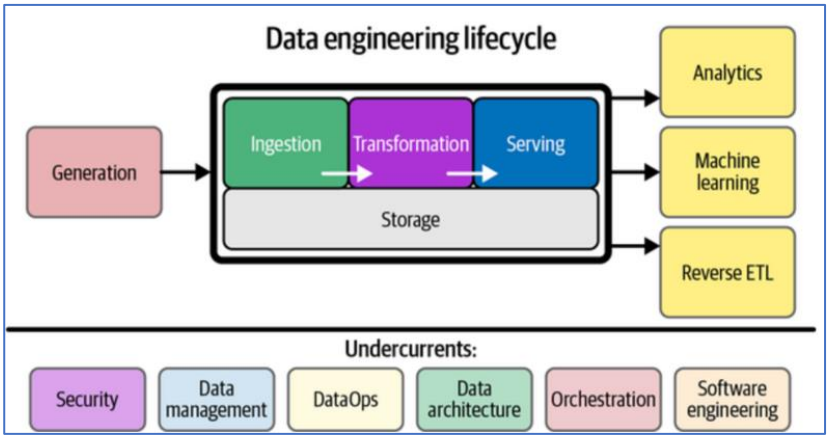
Project Deliverable Part 3

I learnt a lot from this Data Engineer course without having previous experience (just by having some limited tools, sql experience), I understood whole life cycle , background of Data Engineers does in Real time . I have come across several tools like Tableau, Airbyte, Snowflake.

Basically, Data Engineers collect, process, and store vast amounts of data securely and organized manner in the industries, makes data-driven decision-making within organizations. Below is the lifecycle of Data Engineering which involves five stages (Generation, storage, ingestion, transformation, serving data)



Data engineers are responsible for setting up and managing relational and non-relational databases (like SQL, NoSQL, and PostgreSQL), data warehouses (like Redshift and Panoply), and big data systems (like Hadoop and Spark).

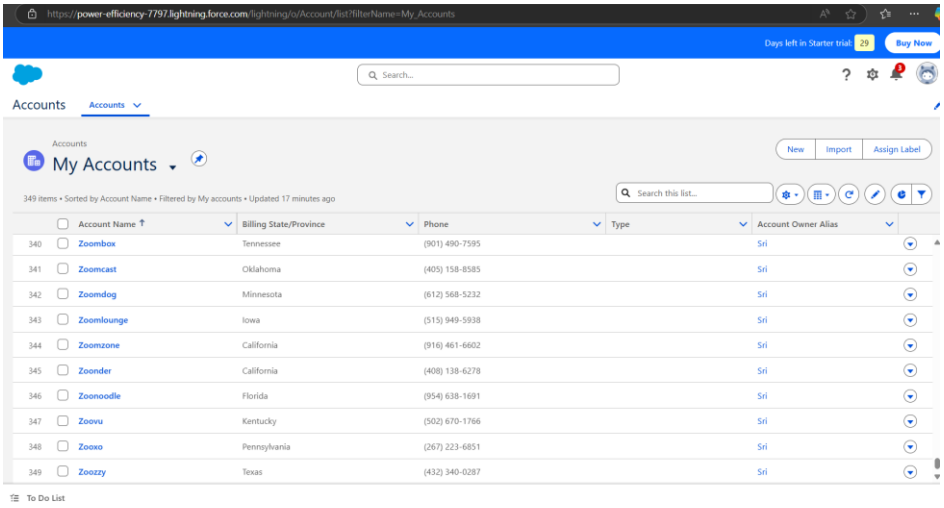
I'll start with basic concepts and how the project is handled in these stages:

Data Collection:

Everything starts with Data collection phase where the data can be from Primary (Statistical methods, Survey polls) to Secondary data collection methods (like Financial, Sales reports).

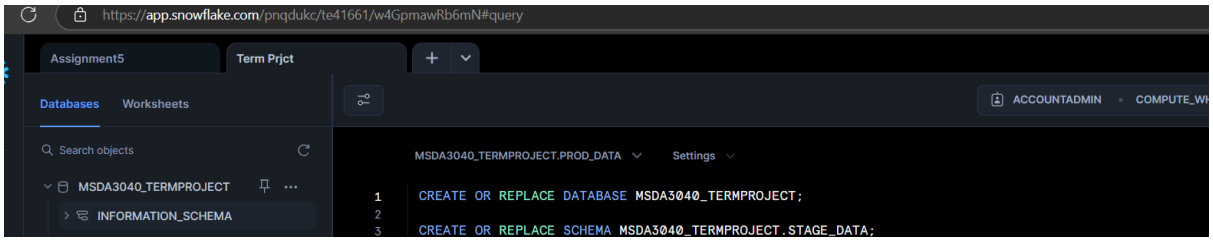
Project involvement: (Data generation, preparation)

Here 4 tables generated with random data Customer dimension, Orderdetails dimension, Product dimension, Orders fact tables from Mockaroo website.



