



University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

# Distributed & Scalable Data Engineering – Technical Report



Fall 2024

# CONTENTS

Project Name .....2

Executive Summary .....2

Technical Report.....3

Highlights of Project .....3

Submitted on:.....3

Abstract .....4

Methodology .....6

Results SectionError! Bookmark not defined.

Discussion.....Error! Bookmark not defined.

Conclusion .....Error! Bookmark not defined.

# Project Name

## Executive Summary

The “**Comprehensive Analysis of Real Estate Market**” project examines the factors influencing residential property prices, focusing on the impact of COVID-19. Using the CAMA dataset and CRISP-DM methodology, the team built linear regression models to predict prices, achieving an R-squared of 0.6539. Key drivers like bedrooms, bathrooms, and gross building area were found to significantly impact prices. COVID-19 led to higher demand for larger homes, rising prices, and increased sales volume. The project utilized AWS S3, Athena, Glue, and SageMaker for data processing and modeling, providing actionable insights for real estate stakeholders.



### Team Members:

**Sai Siddu Vardhan Reddy Annadi**  
**Vinay Kumar Reddy Punuru**  
**Shiva Priya Pillalamarri**

### Questions?

Contact : [sanna10@unh.newhaven.edu](mailto:sanna10@unh.newhaven.edu)  
[vpunu2@unh.newhaven.edu](mailto:vpunu2@unh.newhaven.edu)  
[spill6@unh.newhaven.edu](mailto:spill6@unh.newhaven.edu)

# Technical Report

## *Comprehensive Analysis of Real Estate Market*

### Highlights of Project

The “Comprehensive Analysis of Real Estate Market” project explores key factors affecting residential property prices, especially during and after COVID-19. Using the CAMA dataset and CRISP-DM methodology, the team built linear regression models with predictors like bedrooms, bathrooms, property condition, and gross building area, achieving an R-squared of 0.6539. Analysis revealed that COVID-19 significantly influenced housing prices, driving demand for larger homes and suburban living. Data processing and modeling were conducted using AWS S3, Athena, Glue, and SageMaker, offering valuable insights for real estate stakeholders to adapt to evolving market trends.



**Submitted on : 12-08-2024**

## Abstract

The “**Comprehensive Analysis of Real Estate Market**” project aims to identify key factors influencing residential property prices, with a special focus on the impact of **COVID-19**. Using the **CAMA dataset from Open Data DC** and following the **CRISP-DM methodology**, the project analyzes property characteristics, neighborhood factors, and lifestyle changes to understand price fluctuations. The team developed **linear regression models** with predictors like bedrooms, bathrooms, property condition, and gross building area, achieving an **R-squared of 0.6539**, indicating strong predictive performance. The analysis highlights the influence of COVID-19, which shifted buyer preferences toward larger homes, suburban living, and proximity to recreational spaces. Data was processed and modeled using **AWS tools** like **S3, Athena, Glue, and SageMaker**, ensuring efficient data handling and deployment. The findings offer valuable insights for real estate stakeholders, developers, and policymakers to adapt to changing market dynamics and make data-driven decisions.

Pitch: <https://github.com/siddureddy-DS/Team08-DSCI-6007-01>

## Executive Summary

The “**Comprehensive Analysis of Real Estate Market**” project examines the factors influencing residential property prices, focusing on the impact of COVID-19. Using the CAMA dataset and CRISP-DM methodology, the team built linear regression models to predict prices, achieving an R-squared of 0.6539. Key drivers like bedrooms, bathrooms, and gross building area were found to significantly impact prices. COVID-19 led to higher demand for larger homes, rising prices, and increased sales volume. The project utilized AWS S3, Athena, Glue, and SageMaker for data processing and modeling, providing actionable insights for real estate stakeholders.

## Introductory Section

The “**Comprehensive Analysis of Real Estate Market**” project addresses the evolving nature of residential property prices, particularly in the context of the **COVID-19 pandemic**. Traditionally, factors like property size, location, and neighborhood characteristics have driven real estate prices. However, the pandemic introduced new dynamics, such as increased demand for larger living spaces, remote work flexibility, and a growing preference for suburban living. These changes have reshaped buyer behavior, creating both challenges and opportunities for real estate developers, investors, and policymakers.

This project aims to provide a **data-driven approach** to understanding these shifts by analyzing the **CAMA dataset from Open Data DC**. The study leverages the **CRISP-DM methodology** to explore key predictors of property prices, including the number of bedrooms, bathrooms, property condition, and gross building area. To ensure robust analysis and model development, the project employs a suite of **AWS cloud tools**, including **S3 for data storage, Athena for querying, Glue for ETL, and SageMaker for machine learning**.

The analysis is segmented into three phases—**pre-COVID, during COVID, and post-COVID**—to capture the unique effects of the pandemic on property demand and prices. By building and testing **linear regression models**, the project aims to provide insights into the key factors that influence residential property prices. The results will enable real estate stakeholders to make informed decisions, adapt to new market demands, and seize growth opportunities in the evolving housing market.



