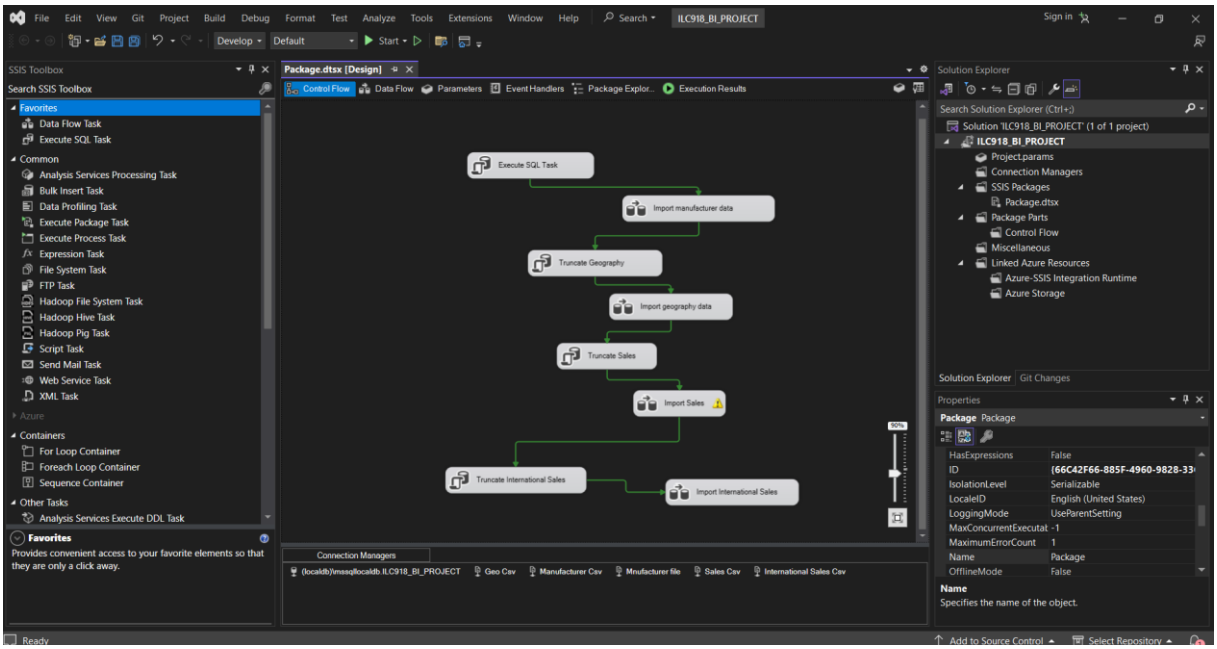
(Extraction-Load-Transform)

Extraction and Load:

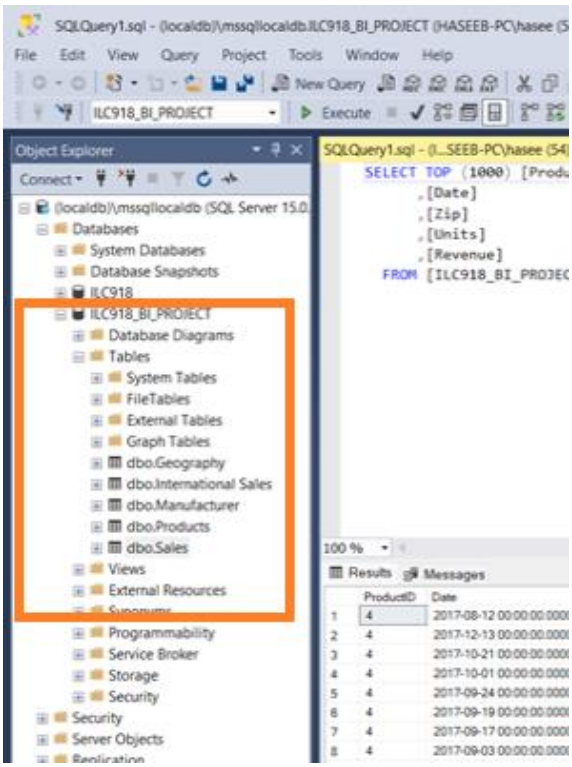
For ELT jobs, **Microsoft SQL Server Integration Services** is used.