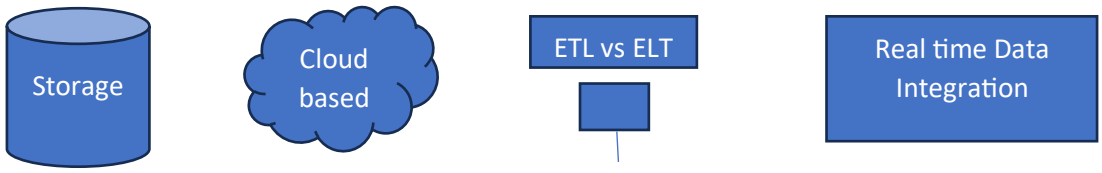
Loaded data in Salesforce

Created these tables objects in snowflake under staging for data preparation.



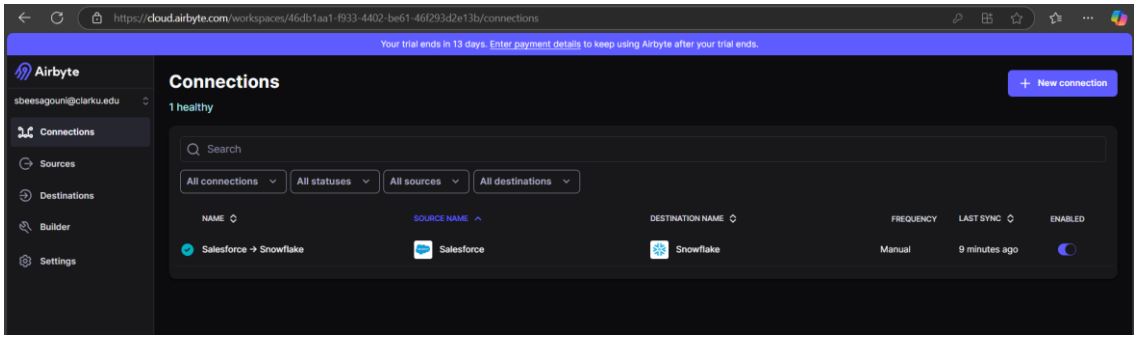
Data Integration:

Once data is collected its most likely going to be transformed and stored in a structured format in data warehouses. Integration involves in cleaning, transforming, consolidating data from various sources into one source manner. During this process we need to decide on storage solution about cloud based , ETL(pull ingestion) , some architectural factors like below should be considered.