## Methodology

CRISP-DM methodology :

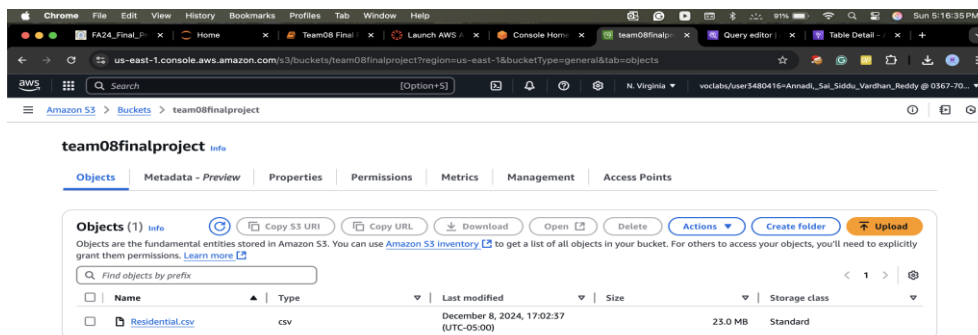
**Title of the Project : Comprehensive Analysis of Real Estate Market**

## Business Understanding:

The primary goal of this project is to provide actionable insights for real estate stakeholders, developers, and policymakers to better understand the changing landscape of residential property prices. By using the CAMA dataset from Open Data DC, the project seeks to address key business questions, including:

1. What factors most significantly influence property prices?
2. How have property sales and price trends changed before, during, and after COVID-19?
3. Which property features, like bedrooms, bathrooms, property condition, and gross building area, impact the selling price the most?
4. How can developers and investors adapt their strategies to meet shifting buyer demands?

## Data Understanding:



Amazon Athena > Query editor

Editor | **Recent queries** | Saved queries | Settings

Workgroup: primary

Recent queries (29)

Execution ID	Query	Start time	Status	Run time
1068ffa4-a36-4b75-840c-18c39b987554	SELECT * FROM "default"."residential_data" limit 10	2024-12-08T17:10:51.002-05...	Completed	602 ms
5dc6087c-e8f9-4174-addc-8010a37dbfa7	CREATE EXTERNAL TABLE IF NOT EXISTS residential_data (ssl ...	2024-12-08T17:10:42.998-05...	Completed	369 ms
75b76a0b-05b8-4c59-befc-0fa5a44ca2	DROP TABLE "residential_data"	2024-12-08T17:09:33.593-05...	Completed	596 ms
76cd7e9b-76ff-4069-b189-fb51f3c870d9	SELECT * FROM "default"."residential_data" limit 10	2024-12-08T17:09:19.938-05...	Completed	588 ms
633e7cf6-beb1-44e3-8b6d-fbb536e1cdf4	CREATE EXTERNAL TABLE IF NOT EXISTS residential_data (ssl ...	2024-12-08T17:09:08.769-05...	Completed	408 ms
03f8aefb-6172-4bcc-b23f-ae11710556c0	SELECT * FROM "default"."residential_data" limit 10	2024-12-08T17:08:52.920-05...	Failed	365 ms
2885d821-df26-4f77-9793-a4bed628216e	DROP TABLE "residential_data"	2024-12-08T17:08:34.151-05...	Failed	1.446 sec
125a0788-ed3c-4ff1-934e-272adc4a1024	SELECT * FROM "default"."residential_data" limit 10	2024-12-08T16:59:39.563-05...	Completed	591 ms
28bceedb-e753-49b9-8d7d-9dc92b7a6a1b	CREATE EXTERNAL TABLE IF NOT EXISTS residential_data (ssl ...	2024-12-08T16:57:24.067-05...	Completed	344 ms
cdd30984-b8a8-437b-b25d-53b602659653	SELECT * FROM residential_data09 LIMIT 10	2024-12-05T15:15:41.046-05...	Completed	572 ms
ca7f2d7d-42fd-4551-aba7-736560ab4000	CREATE EXTERNAL TABLE IF NOT EXISTS residential_data09 ( ...	2024-12-05T15:15:17.291-05...	Completed	406 ms
fe82fa55-5361-4fc0-bd6f-29ef0f20ef91	CREATE EXTERNAL TABLE IF NOT EXISTS residential_data (ssl ...	2024-12-05T15:14:46.543-05...	Completed	356 ms

## Data Preparation: Loading Data to Schema - AWS Glue: Loading Data to Schema

AWS Glue > Tables

Announcing new optimization features for Apache Iceberg tables  
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

**Tables**  
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (1)  
View and manage all available tables.

Filter tables

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
residential_data	default	s3://team08finalprojec	-	-	<a href="#">Table data</a>	<a href="#">View data quality</a>	<a href="#">View statistics</a>

1



**residential\_data**

Table overview | Data quality - new

**Table details**

<b>Name</b> residential_data	<b>Classification</b> -	<b>Deprecated</b> -
<b>Database</b> default	<b>Location</b> s3://team08finalproject/	<b>Column statistics</b> No statistics
<b>Description</b> -	<b>Connection</b> -	
<b>Last updated</b> December 6, 2024 at 22:10:43		

► **Advanced properties**

**Schema** | Partitions | Indexes | Column statistics - new

**Schema (39)**  
View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	ssl	string	-	-
2	bathrm	float	-	-
3	hf_bathrm	float	-	-
4	heat	float	-	-
5	heat_d	string	-	-
6	ac	string	-	-
7	num_units	float	-	-
8	rooms	float	-	-
9	bedrm	float	-	-
10	ayb	float	-	-
11	yr_rmdl	float	-	-
12	eyb	int	-	-
13	stories	float	-	-

## Transforming the Cleaning Data

Chrome

File

Edit

View

History

Bookmarks

Profiles

Tab

Window

Help

<

## Modeling: AWS Sage Maker and Sagemaker Notebook Instances

The image shows two screenshots related to AWS SageMaker. The top screenshot is the AWS SageMaker console, and the bottom screenshot is a Jupyter Notebook instance.

