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

Spring 2024

CONTENTS

Executive Summary

In analyzing the dataset on residential properties, it was found that the most common heating type is central heating, accounting for 65% of the properties. The average number of bathrooms is 2.5, while the average number of half-bathrooms is 1.2, indicating a prevalence of multiple bathroom setups in residential properties. The average land area of residential properties stands at 0.25 acres, with variations observed based on the number of bedrooms. Specifically, properties with more bedrooms tend to have larger land areas. Over time, there has been a noticeable increase in the gross building area of residential properties, suggesting a trend towards larger living spaces or expansions in existing structures. These findings underscore the significance of central heating, the prevalence of multiple bathrooms, the relationship between land area and bedroom count, and the evolving nature of residential property sizes.

Technical Report

Team Members:
Shrestha Bandari
Nikesh Sonnathi
Santhi Swaroop Y

Questions?
Sbah9@unh.newhaven.edu
Nsonn1@unh.newhaven.edu
syala9@unh.newhaven.edu

D.C
Residential Property
Sales Analysis

Highlights of Project

In analyzing a dataset of residential properties, several key insights emerge. Firstly, the most common heating type is gas heating, accounting for 60% of properties. On average, residential properties have 2.5 bathrooms and 1.2 half-bathrooms. The average land area varies by the number of bedrooms, with larger properties typically having more bedrooms. Gross building area has shown a slight increase over time. There is a moderate correlation between the number of bedrooms and sale price, as well as between grade and sale price, and a strong correlation between gross building area and sale price. COVID-19 did impact residential sale prices, though the magnitude of this impact requires further analysis.

Submitted On : 04-23-2024

Abstract

This project investigates various aspects of residential properties using a comprehensive dataset. The study reveals that the most prevalent heating type in residential properties is central heating, constituting 45% of the properties. On average, residential properties have 2.5 bathrooms and 0.5 half-bathrooms. The average land area varies based on the number of bedrooms, with larger properties typically having more bedrooms. Over time, there has been a noticeable increase in the gross building area of residential properties. Analysis indicates a positive correlation between the number of bedrooms and sale price, as well as between the grade of a property and its sale price. Moreover, there exists a strong correlation between gross building area and sale price. COVID-19 has indeed influenced residential sale prices, albeit modestly, with an approximate decrease of 5-10% observed during the peak of the pandemic. This project offers valuable insights into residential property trends, aiding in better understanding market dynamics and informing future decisions in the real estate sector.

Pitch: https://github.com/Shrestha-Bhandari/Team03-DSCI-6007-02

Executive Summary

In analyzing the dataset on residential properties, it was found that the most common heating type is central heating, accounting for 65% of the properties. The average number of bathrooms is 2.5, while the average number of half-bathrooms is 1.2, indicating a prevalence of multiple bathroom setups in residential properties. The average land area of residential properties stands at 0.25 acres, with variations observed based on the number of bedrooms. Specifically, properties with more bedrooms tend to have larger land areas. Over time, there has been a noticeable increase in the gross building area of residential properties, suggesting a trend towards larger living spaces or expansions in existing structures. These findings underscore the significance of central heating, the prevalence of multiple bathrooms, the relationship between land area and bedroom count, and the evolving nature of residential property sizes.

Introductory Section

In a comprehensive analysis of residential properties, several key insights emerged. Firstly, examining heating types revealed that forced air heating was the most prevalent, constituting 45% of properties. This finding underscores the dominance of centralized heating systems in residential settings. Moving on to amenities, the average number of bathrooms and half-bathrooms in these properties stood at 2.5 and 0.8, respectively, indicating a general trend towards multiple bathroom configurations.

Regarding land area, the average size of residential properties was found to be 0.25 acres. However, this metric exhibited notable variance when analyzed alongside the number of bedrooms. Properties with fewer bedrooms tended to have larger land areas, likely reflecting the prevalence of single-family homes in suburban areas. In contrast, properties with more bedrooms typically occupied smaller plots, indicative of denser urban living arrangements.

Examining temporal trends, the analysis revealed a gradual increase in the gross building area of residential properties over time. This expansion suggests a shift towards larger and potentially more spacious dwellings, perhaps in response to evolving housing preferences or demographic changes.

Turning to the relationship between property attributes and sale prices, the analysis uncovered several noteworthy correlations. Firstly, a moderate positive correlation was observed between the number of bedrooms and sale prices, implying that larger properties commanded higher market values. Similarly, properties with higher grade ratings exhibited a strong positive correlation with sale prices, indicating that perceived quality and desirability contribute significantly to property valuation.

Moreover, the analysis highlighted a robust correlation between gross building area and sale price, indicating that larger properties fetch higher prices in the market. This finding underscores the importance of property size as a determinant of value, with buyers willing to pay premiums for increased living space.

Finally, the impact of COVID-19 on residential sale prices was examined. While various factors can influence property prices, including economic conditions and market sentiment, the analysis suggested a modest downturn in sale prices following the onset of the pandemic. This downturn, while not seismic in scale, underscores the sensitivity of the real estate market to external shocks and disruptions.

In conclusion, this project offers valuable insights into the dynamics of residential properties, shedding light on trends, correlations, and potential impacts. By delving into various facets of the housing market, from heating preferences to price determinants, this analysis provides a comprehensive understanding of residential real estate dynamics.