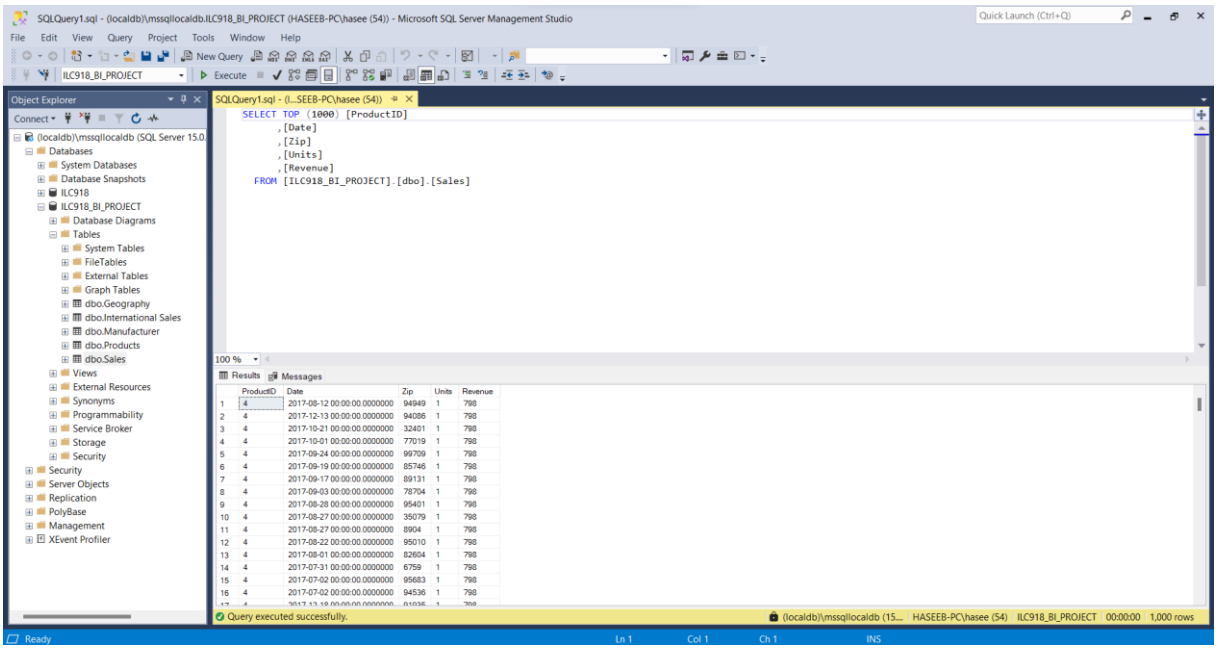
The SSIS workflow:

- extracts data from .csv files
- Performs initial data validation (data type checks and sorting)
- Performs connection with local SQL server database
- Performs truncation of data in respective database table
- Loads sorted data into the table.



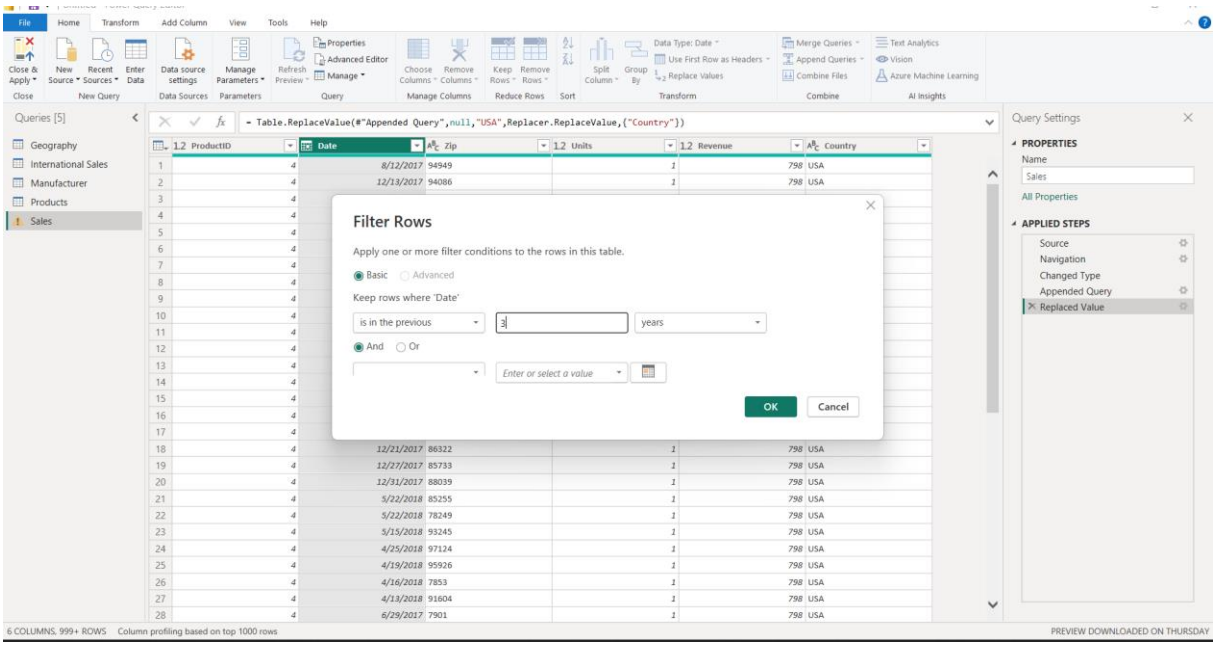
Final Loaded data into SQL server database:



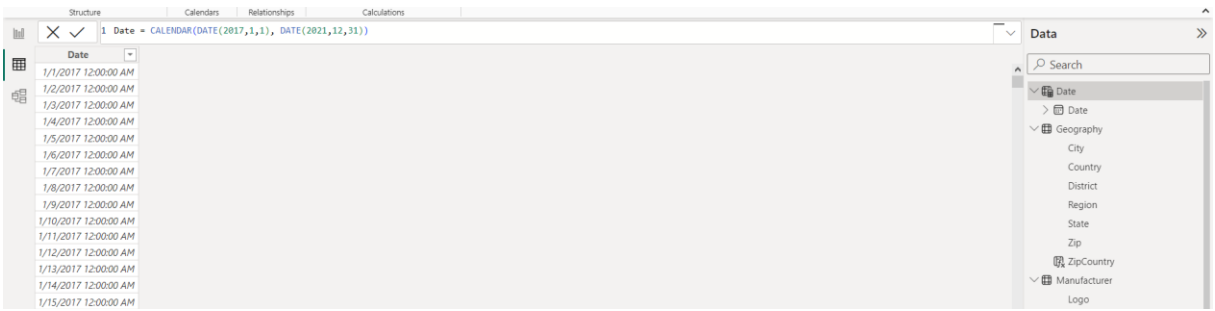


Data Transformation using Power Query:

- Power BI loads data from SQL Database using Power Query
- Data Cleanup (Null values, Fill-Down, Promoting Headers, Removing top rows) is performed
- Data Split based on delimiters (in Products Table) , Currency name and amount was split into two separate columns
- Union operation on USA Sales and International sales data to make one big sales table
- For data exploration , a subset of data was obtained (previous 3 years)



- Created a central calendar table for synchronized time filtering , sorting etc.



Power-BI Data Analysis Expression (DAX):

DAX is used to create dynamic calculation measures for enhanced reporting.

- Previous Year sales as of selected year:

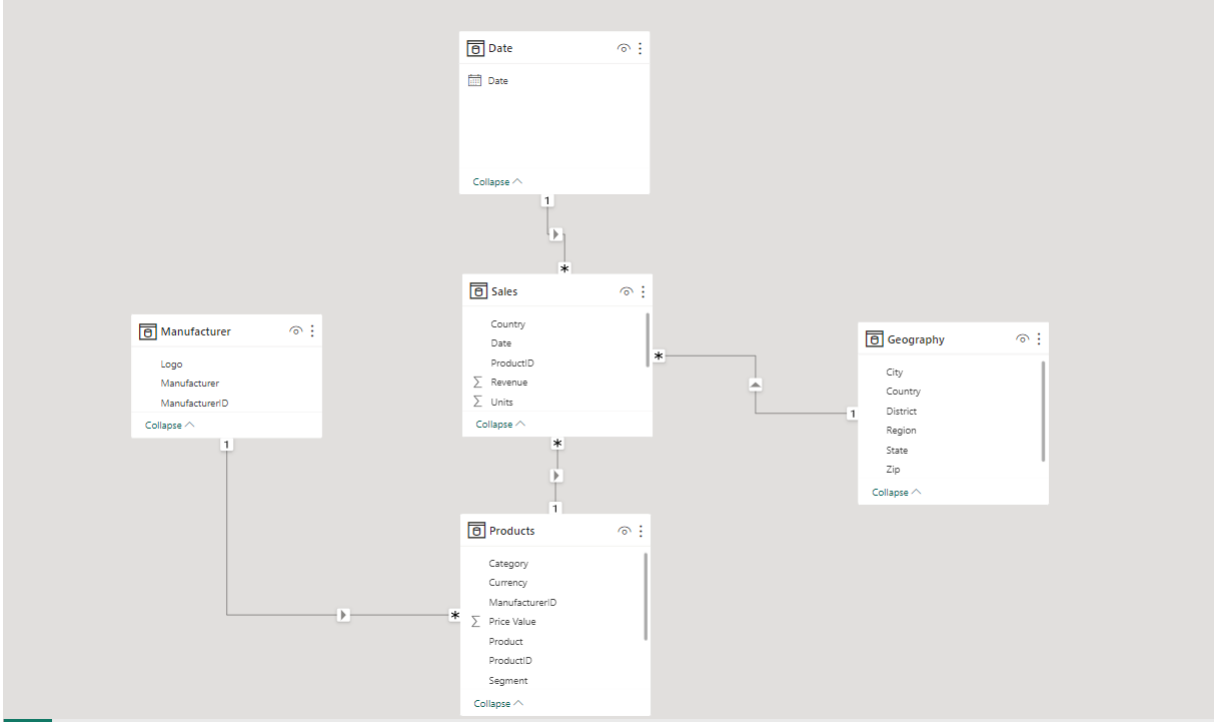
```
LY_Sales =
CALCULATE(
    sum(Sales[Revenue]),
    SAMEPERIODLASTYEAR('Date'[Date])
)
```