**Top Screenshot: AWS SageMaker Console**

The console shows the "Notebook instances" page. A success message at the top states: "Success! Your notebook instance is being created. Open the notebook instance when status is InService and open a template notebook to get started." Below this, a table lists the notebook instances:

Name	Instance	Creation time	Status	Actions
team08finalprojectdeploy	ml.t3.xlarge	12/8/2024, 5:25:45 PM	InService	Open Jupyter   Open JupyterLab
Team08deployment	ml.t3.medium	12/5/2024, 4:54:20 PM	InService	Open Jupyter   Open JupyterLab

**Bottom Screenshot: Jupyter Notebook Instance**

The bottom screenshot shows a Jupyter Notebook instance titled "Team08 Final Project Code". The notebook content includes a title, team members, project description, dataset source, and a code cell with Python imports.

**Final Project**  
**Distributed & Scalable Data Engr - DSCI-6007-01**  
**Comprehensive Analysis of Real Estate Market**  
**Team Members:** Sai Siddu Vardhan Reddy Annadi, Vinay Kumar Reddy Punuru, Shiva Priya Pillalamarri

Our project leverages data from Open Data DC, specifically focusing on the sale history of active properties listed in the District of Columbia's real property tax assessment roll. The dataset includes approximately 108,996 records with 39 columns, capturing various property attributes such as area, number of bedrooms, and other key features. It also provides sale information, including sale prices and transaction dates.

The primary objective of our analysis is to explore the relationship between these property attributes and their influence on sale prices. By applying statistical analysis and predictive modeling, we aim to identify the key drivers of property valuation.

Dataset Source: The dataset is sourced from Open Data DC's Computer Assisted Mass Appraisal - Residential dataset : <https://opendata.dc.gov/datasets/DCGIS:computer-assisted-mass-appraisal-residential/explorer>

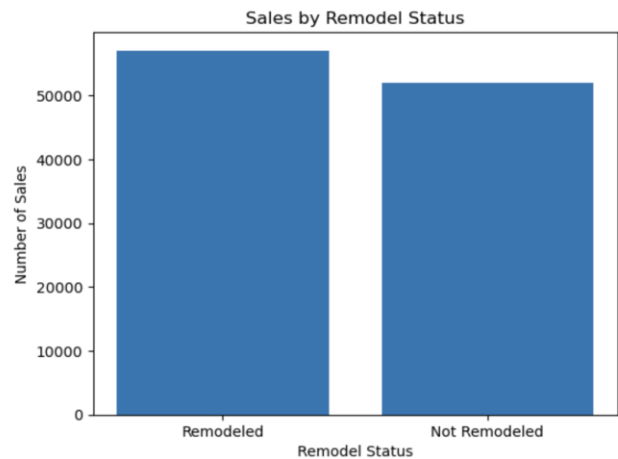
Key Considerations for Analysis: This analysis will cover property sales from 2010 to 2024. Properties with a sale price of \$0 will be excluded from the analysis to ensure accuracy and relevance.

This approach will allow for a comprehensive understanding of how different property characteristics impact sale prices over time.

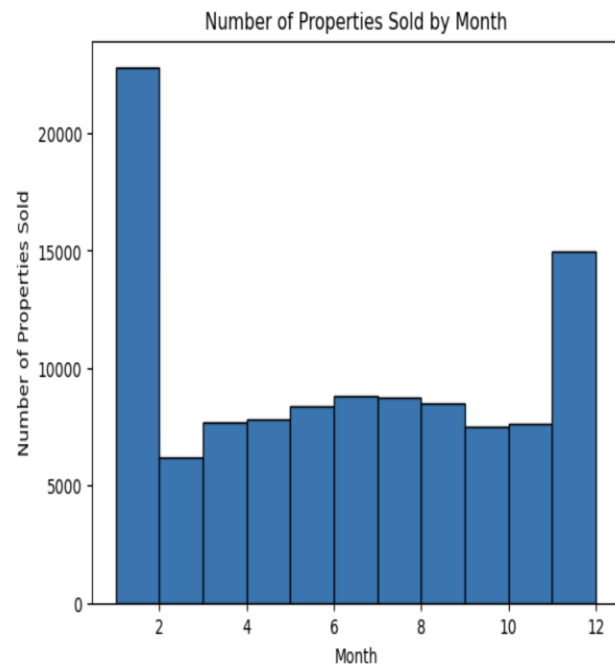
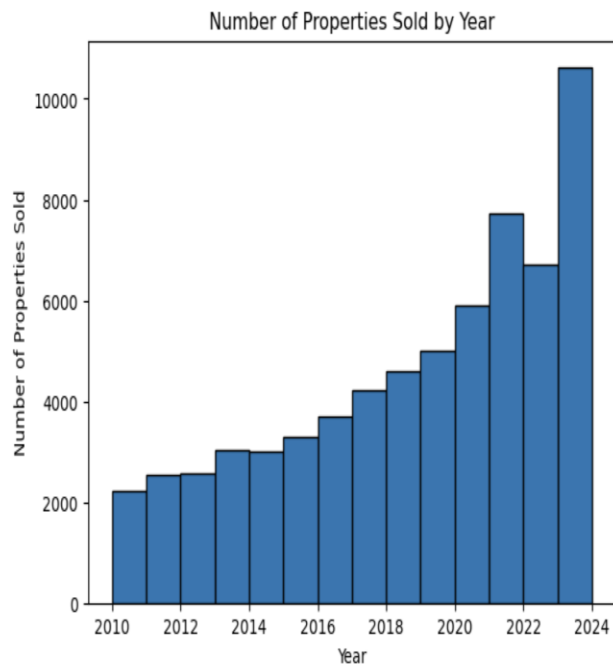
```
[11]: # Import Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import accuracy_score, precision_score, recall_score
from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import plot_confusion_matrix
```