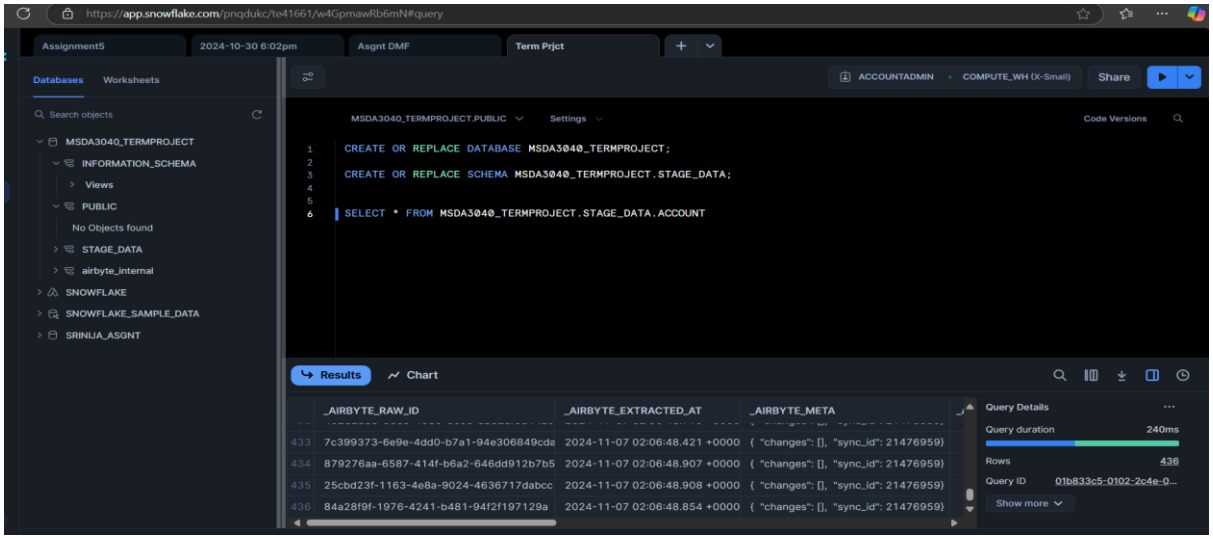


Project involvement: (Data loading, transformation)

Utilized Airbyte to extract data from Salesforce and loaded it into stage tables in Snowflake.



Airbyte connection



Data Ingestion

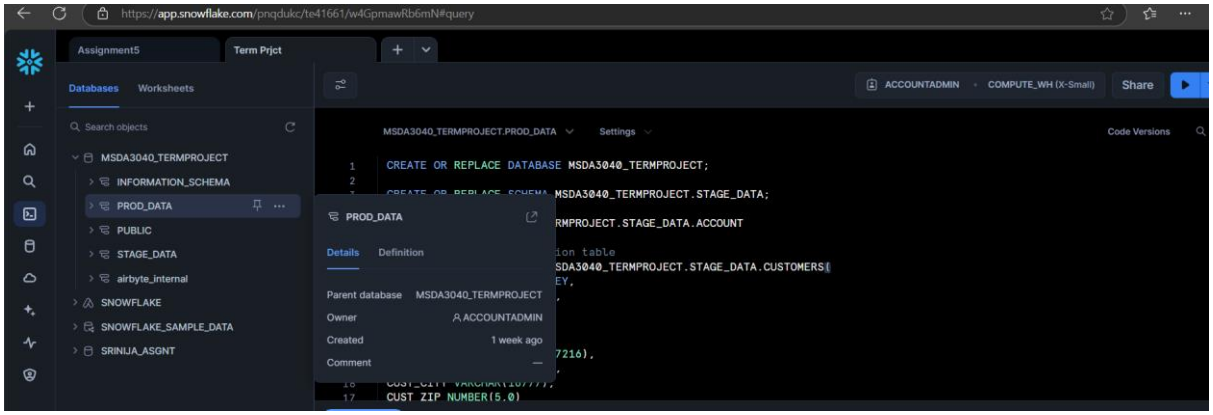
Data Cleaning:

It is the process of converting raw data into the format which is suitable for analysis. This involves some basic transformations to advanced like Merging datasets, rebuilding missing data, standardization, normalization, de-duplication, verification & enrichment, exporting data.

Project involvement: (Data cleaning, transformation)

Transformation - Merged new dataset ‘Zip codes’ with ‘Customers’ for filling missing values based on related fields. Created Profit margin column in ‘Products’ table with cost, msrp, price and some other aggregations.

Validation- Created Production schema, created similar dimension, fact tables pushed staging data to production and performed few quality checks.



Prod in snowflake

Data modelling:

Data modelling mainly focus on designing data schemas efficiently support operational and analytical needs of the business. Relationships like Facts, dimensions tables, attributes.

Data warehousing:

This is one type of data architecture, It is the storage system where it supports business intelligence and reporting , supports scalability of the data. Data warehouse contains tons of data , having structured queries with db design , some challenges can be faced is performance degradation, complex queries, security compliance.

Data analysis:

Here we apply statistical methods machine learning algorithms for analysis (kind of predicting, forecasting data), with this data insights few visualizations can be made for business requirements, stakeholders.

At last, there is one concept called pipelines which make sure the systems are in place to process data, but problems here would be maintaining robust data overtime required data governance.

