# End-to-End Real Estate Data Project

## Project Overview

This document describes the full lifecycle of a real estate system project starting from data collection and preprocessing, building an operational database, creating a Data Warehouse (DWH), performing ETL processes, analyzing data using Tableau, and finally providing recommendations for data-driven decision-making.

## Project Objectives

1. Collect reliable data from multiple sources.  
2. Build an operational database to store transactional data.  
3. Design a Data Warehouse (DWH) using a Galaxy schema.  
4. Implement ETL processes to load data into the DWH.  
5. Analyze and visualize data through Tableau dashboards.  
6. Provide actionable recommendations for improving data quality and business performance.

## Phase 1: Data Collection

- Web Scraping for listing and property

- Generate data using Python code

Listing: 15,000

Leads: 30,000

Sales: 10,000

Agent: 800

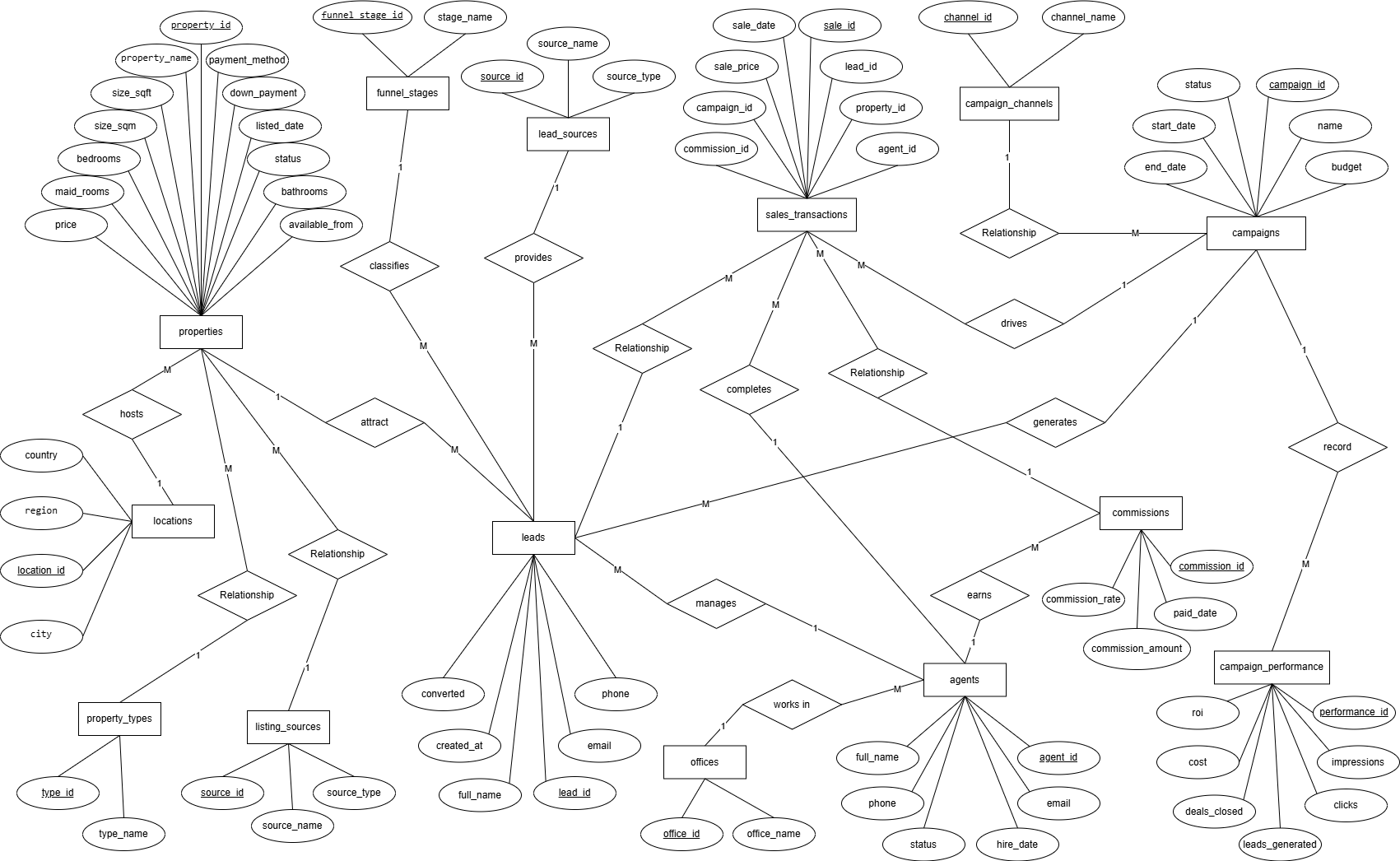
Campaigns: 50 row

## Phase 2: Data Cleaning and Preparation

1.Remove duplicates and Null rows  
2.Standardize formats such as dates, currencies, and measurement units.  
3. Handle missing values