- % Growth as of previous year:

```
Growth =
DIVIDE(
SUM(Sales[Revenue])-[LY_Sales],[LY_Sales]
)
```

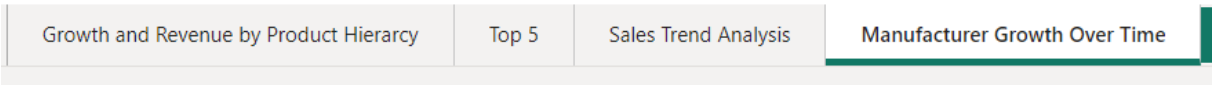
Data-Model in Power BI:

1. Efficient Data Exploration
2. Improved Performance
3. Accurate Reporting
4. Easier Future Maintainability



Reporting and Visualization:

In Power Bi Dashboard, there are 7 different visualizations/reports divided into 4 different tabs.

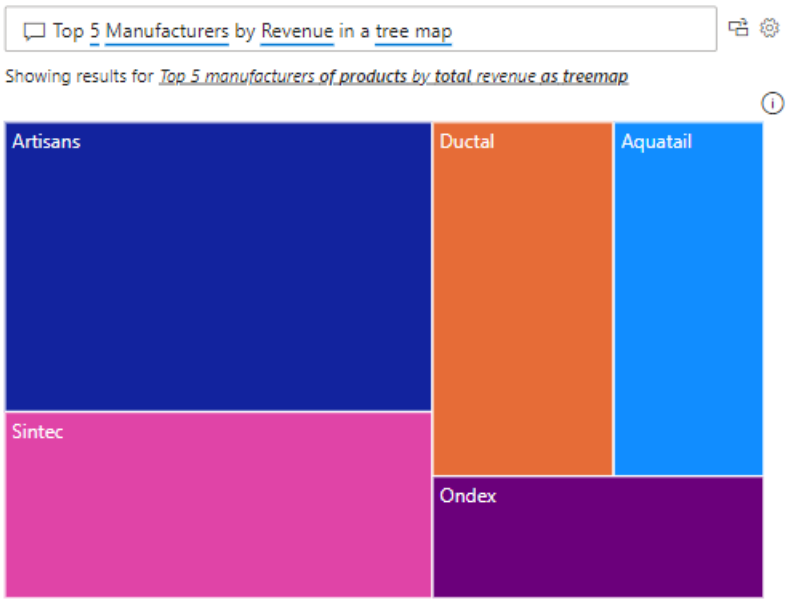


I tried to uncover following and more business insights:

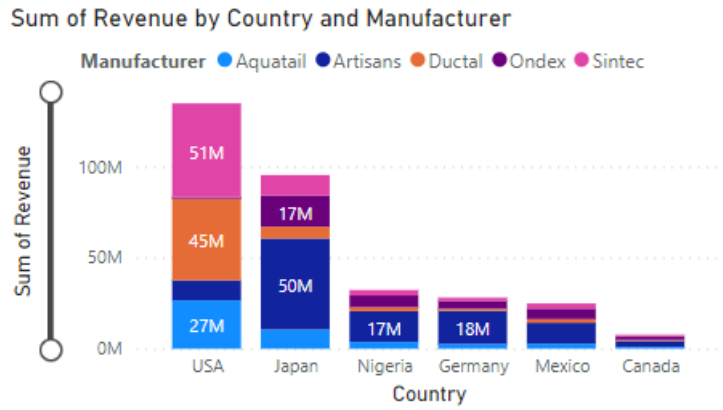
1. **Who are the top competitors generating the most revenue?**
2. **Best performing segments and products**
3. **Growth over time**
4. **Sales compared to previous year.**
5. **Which manufacturer generated the most and the least revenue in a certain time?**
6. **How much is % Growth in Revenue (as of previous year)**

Dynamic Q&A feature of Power BI:

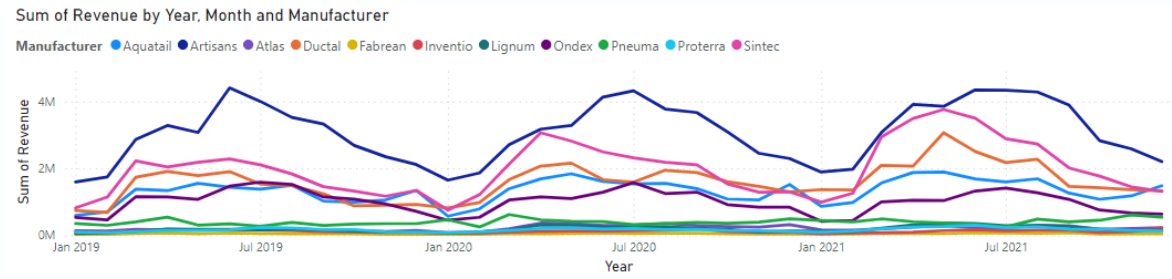
Automatically generate visual based on the question asked and type of visual. Following Tree Map was generated as a response to the prompt “”