## Evaluation - Data Analysis:

### Property Sales by Price and Remodeled Status



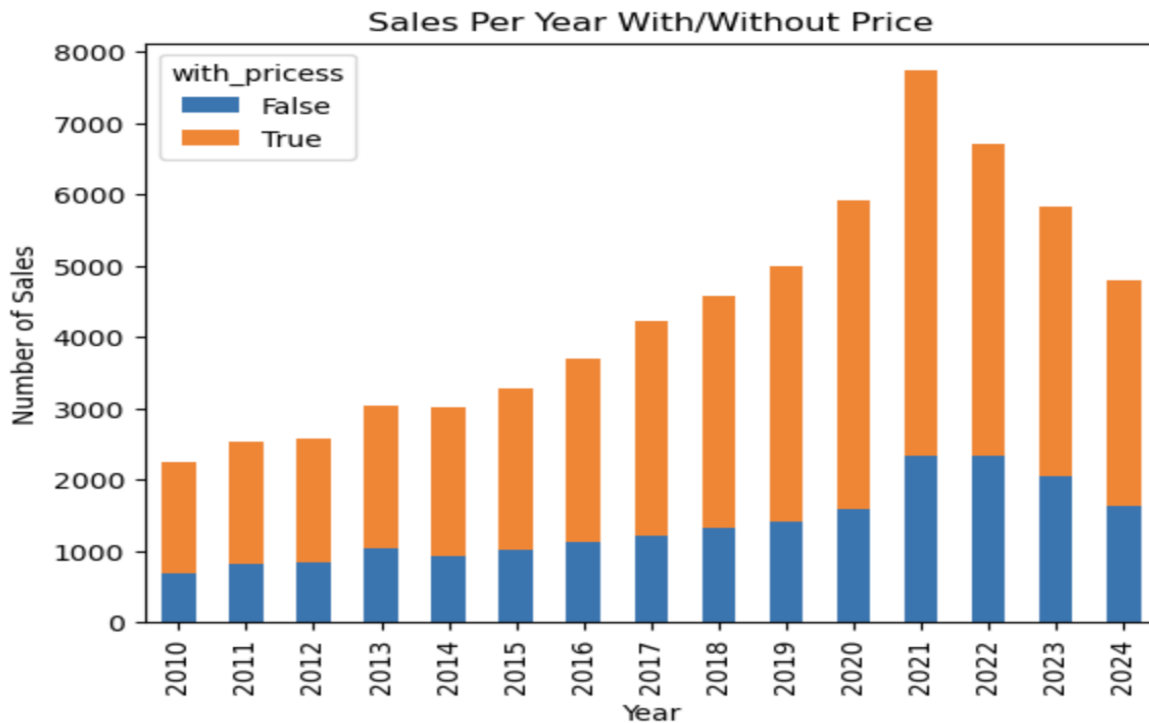
### Annual and Monthly property sales trends



## Distribution of Sale Prices for Trimmed Sales Data with KDE Plot



## Annual Property Sales Comparison with and without Price



## Results:

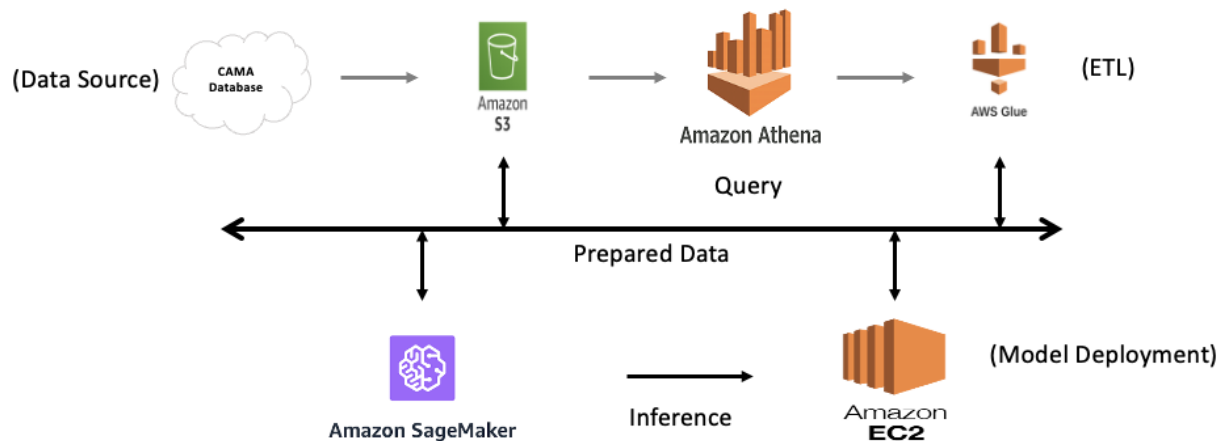
### Model: Build LR model: Adding gross building area another predictor