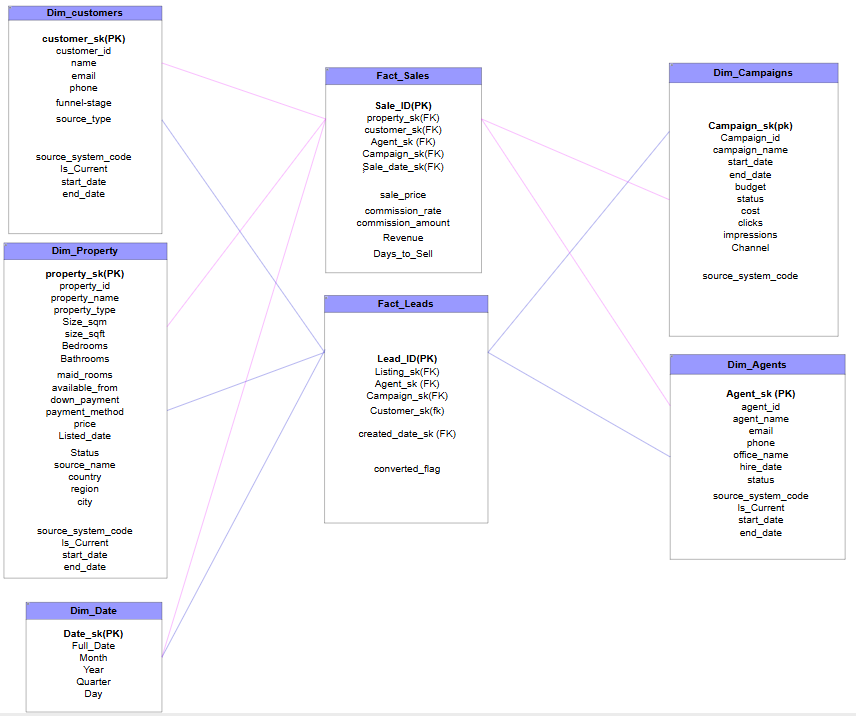
## Phase 3: Operational Database (OLTP)

1. Design an Entity Relationship Diagram (ERD) to represent entities and relationships.  
2. Define data types, primary and foreign keys.  
3. Apply constraints and indexes to improve data integrity and performance.



## Phase 4: Data Warehouse (DWH) Modeling

1. Choose a data modeling approach (Galaxy Schema).  
2. Define Fact and Dimension tables with surrogate keys.  
3. Implement Slowly Changing Dimensions (SCD Type 2).  
4. Create a Time Dimension table.



## Create Dimensions and Facts:

-- CUSTOMER DIMENSION

CREATE TABLE Dim\_Customers (

Customer\_sk INT IDENTITY(1,1) PRIMARY KEY,

Lead\_id INT,

Full\_Name NVARCHAR(100),

Email NVARCHAR(150),

Phone NVARCHAR(150),

source\_type NVARCHAR(50),

stage\_name NVARCHAR(50),

Source\_System\_Code TINYINT NOT NULL

);

-- PROPERTY DIMENSION

CREATE TABLE Dim\_Property (

Property\_sk INT IDENTITY(1,1) PRIMARY KEY,

Property\_id INT,

Property\_Name NVARCHAR(150),

Property\_Type NVARCHAR(50),

Size\_Sqft float,

Size\_Sqm float,

Bedrooms tinyint,

Bathrooms int,

Maidroom INT,

Down\_Payment float,

Payment\_Method NVARCHAR(50),

Price int,

Listed\_Date DATE,

Available\_From DATE,

Status NVARCHAR(50),

Source\_Name NVARCHAR(100),

Country NVARCHAR(100),

Region NVARCHAR(100),

City NVARCHAR(100),

Source\_System\_Code TINYINT NOT NULL,

Start\_Date DATETIME,

End\_Date DATETIME,

Is\_Current TINYINT NOT NULL

);

-- AGENT DIMENSION

CREATE TABLE Dim\_Agents (

Agent\_sk INT IDENTITY(1,1) PRIMARY KEY,

Agent\_id INT,

Agent\_Name NVARCHAR(100),

Email NVARCHAR(150),

Phone NVARCHAR(50),

Office\_Name NVARCHAR(100),

Hire\_Date DATE,

Status NVARCHAR(50),

Source\_System\_Code TINYINT NOT NULL,

Start\_Date DATETIME,

End\_Date DATETIME,

Is\_Current TINYINT NOT NULL

);

-- CAMPAIGN DIMENSION

CREATE TABLE Dim\_Campaigns (

Campaign\_sk INT IDENTITY(1,1) PRIMARY KEY,

Campaign\_id INT,

Campaign\_Name NVARCHAR(150),

Start\_Date DATE,

End\_Date DATE,

Budget DECIMAL(15,2),

Cost DECIMAL(15,2),

Impressions INT,

Clicks INT,

Channel NVARCHAR(100),

Status NVARCHAR(50),

Source\_System\_Code TINYINT NOT NULL,

);

-- FACT LEADS

CREATE TABLE Fact\_Leads (

Lead\_ID\_pk INT IDENTITY(1,1) PRIMARY KEY,

Property\_sk INT,

Agent\_sk INT,

Campaign\_sk INT,

Customer\_sk INT,

Created\_Date\_sk INT,

Converted\_Flag BIT

constraint property\_fk foreign key(Property\_sk) references Dim\_Property(Property\_sk),

constraint Lead\_ID\_fk foreign key(Customer\_sk) references Dim\_Customers(Customer\_sk),

constraint agent\_fk foreign key(Agent\_sk) references Dim\_Agents(Agent\_sk),

constraint campaign\_fk foreign key(Campaign\_sk) references Dim\_Campaigns(Campaign\_sk),

constraint date\_fk foreign key(Created\_Date\_sk) references DimDate(DateSK)