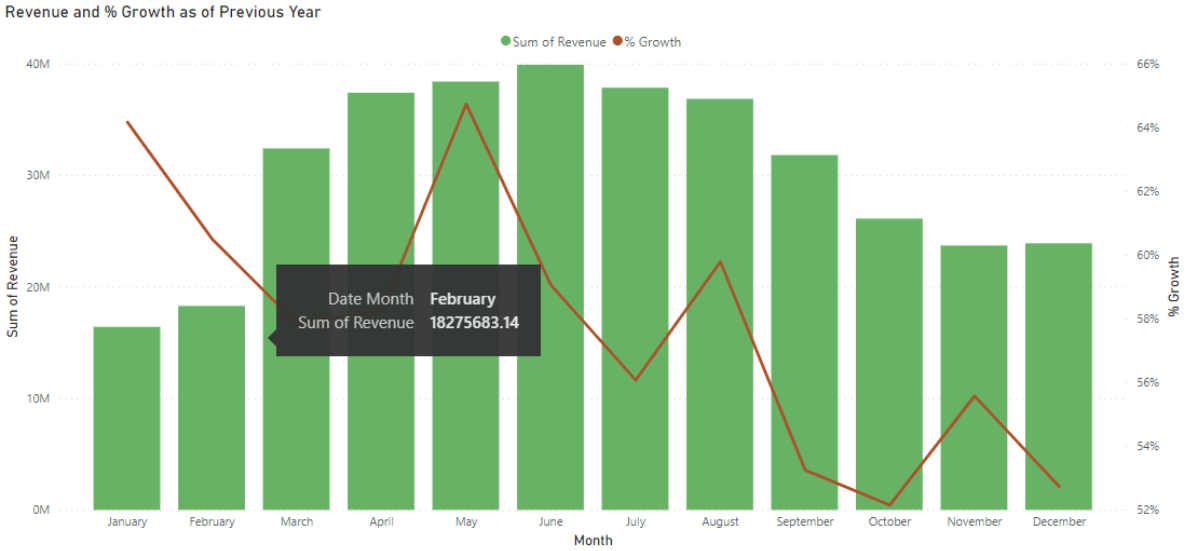
Revenue across Countries by Top 5 Manufacturers



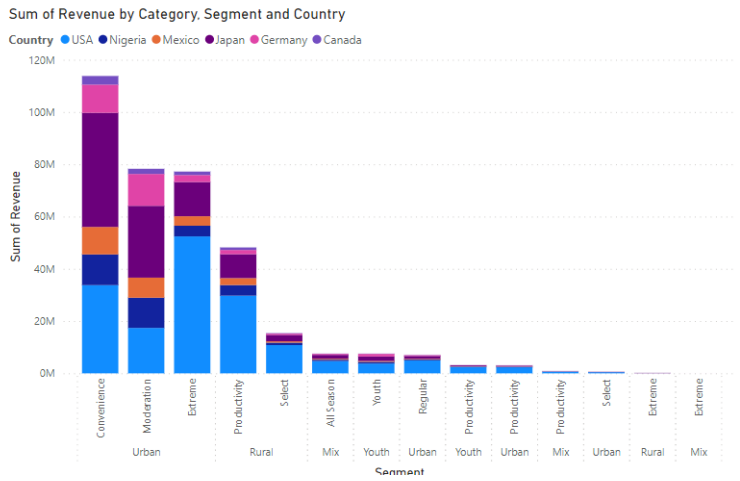
Sales Trend by Year and Month by all or any selected manufacturer



% Growth in Revenue as of previous year



Sum of Revenue by Category, Segment and Country:



Revenue, Previous Year Revenue and %Growth Analysis by Product categories and operating countries