Project involvement: (Data analysis, visualization)

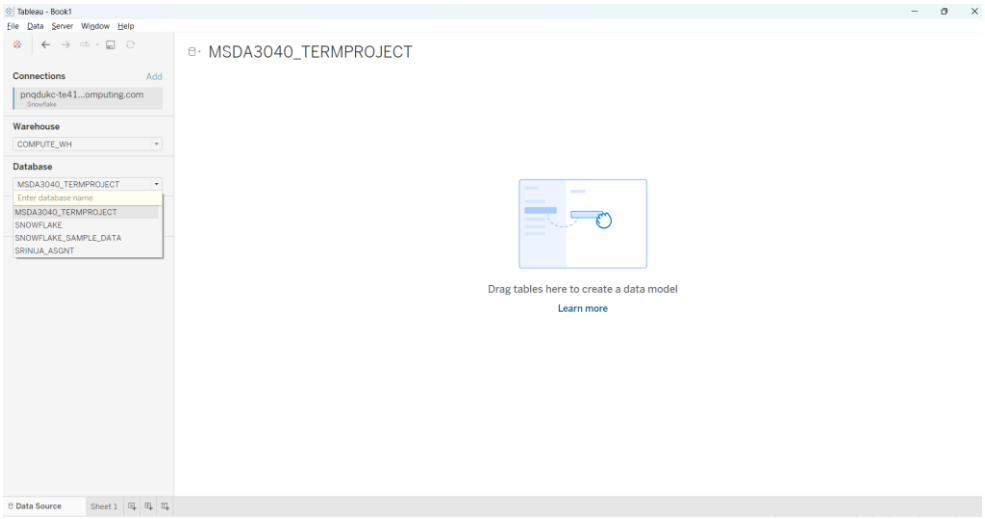
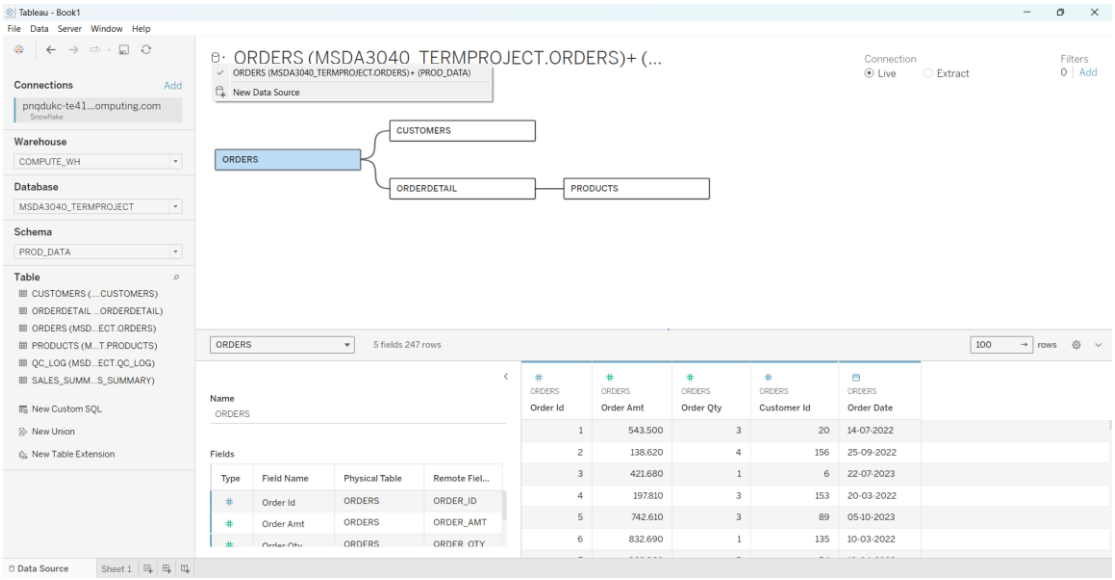
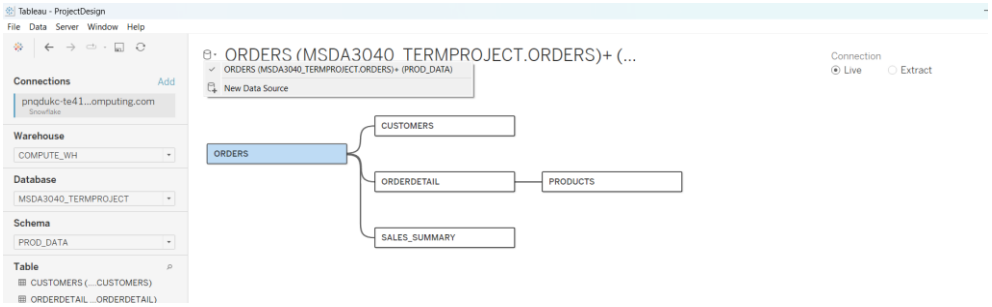