);

-- FACT SALES

CREATE TABLE Fact\_Sales (

Sale\_ID INT IDENTITY(1,1) PRIMARY KEY,

Property\_sk INT,

Customer\_sk INT,

Agent\_sk INT,

Campaign\_sk INT,

Sale\_Date\_sk INT,

Sale\_Price DECIMAL(15,2),

Commission\_Rate DECIMAL(5,2),

Commission\_Amount DECIMAL(15,2),

Revenue DECIMAL(15,2),

Days\_to\_Sell INT

constraint sales\_property\_fk foreign key(Property\_sk) references Dim\_Property(Property\_sk),

constraint sales\_customer\_fk foreign key(customer\_sk) references Dim\_Customers(Customer\_sk),

constraint sales\_agent\_fk foreign key(Agent\_sk) references Dim\_Agents(Agent\_sk),

constraint sales\_campaign\_fk foreign key(Campaign\_sk) references Dim\_Campaigns(Campaign\_sk),

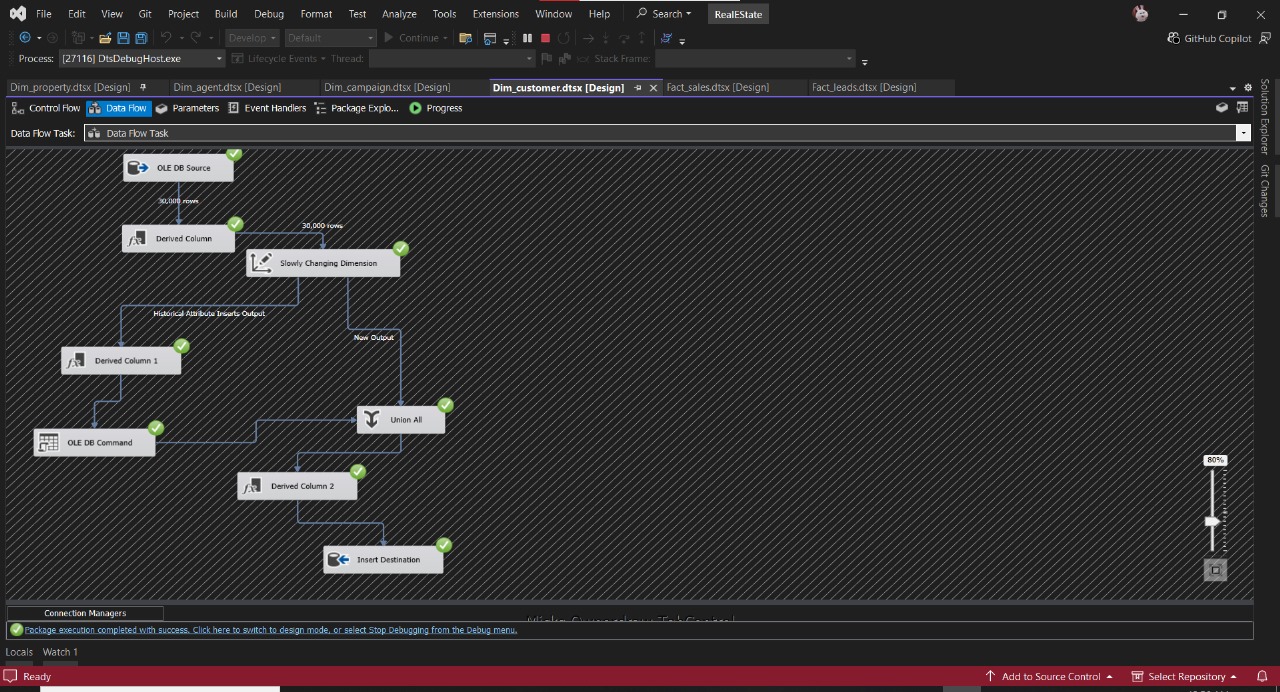
constraint sales\_date\_fk foreign key(Sale\_date\_sk) references DimDate(DateSK)

);