Country	Revenue	LY_Sales	Growth
Canada	**8.15M**	**$5.14M**	**58.63%**
Germany	**29.65M**	**$19.53M**	**51.76%**
Mix	0.41M	$0.25M	62.05%
Rural	2.13M	$1.48M	43.93%
Urban	26.04M	$17.13M	52.07%
Youth	1.06M	$0.67M	57.24%
Japan	**101.19M**	**$66.20M**	**52.85%**
Mix	1.59M	$0.96M	66.08%
Rural	11.70M	$7.93M	47.60%
Urban	85.82M	$56.01M	53.22%
Youth	2.08M	$1.31M	59.17%
Mexico	**26.18M**	**$16.48M**	**58.82%**
Mix	0.37M	$0.23M	57.66%
Rural	3.16M	$2.19M	44.15%
Urban	22.20M	$13.78M	61.14%
Youth	0.45M	$0.28M	60.64%
Nigeria	**34.36M**	**$23.07M**	**48.98%**
Mix	0.59M	$0.37M	56.79%
Rural	5.04M	$3.64M	38.58%
Urban	27.91M	$18.52M	50.72%
Youth	0.83M	$0.54M	53.65%
USA	**163.36M**	**$99.56M**	**64.08%**
Mix	5.32M	$3.26M	63.36%
Rural	40.55M	$25.82M	57.05%
Urban	111.33M	$66.86M	66.50%
Youth	6.16M	$3.62M	70.07%
Total	**362.89M**	**$229.99M**	**57.79%**

Exploratory Data Analysis

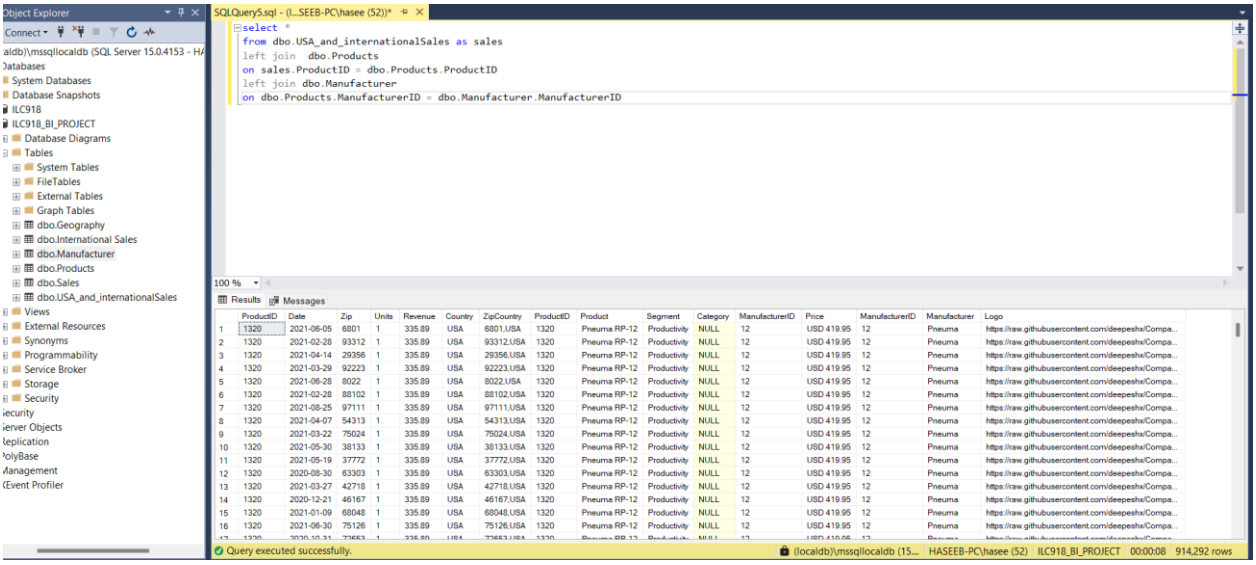
Python

For Exploratory Data Analytics, Google Co-lab code can be viewed at this link:
https://colab.research.google.com/drive/13yd-fc7r4eM5In_EMXPZMlUF1slR2pti?usp=sharing

Libraries Used:

- Python
- Pandas
- Seaborn
- Matplotlib

For analysis, I created one big facts table joined with all the mapping tables in MSSQL management studio, and saved results as .csv file to be consumed by Pandas (in Google Colab)



The Query:

```
select *
from dbo.USA_and_internationalSales as sales
left join  dbo.Products
on sales.ProductID = dbo.Products.ProductID
left join dbo.Manufacturer
on dbo.Products.ManufacturerID = dbo.Manufacturer.ManufacturerID
```

Data Overview

The data consists of 16 columns, reflecting different aspects of sales transactions. The dataset is primarily focused on transactions within the USA, with detailed product information and pricing. Key columns include ProductID, Date, Revenue, Country, Category, and Manufacturer details.

Objectives

The main objectives of the analysis were to:

1. Understand the distribution and characteristics of key variables like revenue and units sold.
2. Identify missing values and data inconsistencies.
3. Explore the data's categorical variables and their distributions.
4. Examine trends over time and correlations between different variables.

Methods and Results

Data Loading and Inspection

The dataset was loaded using pandas, and initial exploration was conducted to understand the structure and size:

```
import pandas as pd

data = pd.read_csv("Facts_Table_For_EDA.csv")

# Head function tells high level information about the data frame

data.head()

data.shape
```

Key Findings:

- The 'Category' column contains several null values.