- X : bathrm, bedrm, grade, heat, cndtn, gba
- Y: price

=== Linear Regression Summary ===

- Independent variables: ['bathrm', 'bedrm', 'grade', 'heat', 'cndtn', 'gba']
- Dependent variable: price
- Training data size: 24499
- Test data size: 6125
- Mean squared error: 212482373114.2765
- R-squared: 0.6539350923078139
- Coefficients: [ 45474.64703669 -72781.65407855 221173.8025435  
857.96809526 257221.10348558 434.47093989]

## Data Engineering Pipeline:



### 1. Data Ingestion:

- Data is sourced from databases or data lakes.
- This data is then stored in **Amazon S3**

### 2. Data Storage:

- **Amazon Athena** is used to query and analyze the data directly stored in Amazon S3. Athena enables running SQL queries on this data without managing a data warehouse.

### 3. Data Processing:

- The processed data is prepared for machine learning using **Amazon SageMaker**, which is used for building, training, and deploying machine learning models.

### 4. Model Deployment:

- The models are deployed using **Amazon EC2** instances to perform inferences on the prepared data.

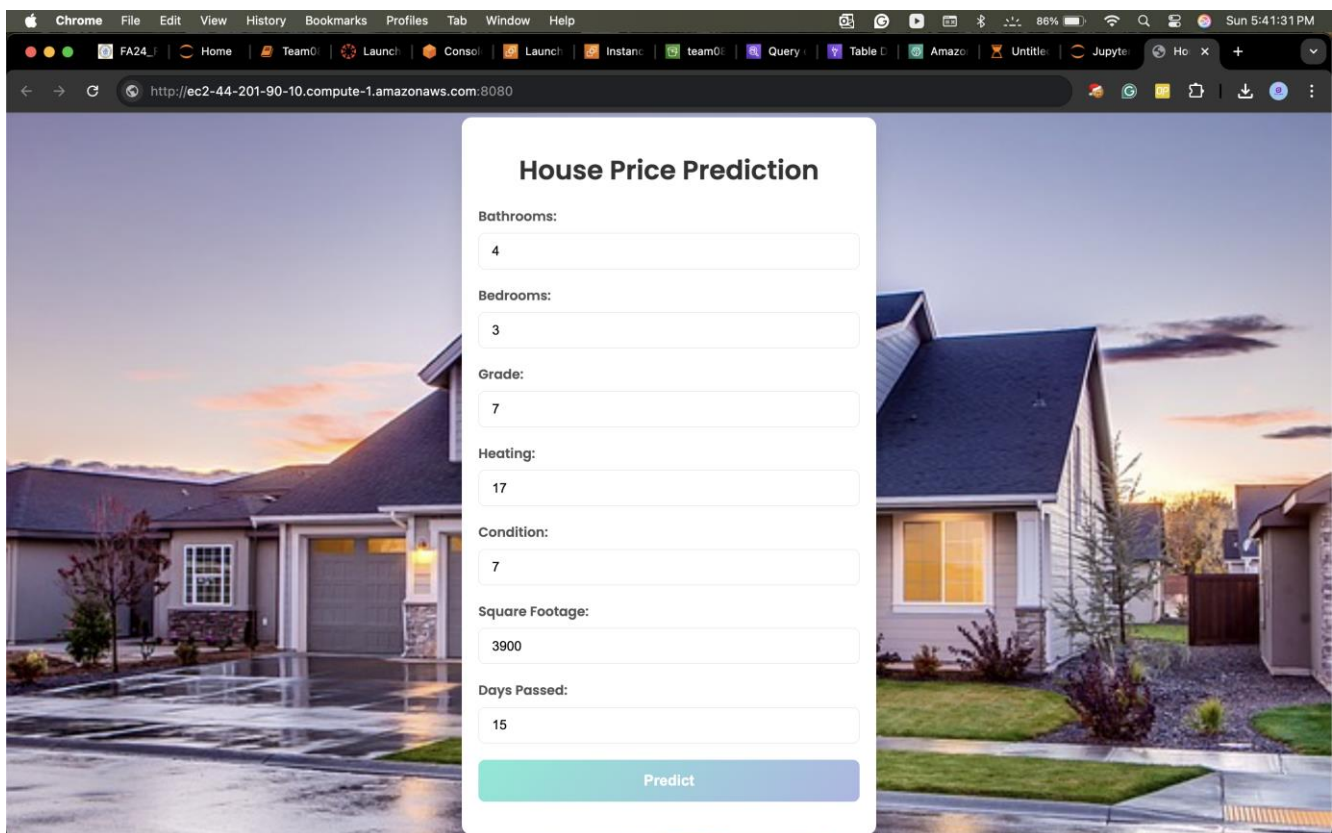


- Additionally, **AWS Glue** is used to handle ETL (Extract, Transform, and Load) processes, transforming data between various sources and destinations.