Tableau , Snowflake connection

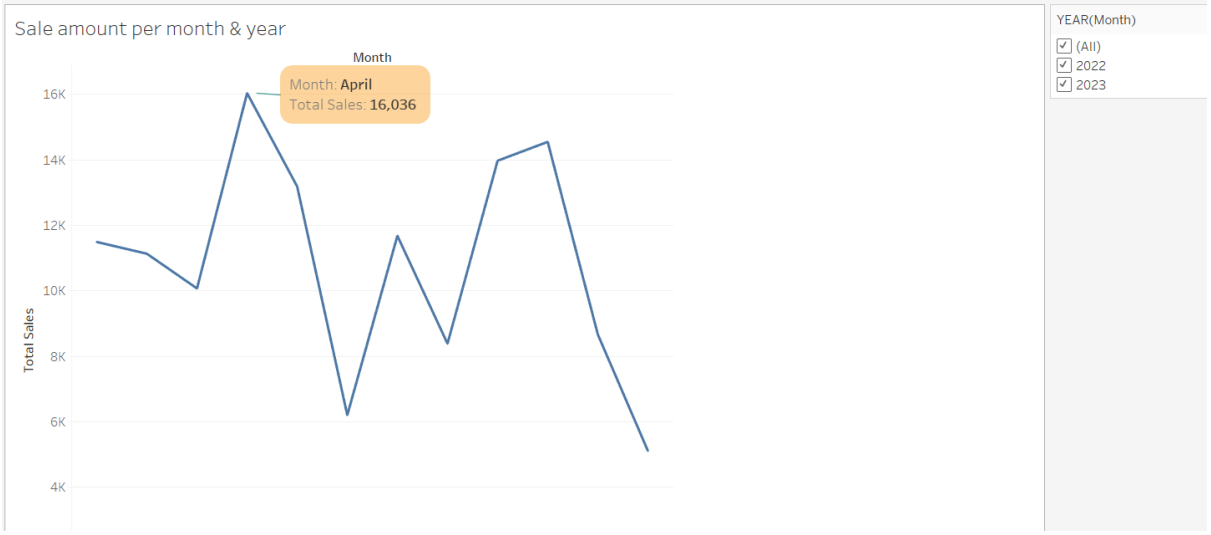


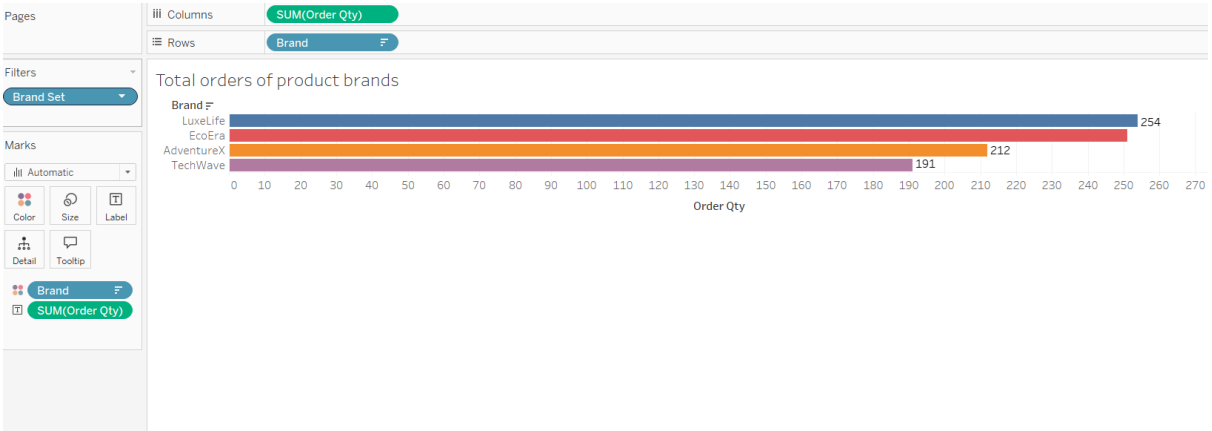
Joins in Tableau

