Summary of project Documents, by order

(1) Dataset :
Original file of raw dataset

(2) SQLite :
SQL script file for data exploration and basic Understanding.

(3) Python scripts :
for data cleaning, analysis and visualization of customer purchase data.

(4) Dataframe - from Python :
This is the csv file that was extracted after cleaning and transformation by Python, so as to be furtherly explored.

(5) Sales visualization :
Power BI dashboards, showcasing customer Trends and insights.

(6) Final Report (here in this PDF file, next page) :
Includes overview and the main Findings and recommendations to this project. However, some additional insights and suggestions are stated and highlighted inside the file of Python script as comments per each related code.

Final Report

Data Exploration & Understanding (Initial remarks)

First upon exploring the main Dataset file and to better understand the given Columns ; I was checking some points for Validation :

- ORDERNUMBER VS ORDERLINENUMBER

  In order to figure out the difference between them and to find the correct indication of ORDERLINENUMBER ; I checked on random sample of OrderNumbers VS their OrderLineNumbers:
  - ➢ It's found that each OrderNumber includes several ORDERLINENUMBERs that indicate one LineNumber for each PRODUCTCODE in this OrderNumber.
  - ➢ So 1 ProductCode cannot be repeated inside the same ORDERNUMBER.

- It's also noticed that Year 2005 does not include the data till end of year, since Quarter 3,4 for example are missing in Data, this can be for example because 2005 had not ended yet upon making this analysis.

- "TERRITORY" column is noticed to include "Japan" among the territories, although Japan is not a territory and is known to belong to the "APAC" Territory.
  Also the written "Japan" Territory corresponds to Only 3 Countries : Singapore, Philippines, Japan… which are all known to be part of the "APAC" Territory countries.
  so "Japan" should be replaced to be "APAC" .

- **Checking for Blank cells (NA) in the data**, it's found that "STATE" is the column which has missing values, and this is Logic because States are usually in the USA but other countries like France for example; do not use the term "State".

- Meanwhile, the 'NA' mentioned in the Column of "Territory" stands for "**North America**" territory and is not indicating any blanks. This was concluded when looking into the corresponding Countries and found that they're all "USA, Canada" .
  So later on I shall make sure that "NA" in "TERRITORY" is not read as Null.

Conclusions & Recommendations with details (Power BI section)

Average Order Value (AOV) when it's analyzed along time, it's found that the highest AOV was on Feb-2003 of value 47K. And when noticing that this month had only 3 Orders, so this means that they were of high values and for 3 high value customers.
Even if these customers are of low-frequency purchases, they still have high potential to make large Purchase order values. So this is a point of improvement, that its recommendation will be advised in the Customer analysis section below.

Product performance over time, Looking into Product lines trends along time, it's noticed to occur that some Categories are not sold at all at some months along the given years. Except for the "Cars" category; it's always sold in each month along the Years. Next in order of frequent selling, are the Lines of "Motorcycles" and "Trucks & Buses".

➢ Recommendation (no.1):

- **"Cars"** is the winning constantly sold category, so the management can better focus on this Line and might expand its items and types, and invest in developing this category.
- Also **"S18_3232"** Product code, since it's the most sold item then it's advised that management would also focus on this item and expand it. They can use it in "**Product Bundle offers**" to boost sales of other least selling items like in "Train" category. (detailed in Python practice)

Analyzing Customer purchase behavior & Buying patterns (via Power BI)

- In the "Customers" page, the Customers buying behavior is clarified and analyzed in the visual "**Customers Classification by Value & Frequency**". It's found that the majority of Customers (54) are "Low-Frequency & Low-Value", and only (2) customers are "High-Frequency & High-Value".
- It's detailed in Python practice that segmentation of Customer Frequency was based on limit 14 Orders, while Customer Value segmentation was based on limit 100,000.

- **It's an Opportunity to** resolve the drawback that almost all customers are of Low-Frequency (90 customers out of 92).

➢ Recommendation (no.2)…

to management is to attract Customers to make more Orders and be more frequent buyers.
This can be done by offering them discounts on their Next purchase Order,
or by making a **"Points system"** where points increase by the number of Orders, and shall be translated to discount at a certain limit. Noha Ahmed Shehata Kawashti July 14, 2024

Analyzing trends over time to identify seasonal buying patterns or changes in customer behavior (through Python)

- **Finding is that** there's a massive shift in demand in **"November"** month of each year, where this month has the highest purchase pattern in terms of Purchase value and also Number of Orders.
- It's assumed that this can be due to "November offers and deals" like "**Black Friday**" that are usually carried out in November yearly worldwide, as a clearance sale process before end-of-year.
  For example in the USA, Black Friday is the Friday after Thanksgiving in the United States, it traditionally marks the start of the **Christmas shopping season**. Many stores offer highly promoted items at discounted prices.