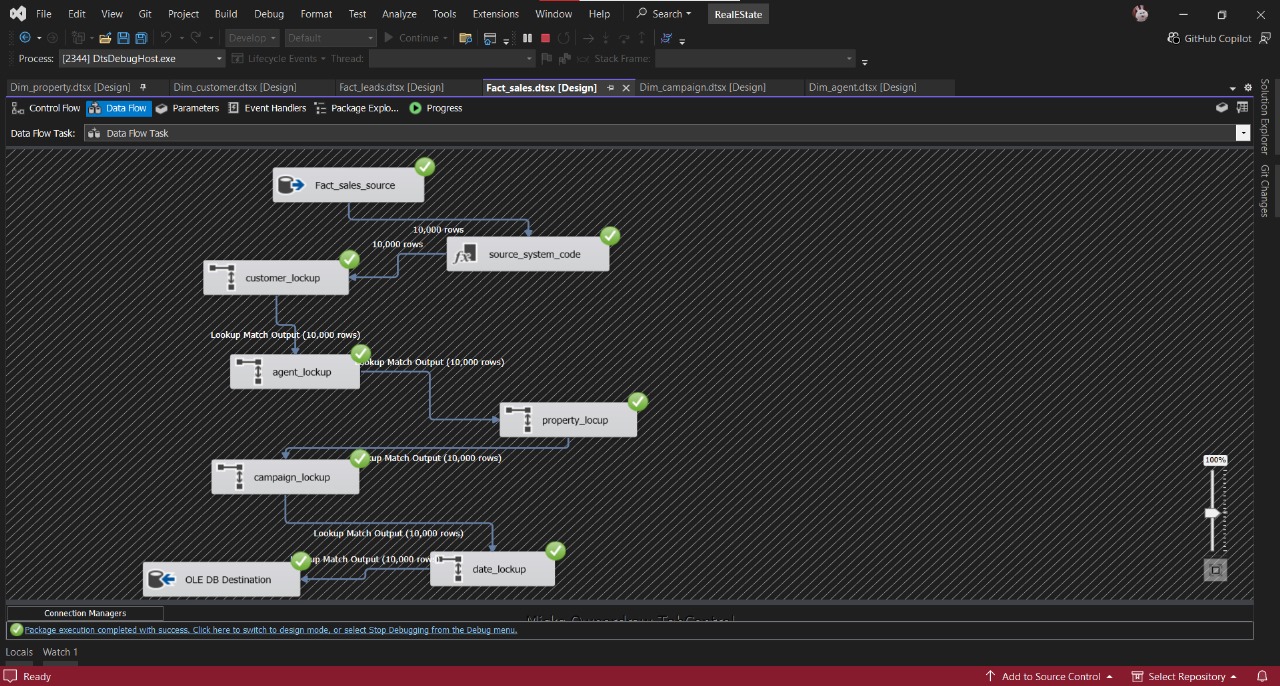
## Phase 5: ETL Process

* Create package for each table, each package independently performs the extraction, transformation, and loading processes
* The transformation logic includes data cleaning, field mapping, and enrichment to align with the Data Warehouse (DWH) structure.
* Load processed data into DWH Fact and Dimension tables.

**Customer Dimension**



**Fact Sales**



## Phase 6: Analysis Using SQL Queries

-- Property Performance

create view property\_KPIs as (

select format(avg(cast(s.sale\_price as float)) , 'N0') 'Average Property Value Sold',

avg(case when status='Sold' then datediff(day,listed\_date ,sold\_date) end) 'Average Time on Market',

count(case when status='Sold' then p.property\_id end ) \*100 / count(p.property\_id) 'Property Occupancy Rate',

count(case when status='Sold' then p.property\_id end) 'sold properties',

count(p.property\_id) 'No of properties'

from properties p left join sales\_transactions s

on p.property\_id=s.property\_id

)

GO

select \* from property\_KPIs ;

Go

-- Listings status by Region

create procedure ListingsStatusByRegion

@Status varchar(50)

as

begin

select l.region, count(p.property\_id) 'No of Listing listings'

from properties p join locations l

on l.location\_id= p.location\_id

where p.status= @Status

group by l.region

end;

GO

EXEC ListingsStatusByRegion 'Active' ;

GO

-- Agent performance

-- deals closed per agent (top 5) in a specific date

alter procedure GetTopAgentsByClosedDeals

@target\_month int , @target\_year int

as

begin

select top 5 a.full\_name ,

count(l.lead\_id) 'deals closed',

format(avg(st.sale\_price), 'N0') 'Average Revenue',

format(avg(c.commission\_amount) , 'N0') 'Average Commession'

from agents a join leads l

on l.agent\_id=a.agent\_id join sales\_transactions st

on a.agent\_id=st.agent\_id join commissions c

on st.commission\_id=c.commission\_id

where l.converted=1 and month(st.sale\_date)= @target\_month and year(st.sale\_date)=@target\_year

group by a.full\_name

order by 2 desc

end;

GO

EXEC GetTopAgentsByClosedDeals 9,2024 ;

GO

-- Market insights and pricing distribution

create view Market\_insights as(

select loc.region,pt.type\_name,

count(st.sale\_id) 'sold properties',

format(avg(st.sale\_price) , 'N0') 'Average price' ,

format(avg(p.size\_sqm) , 'N0') 'Average of Square Meter'

from properties p join sales\_transactions st

on st.property\_id= p.property\_id join property\_types pt

on p.type\_id=pt.type\_id join locations loc

on loc.location\_id=p.location\_id

where p.status='Sold'

group by loc.region,pt.type\_name

);

GO

select \* from Market\_insights

order by 3 desc

-- Total sales and revenue trends

create view SalesTrends as (

select year(sale\_date) 'Year',

month(sale\_date) 'month',

count(st.sale\_id) Sales,

format(sum(cast(st.sale\_price as float)), 'N0') Revenue,

format(sum(cast(st.sale\_price as float)-c.commission\_amount) , 'N0') Net\_Revenue

from sales\_transactions st join commissions c

on st.commission\_id=c.commission\_id

group by year(sale\_date), month(sale\_date)

);

select \* from SalesTrends

order by 1 , 2;

-- Month-over-Month Growth

create procedure MOMGrowth

@year int

as

begin

select [month],

format(Revenue, 'N0') Revenue,

format(Revenue- lag(cast(Revenue as float)) over(order by [month]) , 'N0') MOMGrowth

from(

select month(sale\_date) 'month',

sum(cast(sale\_price as float)) Revenue

from sales\_transactions

where year(sale\_date) = @year

group by month(sale\_date)