## 5. Data Visualization:

- Results from the analysis, such as property sales trends and prices, are visualized for interpretation and reporting.

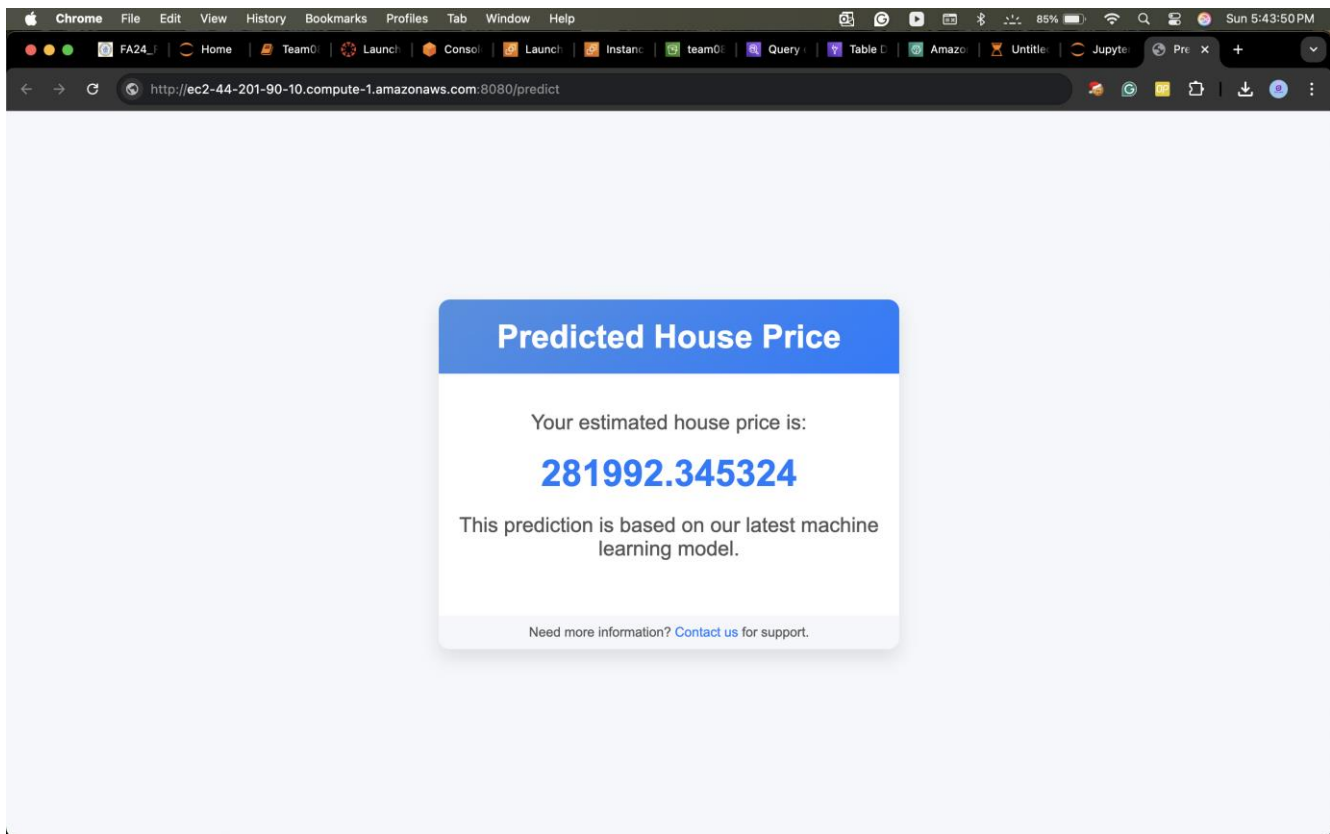
## Deployment



The screenshot shows a web browser window with a URL bar indicating an Amazon EC2 instance. Overlaid on the browser is a 'House Price Prediction' form. The form has a title 'House Price Prediction' and several input fields with labels and values:

Field Label	Value
Bathrooms:	4
Bedrooms:	3
Grade:	7
Heating:	17
Condition:	7
Square Footage:	3900
Days Passed:	15

At the bottom of the form is a green button labeled 'Predict'.