- The dataset comprises 16 columns and multiple rows indicating individual transactions.

Data Cleaning and Validation

Missing values in the 'Category' column were noted for later treatment in visualization tools like Power BI. The 'Date' column was converted to datetime format to facilitate time-based analysis:

```
data['Date'] = pd.to_datetime(data['Date'])
data.dropna(subset=['Date'], inplace=True)
```

Descriptive Statistics and Data Distribution

Descriptive statistics provided insights into central tendencies and variability:

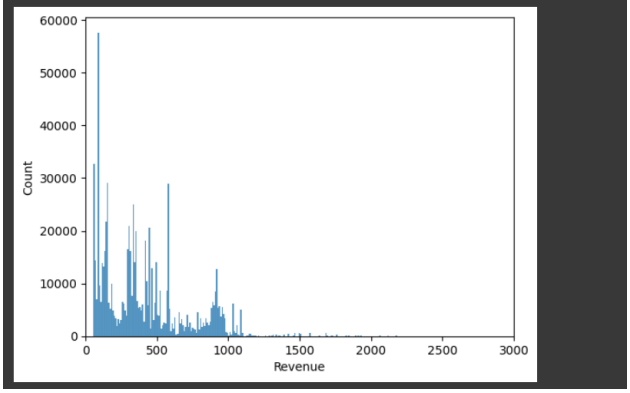
```
#Descriptive Statistics of the data

data.describe()
```

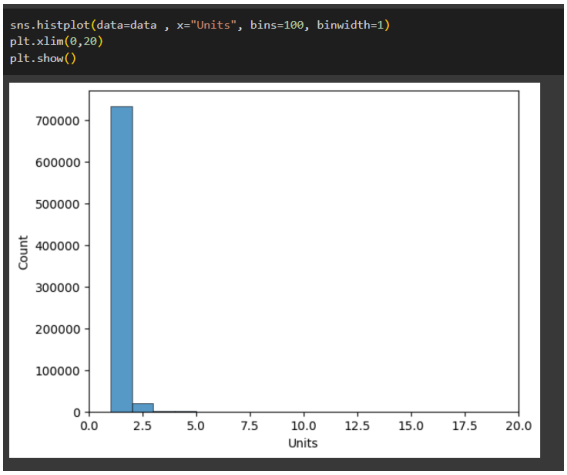
	ProductID	Units	Revenue	ProductID.1	ManufacturerID	ManufacturerID.1
count	757401.000000	757400.000000	757400.000000	757400.000000	757400.000000	757400.000000
mean	1412.786360	1.042861	412.389458	1412.788211	6.384645	6.384645
std	673.942616	0.391748	336.603787	673.941136	3.240722	3.240722
min	3.000000	1.000000	17.330000	3.000000	1.000000	1.000000
25%	791.000000	1.000000	146.950000	791.000000	4.000000	4.000000
50%	1186.000000	1.000000	341.200000	1186.000000	7.000000	7.000000
75%	2067.000000	1.000000	577.450000	2067.000000	10.000000	10.000000
max	2412.000000	76.000000	25436.250000	2412.000000	14.000000	14.000000

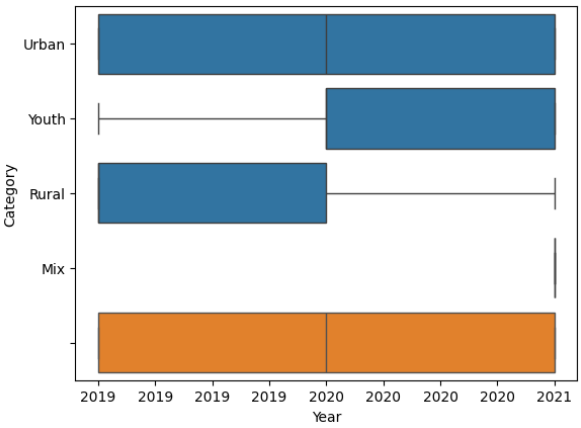
Histograms and boxplots visualized the distributions of 'Revenue' and 'Units':

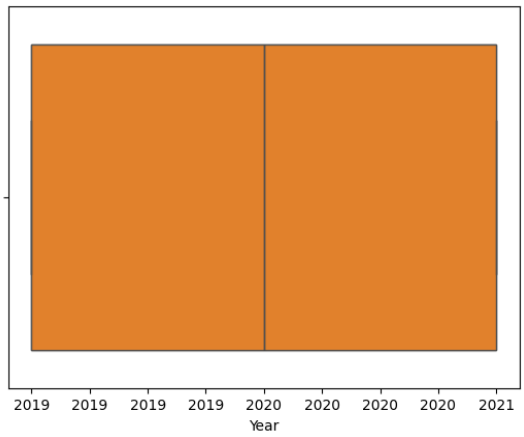
```
sns.histplot(data=data , x="Revenue",bins='auto')
plt.xlim(0,3000)
plt.show()
```




```
sns.histplot(data=data , x="Units", bins=100, binwidth=1)
plt.xlim(0,20)
plt.show()
```







Revenue showed a significant concentration at lower values, while the spread of units sold varied, with most sales involving few units.