) as Revenue\_Growth

end ;

EXEC MOMGrowth 2024;

-- top-performing campaigns

create view CampaignPerformance as(

select c.campaign\_id,

c.name,

datediff(day,c.start\_date,c.end\_date ) campaign\_duration,

c.budget,

cp.clicks,

cp.cost,

cp.deals\_closed,

cp.leads\_generated,

cp.impressions,

cp.roi

from campaigns c join campaign\_performance cp

on c.campaign\_id=cp.campaign\_id

)

select \* from CampaignPerformance

-- top-performing offices

create view OfficesPerformance

as(

select o.office\_name ,

count(distinct a.agent\_id) AgentsCount,

count(l.lead\_id) LeadsCount,

count(case when l.converted=1 then lead\_id end ) Sales

from offices o join agents a

on a.office\_id=o.office\_id join leads l

on a.agent\_id=l.agent\_id

group by o.office\_name

) ;

select top 5\* from OfficesPerformance

order by Sales desc

-- Number of leads and Conversion\_Rate

create function LeadsAndConversion()

returns table

as

return(

select count(lead\_id) 'Total Leads Generated',

count(case when converted=1 then lead\_id end) \*100 / Count(lead\_id) Conversion\_Rate

from leads l )

;

select \* from dbo.LeadsAndConversion()

--Cost per Lead

create function Cost\_per\_Lead ()

returns nvarchar(50)

as

begin

declare @totalCost float, @TotalLeads int, @CostPerLead nvarchar(50)

set @totalCost =( select sum(cast(cp.cost as float))

from campaigns c join campaign\_performance cp

on cp.campaign\_id=c.campaign\_id ) ;

set @TotalLeads = (select count(lead\_id) from leads) ;

set @CostPerLead = @totalCost/@TotalLeads

Return @CostPerLead ;

end ;

select dbo.Cost\_per\_Lead() as CostPerLead

--Lead Source Effectiveness

create view LeadSourceEffectiveness

as(

select ls.source\_name,

cc.channel\_name,

count(l.lead\_id) 'No Of Leads'

from leads l join campaigns c

on l.campaign\_id=c.campaign\_id join campaign\_channels cc

on c.channel\_id=cc.channel\_id join properties p

on p.property\_id=l.property\_id join listing\_sources ls

on ls.source\_id=p.source\_id

group by ls.source\_name,cc.channel\_name

) ;

select \* from LeadSourceEffectiveness ;

-- Average Sales Cycle Duration

create function AvgSalesCycleDuration()

Returns int

as

begin

declare @AvgDuration int

set @AvgDuration= (select avg(datediff(day, l.created\_at,st.sale\_date) )