## Discussion

### COVID-19 Impact

Time Period Segmentation:

- Pre-COVID (ExAnte) Period: January 1, 2019 – February 29, 2020
- During COVID Period: March 1, 2020 – July 31, 2021
- Post-COVID (ExPost) Period: August 1, 2021 – December 31, 2022

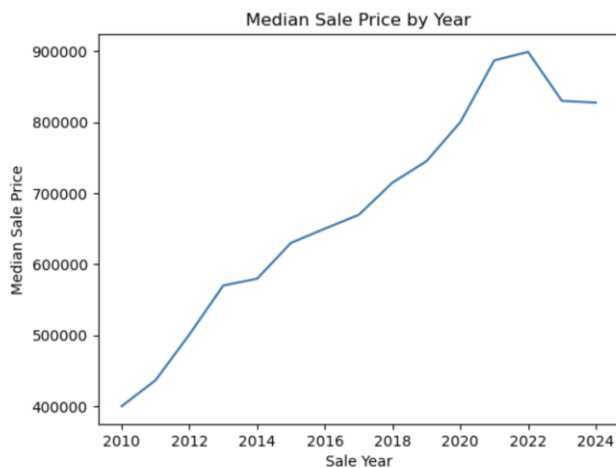
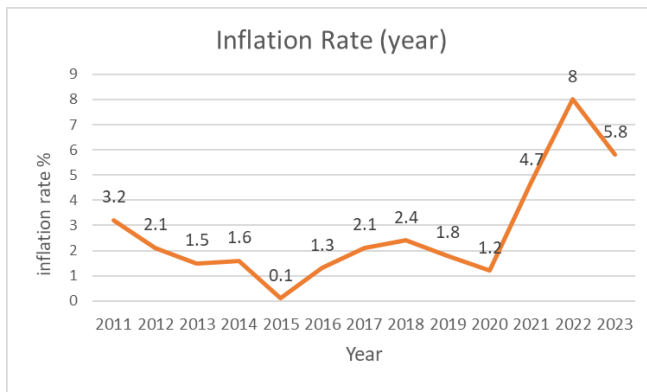
### Price Fluctuations:

Significant shifts in residential property prices were observed before, during, and after the COVID-19 pandemic. Noticeable differences in price

trends highlight the impact of the pandemic on housing demand and market conditions.

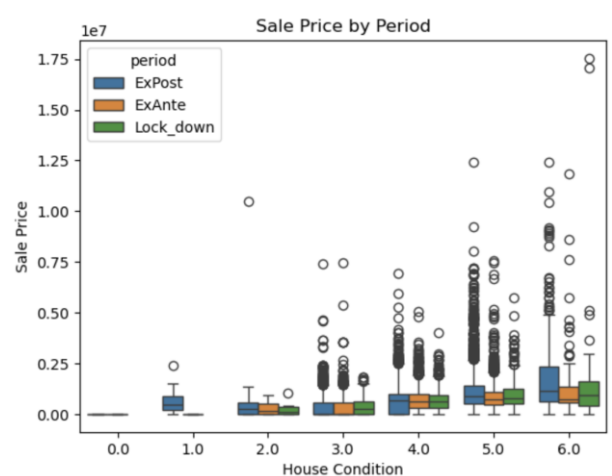
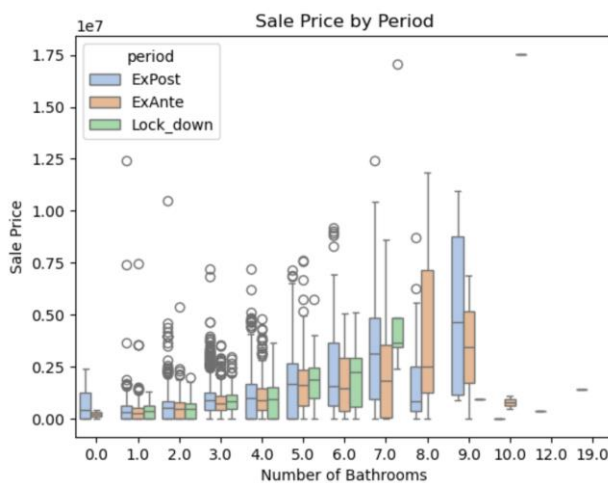
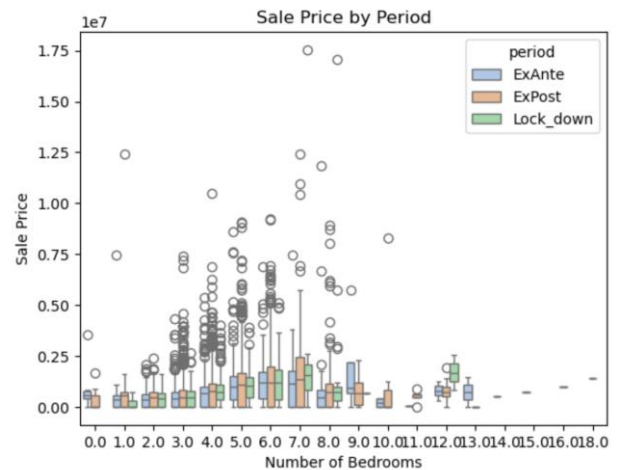
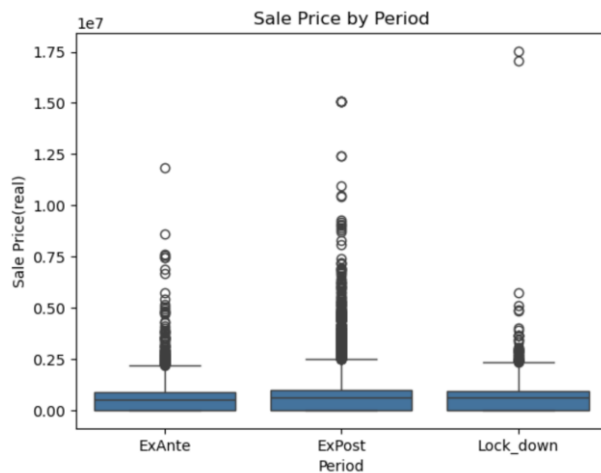
### Inflation Adjustment:

To ensure a more accurate comparison of house prices across the three periods, property prices were deflated using the inflation rate, minimizing the influence of general price fluctuations.