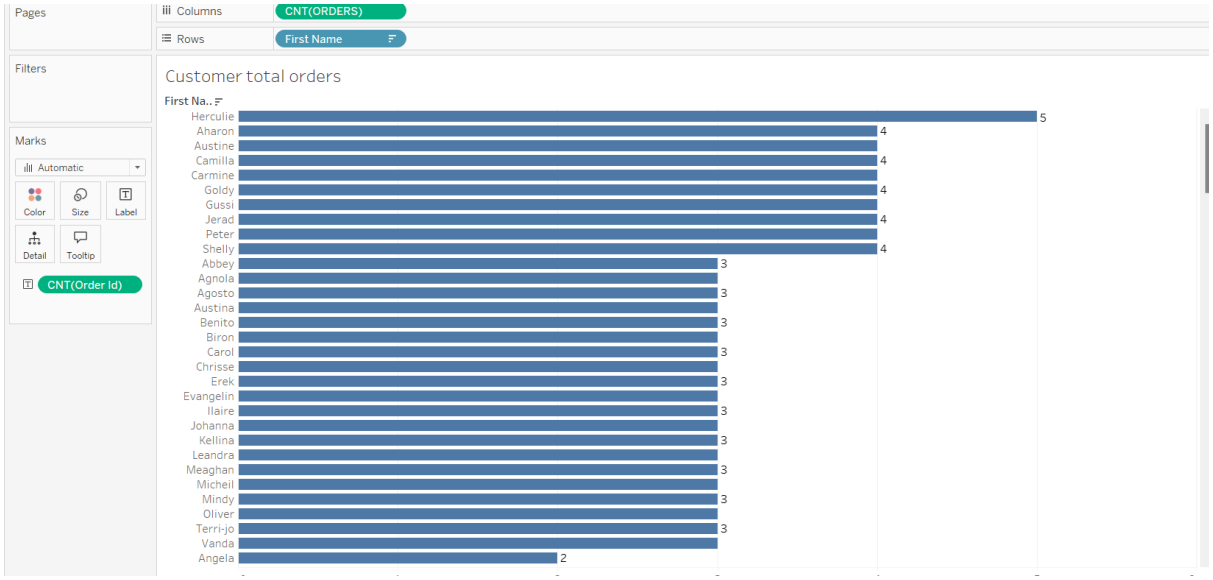


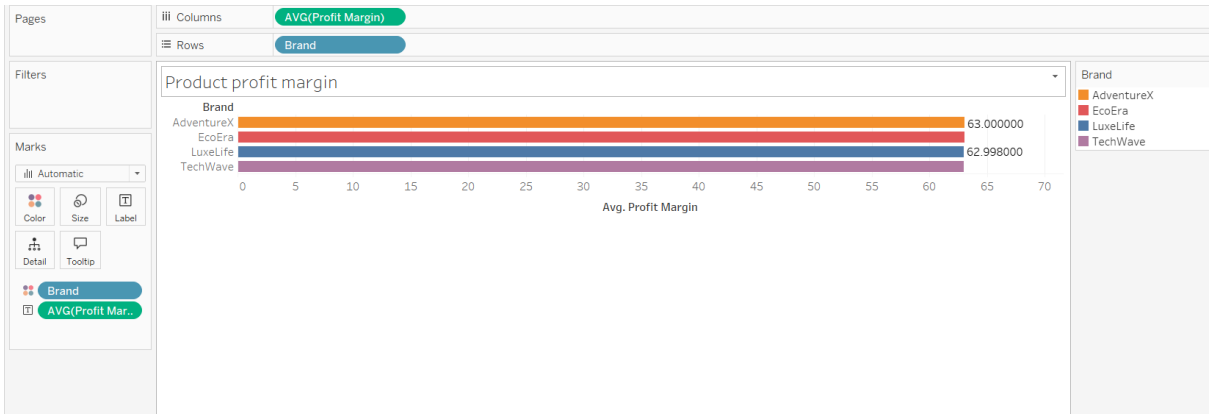
Join with additional table for monthly summary