from leads l join sales\_transactions st

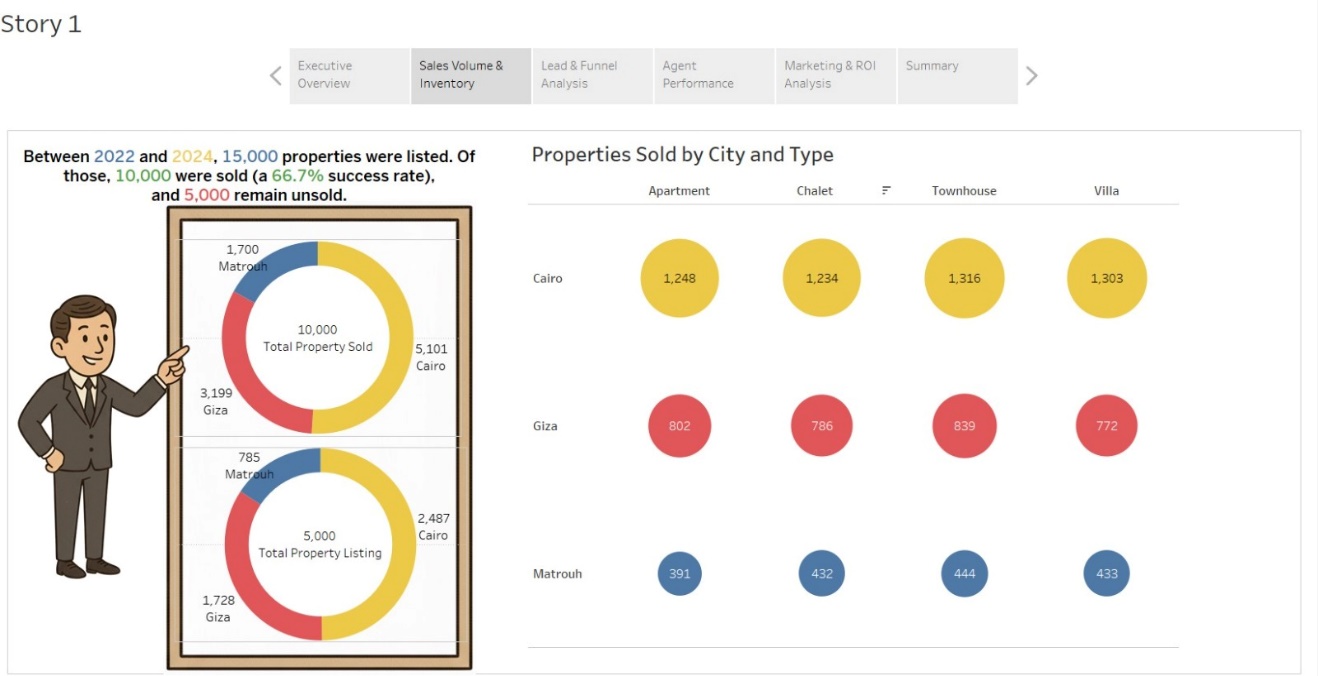
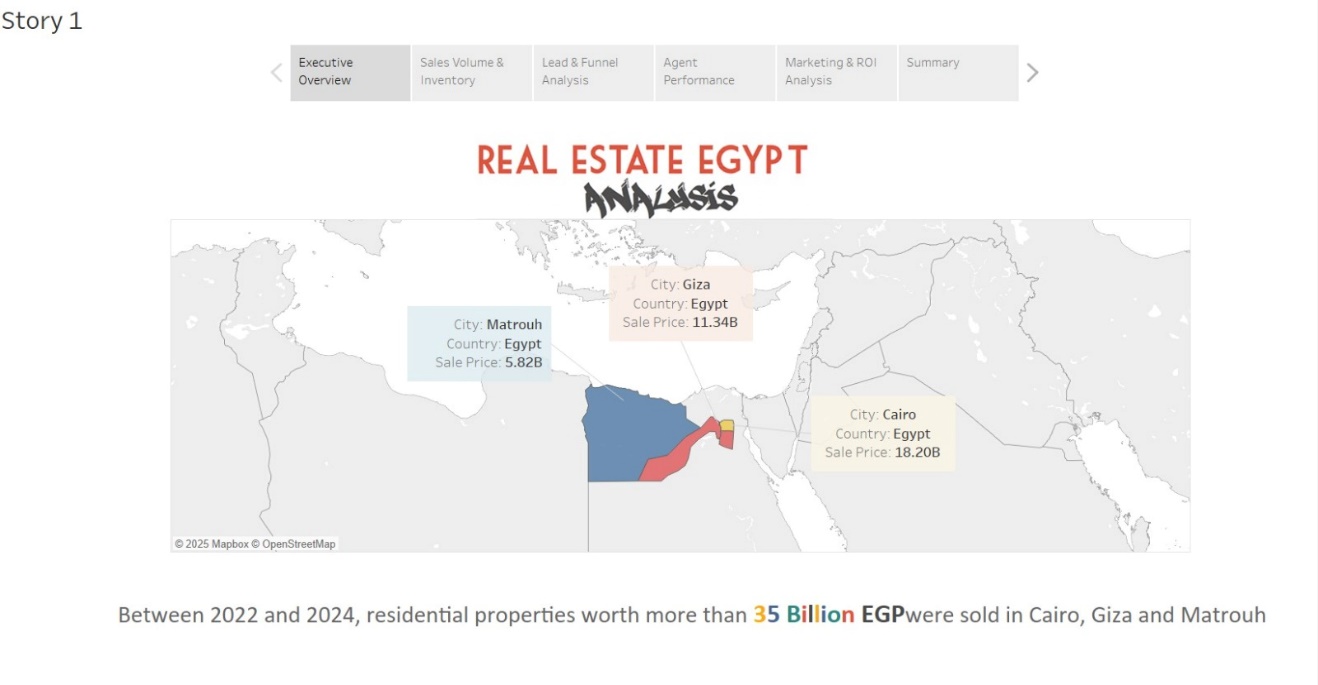
on l.lead\_id=st.lead\_id)

return @AvgDuration

end ;

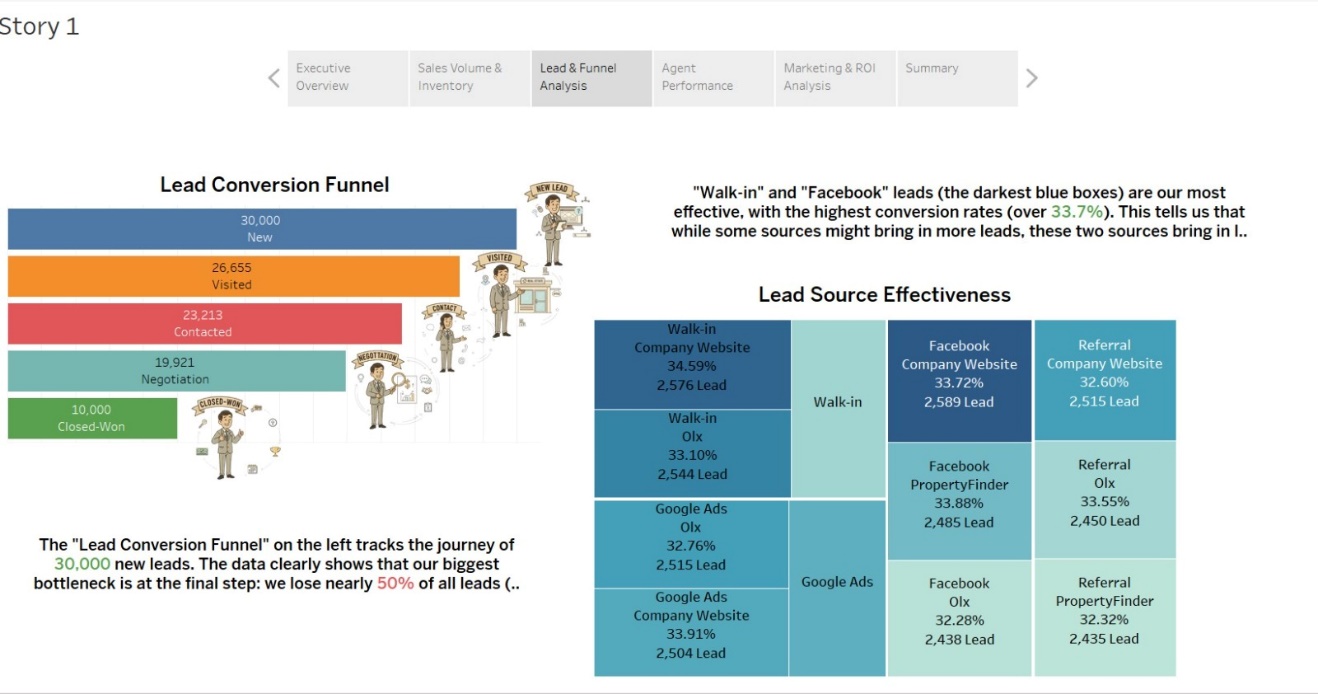
select dbo.AvgSalesCycleDuration()