## Price Comparison by Property Features:

Analyzing house price levels by the number of bedrooms, number of bathrooms, and overall property condition reveals distinct differences in pricing trends across the three periods. Properties with certain features, such as larger living spaces and better conditions, experienced greater price fluctuations, reflecting shifts in buyer preferences driven by the COVID-19 pandemic.



## Generalized Linear Model Regression Results

Dep. Variable:	Real_price	No. Observations:	24579
Model:	GLM	Df Residuals:	24572
Model Family:	Gaussian	Df Model:	6
Link Function:	Identity	Scale:	3.8611e+11
Method:	IRLS	Log-Likelihood:	-3.6275e+05
Date:	Sun, 08 Dec 2024	Deviance:	9.4876e+15
Time:	13:17:26	Pearson chi2:	9.49e+15
No. Iterations:	3	Pseudo R-squ. (CS):	0.3735
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
Intercept	-1.069e+06	2.34e+04	-45.693	0.000	-1.11e+06	-1.02e+06
C(period) [T.ExPost]	2.867e+04	9563.311	2.998	0.003	9930.107	4.74e+04
C(period) [T.Lock_down]	2.845e+04	1.58e+04	1.800	0.072	-2524.816	5.94e+04
bathrm	8.556e+04	5711.466	14.980	0.000	7.44e+04	9.67e+04
bedrm	-4.15e+04	4697.744	-8.833	0.000	-5.07e+04	-3.23e+04
cndtn	2.709e+05	5401.697	50.154	0.000	2.6e+05	2.82e+05
gba	330.5307	6.871	48.108	0.000	317.065	343.997

## Generalized Linear Model Regression Results

Dep. Variable:	sale_num	No. Observations:	24579
Model:	GLM	Df Residuals:	24572
Model Family:	Gaussian	Df Model:	6
Link Function:	Identity	Scale:	2.2144
Method:	IRLS	Log-Likelihood:	-44642.
Date:	Sun, 08 Dec 2024	Deviance:	54412.
Time:	13:17:26	Pearson chi2:	5.44e+04
No. Iterations:	3	Pseudo R-squ. (CS):	0.07249
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
Intercept	2.5115	0.056	44.836	0.000	2.402	2.621
C(period) [T.ExPost]	0.1999	0.023	8.727	0.000	0.155	0.245
C(period) [T.Lock_down]	0.1019	0.038	2.693	0.007	0.028	0.176
bathrm	0.2238	0.014	16.361	0.000	0.197	0.251
bedrm	-0.0063	0.011	-0.559	0.576	-0.028	0.016
cndtn	0.3469	0.013	26.818	0.000	0.322	0.372
gba	-0.0004	1.65e-05	-24.648	0.000	-0.000	-0.000

The COVID-19 period had a positive impact on housing prices and sales volume.

- Increase in Housing Prices: Compared to the Pre-COVID (ExAnte) period, housing prices rose significantly during and after the COVID period, reflecting increased demand for residential properties.
- Increase in Sales Volume: Similarly, sales volume saw a noticeable rise during and after COVID compared to the ExAnte period, indicating higher buyer activity and market engagement.

The “Comprehensive Analysis of Real Estate Market” project provides a data-driven understanding of the factors influencing residential property prices, with a focus on the impact of COVID-19. By leveraging the CAMA dataset and utilizing the CRISP-DM methodology, the project explored how property features, market trends, and pandemic-induced lifestyle changes shaped the real estate market. Key predictors such as bedrooms, bathrooms, property condition, and gross building area were identified as significant factors impacting property prices. The development of linear regression models with an R-squared of 0.6539 demonstrated the effectiveness of incorporating additional predictors for more accurate price forecasting.

The analysis revealed a significant rise in housing prices and sales volume during and after COVID, driven by shifts in buyer preferences toward larger homes, suburban living, and better living conditions. The segmentation of the data into Pre-COVID, During-COVID, and Post-COVID periods highlighted clear differences in pricing and demand trends. The use of AWS tools (S3, Athena, Glue, SageMaker) facilitated efficient data handling, analysis, and model deployment.

This project offers valuable insights for real estate developers, investors, and policymakers, enabling them to make informed, data-driven decisions. The findings emphasize the importance of adapting to evolving market



demands, especially as buyer preferences continue to shift in the post-pandemic era. These insights can be used to optimize development strategies, identify emerging investment opportunities, and create more resilient market responses to future disruptions.

## Contributions/References

1. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media. ISBN: 978-1-492-03264-9.
2. Duca, J. V., & Murphy, A. (2021). Why house prices surged as the COVID-19 pandemic took hold. *Federal Reserve Bank of Dallas*.  
<https://www.dallasfed.org/research/economics/2021/1228>
3. Schwartz, A. E., & Wachter, S. (2022). COVID-19's impacts on housing markets: Introduction. *Journal of Housing Economics*.