Categorical Analysis

The distribution of categories and segments was examined to understand the product mix:

```
# Categorical columns : Lisitng out data points for each category
data.value_counts('Category')
```

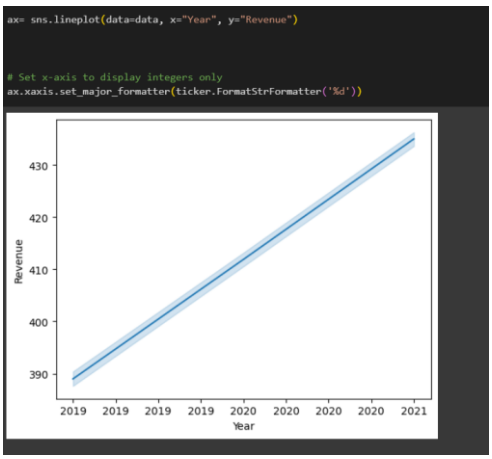
```
Category
Urban    20554
Rural     6811
Youth     2816
Mix          3
Name: count, dtype: int64
```

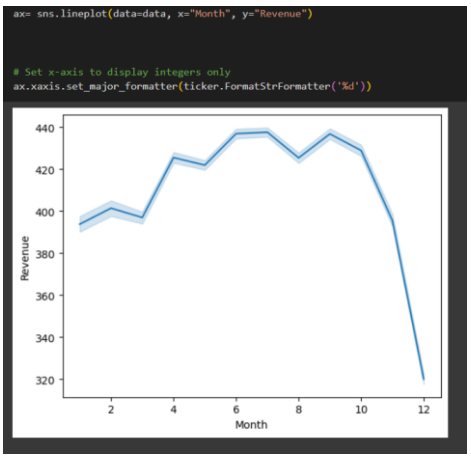
```
# Categorical columns : Lisitng out data points for each category
data.value_counts('Segment')
```

```
Segment
Productivity    271159
Convenience     175805
Extreme         141031
Moderation       76033
Select           38171
Youth            25584
All Season       15788
Regular          13829
Name: count, dtype: int64
```

Temporal Trends

Sales data was analyzed over time to identify patterns:





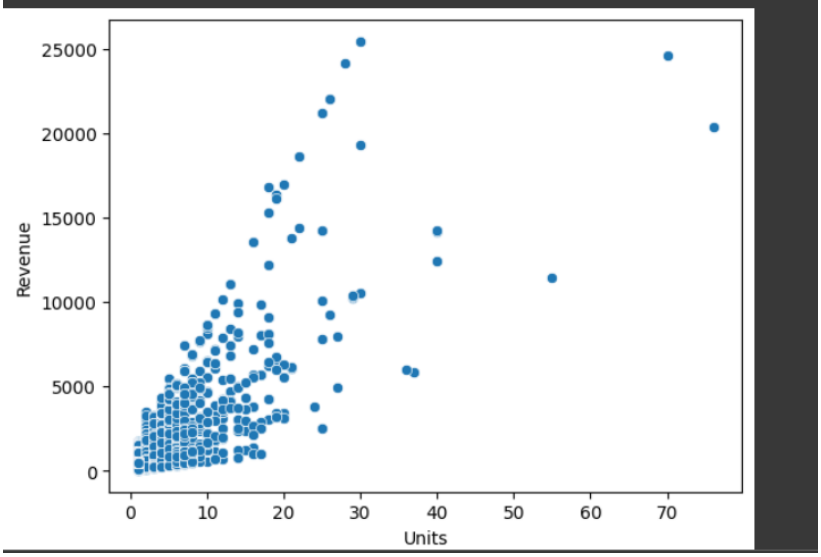
This analysis highlighted trends in sales revenue across different years and months.

Correlation Analysis

Correlations between numerical variables were explored to identify relationships:

```
ax= sns.scatterplot(data=data, x="Units", y="Revenue")


# Set x-axis to display integers only
ax.xaxis.set_major_formatter(ticker.FormatStrFormatter('%d'))
```



This revealed relationships and trends in how different time periods and sales units interact.

Exploratory Data Analysis Conclusion:

The exploratory data analysis provided valuable insights into the sales dynamics, helping identify key areas for further detailed analysis and potential areas of improvement in sales strategies. Patterns over time and correlations between sales metrics offer actionable intelligence for business strategy adjustments.