## Phase 7: Tableau Analysis and Dashboards



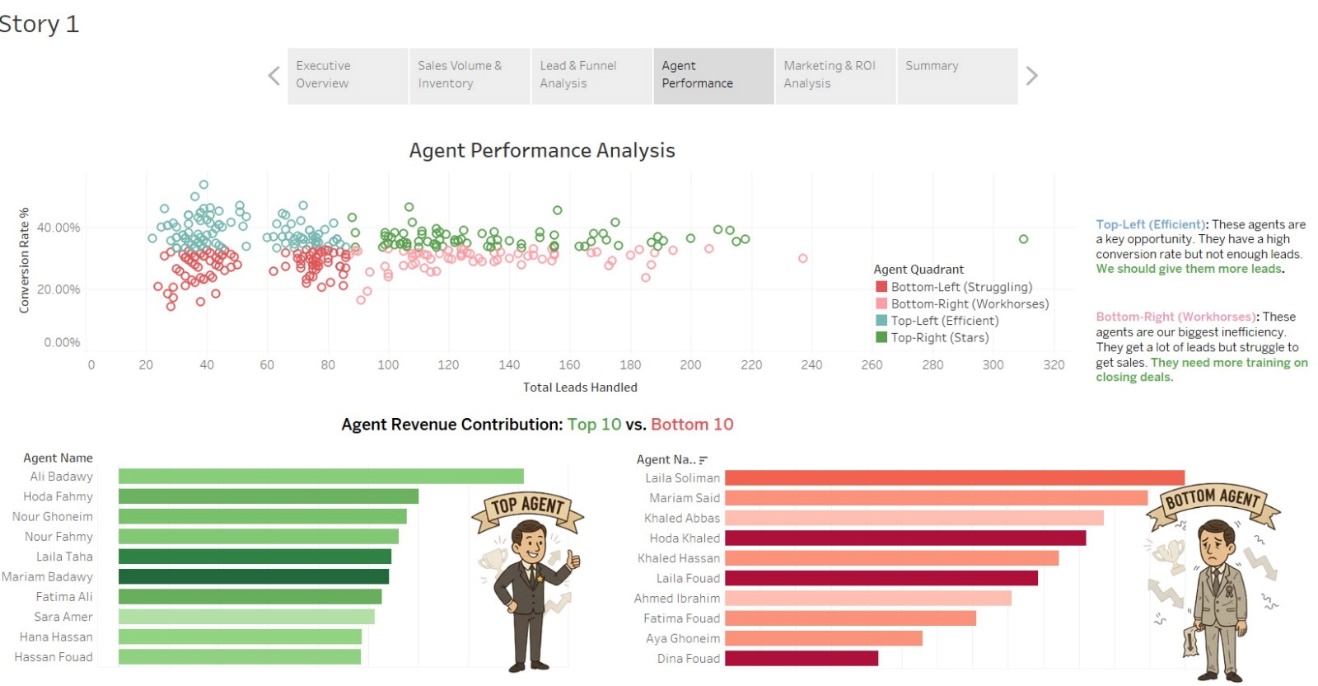
**Insights**

* **Cairo** achieved the highest property sales, representing more than **50% of total sold units**, making it the strongest real estate market.
* **Giza** came second in total sales, while **Matrouh** recorded the lowest but most balanced sales across property types.
* The overall **sales success rate is 66.7%**, with **10,000 sold properties** out of **15,000 listed** between 2022 and 2024.
* **5,000 properties remain unsold**, mostly concentrated in **Cairo and Giza**, indicating a need for better marketing or pricing strategies.
* **Townhouses and Villas** are the most popular property types across all cities