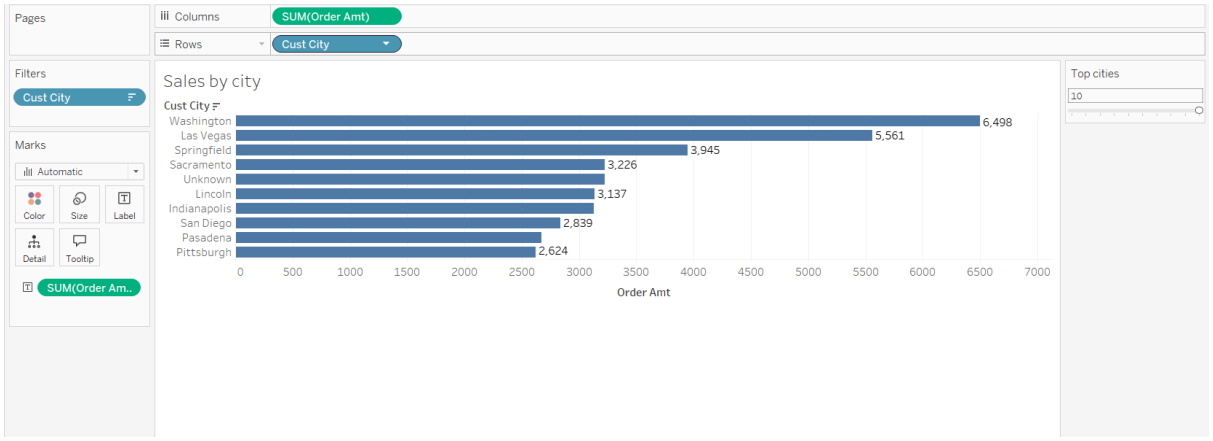
Visualizations (based on analysis):