➢ So Recommendation (no.3) :

It's recommended to apply strong Marketing campaigns more frequently in other months, not just once a year in October and November.

Analyzing sales and orders by Area of customers ;

Number of orders per Country shows that **USA** is the massively highest Country in the number of purchase Orders, then next in order are **France and Spain**.

Territory Sales percentage by Category : it's found the category per each territory where it has the highest Sales percentage…

- **North America** has the highest sales percentage (45%) among other areas, in the category of "**Motorcycles**"
- **EMEA** has (61%) of sales among other areas in the category of "**Trains**".
- **APAC** area is generally of weak contribution of Sales percentage among other territories, however its highest sales Percentage was (16%) in products of "**Trucks and Buses**".

Analyzing territories, "APAC" is found to have the least Purchases allover the given Time scale, to the point that it's Zero purchases in APAC at certain months (Jan,Feb,Jun,Aug,Dec-2003, Mar,May,Oct-2004).

➢ Recommendation (no.4):

"APAC" is an opportunity to increase Purchase frequency and Loyalty, so apply on this territory the strategies advised in recommendation (no.2).
As for the findings in "Territory Sales percentage by Category", it's recommended to use the same strategy mentioned in recommendation no.1 "**Product Bundle offers**".

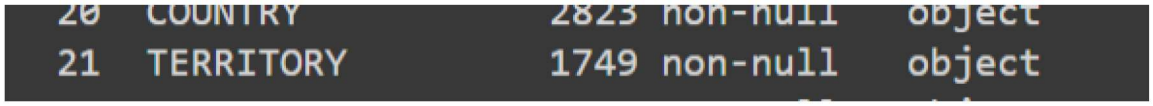
Data Exploration & Understanding : by SQL

Main Findings (SQL):

- **S18_3232** is the most frequently purchased item Product code, it's far ahead of the rest of products (Frequency is 52 times, and also highest Quantity 1774 units).
- **"Classic cars"** is the most frequently purchased item Product Line, (Frequency is 967 times, and also highest Quantity 33992 units)
- **DEALSIZE** of "Medium" is the highest percentage among other Sizes. It's assumed that Deal size is based on the Sales of this ORDERLINENUMBER (noticed upon data exploration on random sample of rows).
- **USA** has the highest Sales, while Ireland has the least sales. So it's recommended to increase Marketing campaigns in Ireland.
- **S18_3320** is the most sold item in California, and "Cars" products are the most sold.
- **EMEA (Middle East)** is the territory where "Classic Cars" were sold the most.

Python

- Discovering the data info. using Python ; it's noticed that it counts the NA (North America) in the "TERRITORY" column as Null – per below screenshot, which is wrong.
- So in order to be read as a value (Non-null record) ; then I thought it's better to change it from "NA" to "North America".



- This was successfully done by transforming the Territory from "1749 non-null" to be "2823 non-null", screenshots shown here.

From the Statistical overview ;

- I wanted to know which customer made the highest Sales transaction OrderLine (the MAX = 14082.8).
  So it's found that the Customer "The Sharp Gifts Warehouse" has the highest sales Transaction line as a part of ORDERNUMBER 10407 on 22-Apr-2005, the whole order of 12 lines has all its items "Cars". Customer is located in USA, in state of California.
- **The average (Mean) Sales** is 3553.88

Upon exploring the DEALSIZE aspect...

- (Large, Medium, Small), it's concluded to be based on the Sales value of this OrderLineNumber. I assumed that:
- Deal size "Small" (Low-value) transactions are for OrderLines of Sales value less than 3000 USD, while "Large" (high-value) transactions are for OrderLines of Sales value more than 7000 USD, if currency is Dollars.
- When checked the No. of times "Large" Deal size was most repeated per Customers; it's found that 3 Customers had the most repeated "Large" Deal Sizes, so they are **High-value customers which are :**
  "Euro Shopping Channel", "Mini Gifts Distributors Ltd", "The Sharp Gifts Warehouse".

- **A Finding** along analyzing by Python, is that Madrid City in Spain has the highest Number of Orders (31), while Munich in Germany has only 1 purchase Order.(found from Pivot table).

➢ **For Python practice section ; there are other details, insights and recommendations available inside the Python script as comments per each Related code.**