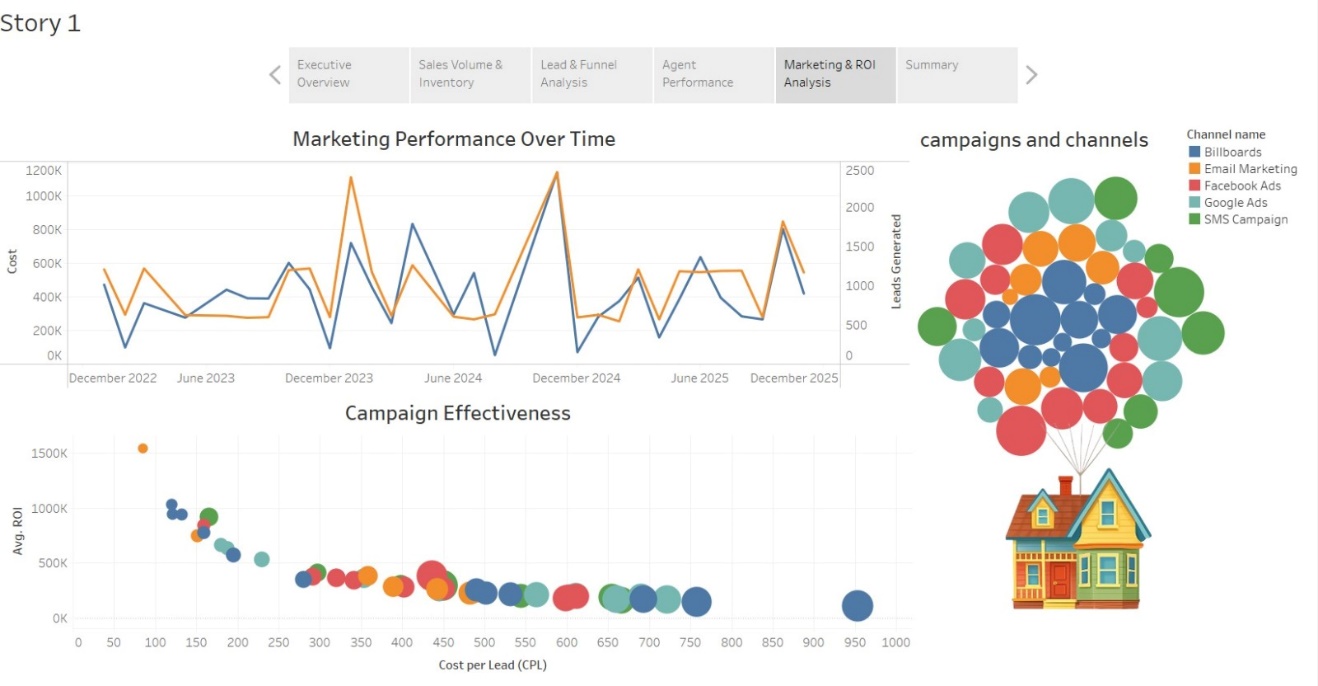
**Insights**

* **30,000 new leads** entered the funnel, but only **10,000 were converted** into closed deals — a **conversion rate of 33.3%**.
* The **biggest drop** in the funnel occurs between **Negotiation (19,921)** and **Closed-Won (10,000)** — showing the need for stronger follow-up and closing strategies.
* **“Walk-in”** and **“Facebook”** leads are the **most effective sources**, each achieving a conversion rate **above 33%**.
* **Referral** and **Google Ads** sources have **slightly lower conversion rates (around 32%)**, meaning they bring good volume but need **better nurturing** to improve closing rates.
* Overall, the data suggests the team should **focus on quality leads rather than quantity**, especially by **investing more in Walk-in and Facebook campaigns**.



**Insights**

* Top-performing agents **“Stars”** handle a large number of leads and maintain a high conversion rate
* **Top-left quadrant (Efficient agents)** have **strong conversion rates** but **fewer leads**.  
  ➤ They are an **untapped opportunity** — they should be given **more leads** to maximize performance.
* **Bottom-right quadrant (Workhorses)** handle **many leads** but have **low conversion rates**.  
  ➤ They represent **inefficiency** in the sales process and need **additional training in closing deals**.
* **Bottom-left quadrant (Struggling agents)** show both **low leads** and **low conversion rates**, indicating a need for **coaching or reassignment.**
* **Ali Badawy** is the **top-performing agent** in revenue contribution, followed by **Hoda Fahmy** and **Nour Ghoneim**.  
  ➤ These agents consistently convert leads into sales at a high rate.



**Insights**

* **Facebook Ads** and **Google Ads** are the **top-performing channels** in generating a high number of leads, while **Billboards** and **Email Marketing** show **lower efficiency** despite high costs
* **SMS Campaigns** and some **Facebook campaigns** provide **the best ROI-to-cost ratio**, meaning they are both **budget-friendly and effective** in converting leads

## Recommendations

* Focus on maintaining strong marketing and promotional efforts in **Cairo and Giza**, as they have the highest sales performance.
* Launch **targeted marketing campaigns in Matrouh** to boost property demand and improve regional sales balance.
* Improve **follow-up strategies** and provide **sales training** to reduce the drop in the **negotiation stage** of the sales funnel.
* **Increase investment** in high-performing lead sources such as **Facebook** and **Walk-in leads**, while **reassessing underperforming channels**.
* **Distribute leads more efficiently** by assigning more to top-performing agents and offering **training to low-performing ones**.
* **Allocate a larger share of the marketing budget** to **digital channels** (Facebook, Google Ads, SMS) with higher ROI.
* **Reduce spending** on low-return traditional channels like **Billboards** and **Email Marketing**.