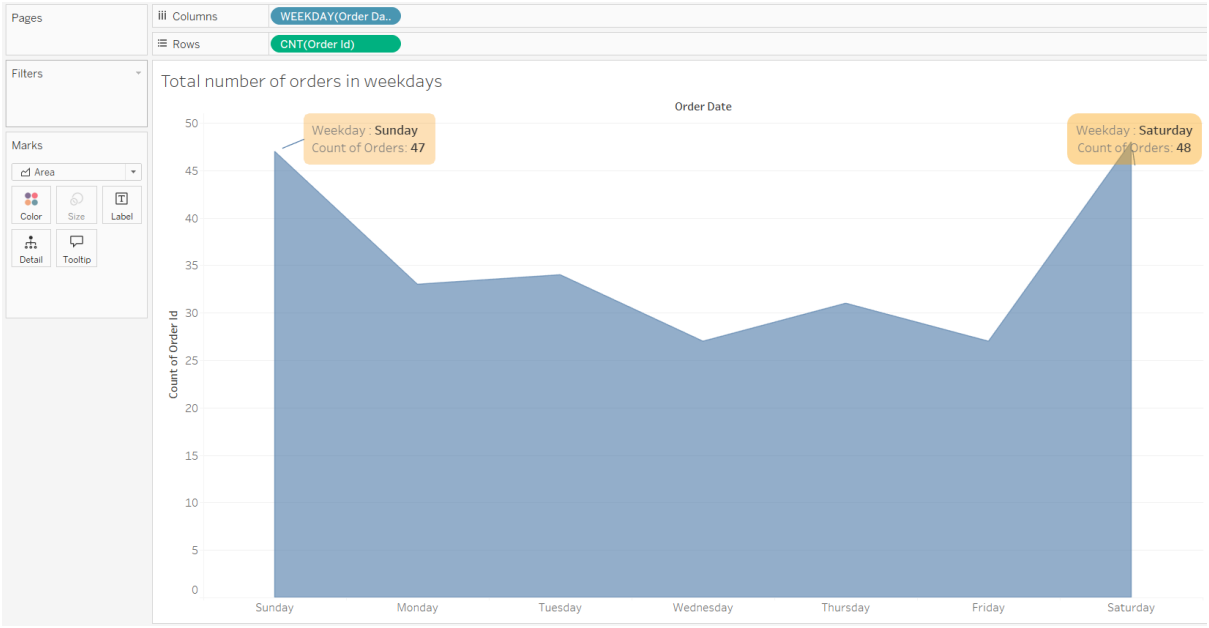




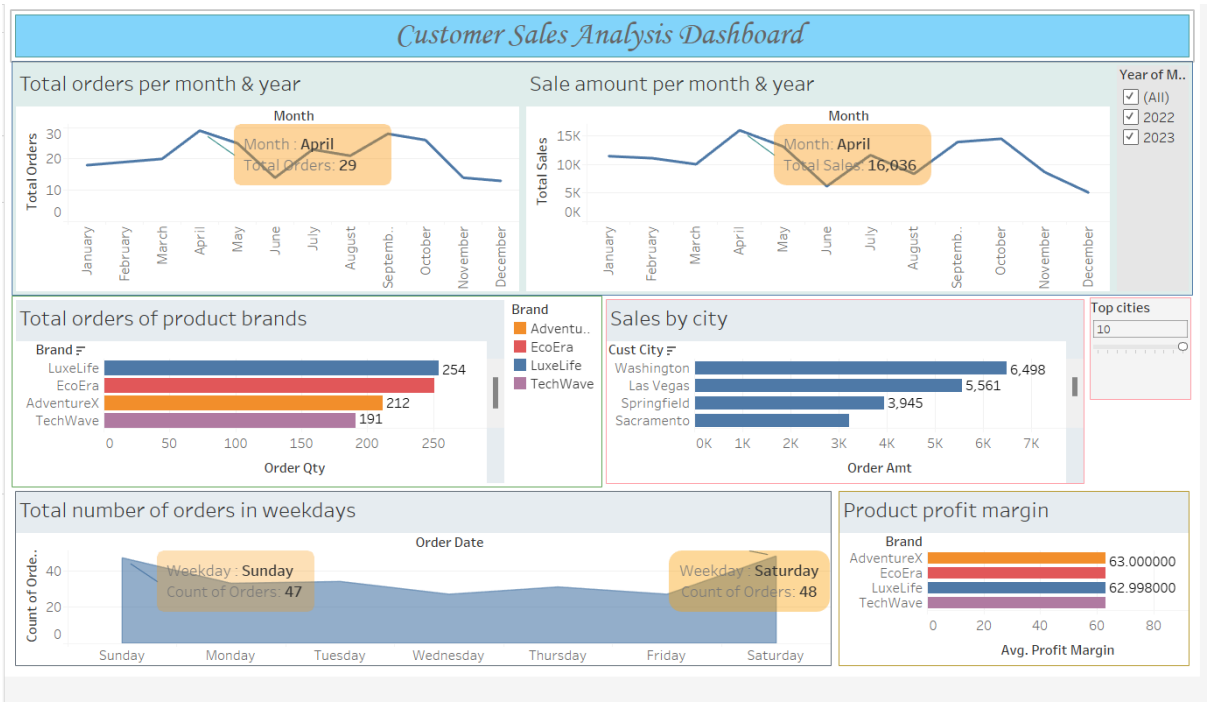

Pages
Columns
CNT(ORDERS)
Rows
First Name
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
CNT(Order Id)
Customer total orders
First Na..
Herculie 5
Aharon 4
Austine 4
Camilla 4
Carmine 4
Goldy 4
Gussi 4
Jerad 4
Peter 4
Shelly 4
Abbey 3
Agnola 3
Agosto 3
Austina 3
Benito 3
Biron 3
Carol 3
Chrisse 3
Erek 3
Evangelin 3
Ilaire 3
Johanna 3
Kellina 3
Leandra 3
Meaghan 3
Micheil 3
Mindy 3
Oliver 3
Terri-jo 3
Vanda 3
Angela 2


Pages
Columns
AVG(Profit Margin)
Rows
Brand
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Brand
AVG(Profit Mar..
Product profit margin
Brand
AdventureX 63.000000
EcoEra
LuxeLife 62.998000
TechWave
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70
Avg. Profit Margin
Brand
AdventureX
EcoEra
LuxeLife
TechWave


Pages
Columns
SUM(Order Amt)
Rows
Cust City
Filters
Cust City
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Order Am..
Sales by city
Cust City
Washington 6,498
Las Vegas 5,561
Springfield 3,945
Sacramento 3,226
Unknown
Lincoln 3,137
Indianapolis
San Diego 2,839
Pasadena
Pittsburgh 2,624
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500 6000 6500 7000
Order Amt
Top cities
10



Dashboard:



Project link:
https://public.tableau.com/views/ProjectDesign_17321639004350/CustomerSales?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

So far what I learnt was Effective Visualization Planning, Parameter Utility in Tableau, Importance of Flexibility (Every dataset and business question is unique. Designing parameters and visualizations that adapt to different contexts ensures scalability and usability.)

What surprised me is Complexity from Simple Metrics, Interplay of Relationships (Working across datasets [Customer, Order, Order Details, Product] highlights how joining the right tables unlocks deeper insights, such as tracking revenue per product across customer demographics), Customization Potential in Tableau.

What I'll Apply in Future Work is Deeper Metric Analysis, Iterative Communication, Focus on User Interaction (Incorporate parameters early when designing dashboards, ensuring the end-user has the tools to customize views based on their specific questions or thresholds.)

Thank you.