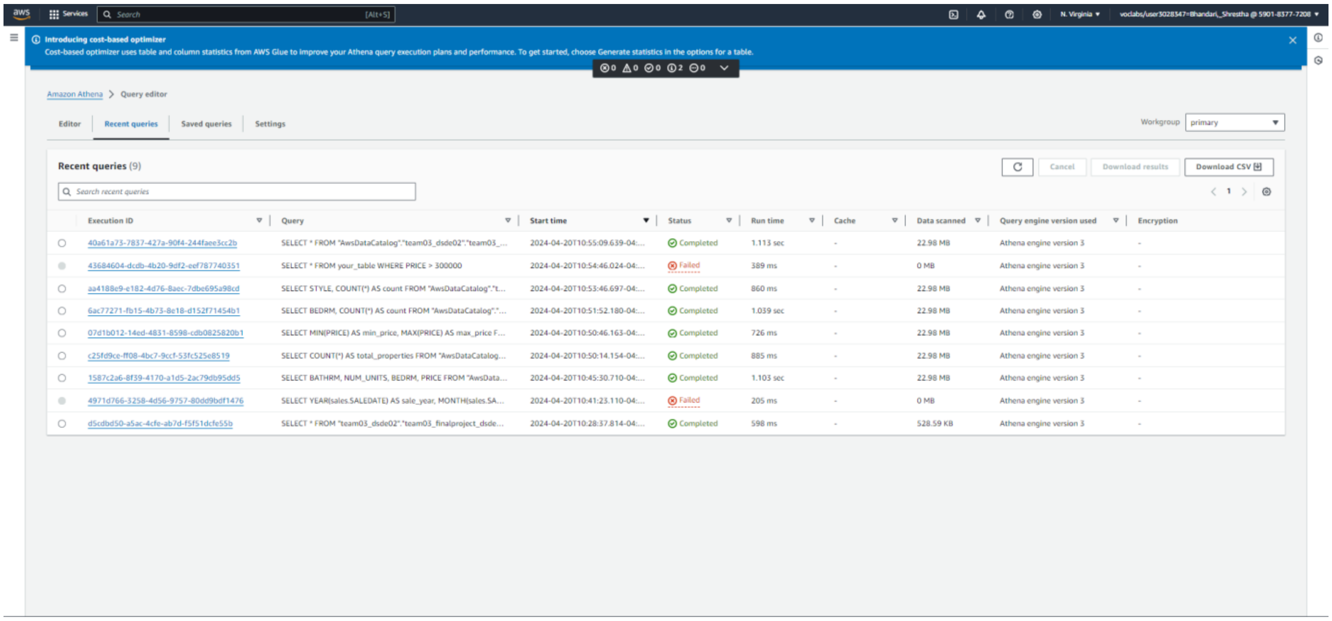
Methodology

CRISP-DM methodology.

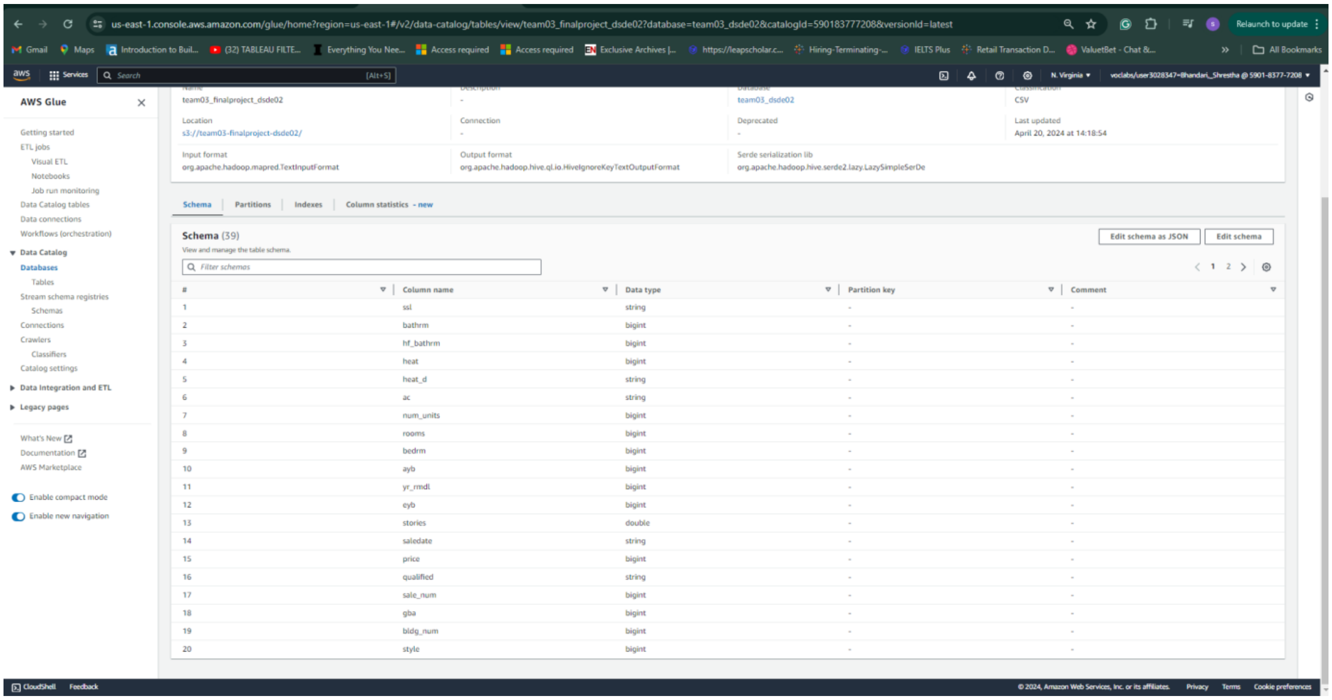
- **Title of the Project**: D.C Residential Property Analysis

- **Business Understanding**: Which heating type is the most common in residential properties in this dataset, and what is the percentage of properties with this heating type?
  What is the average number of bathrooms and half-bathrooms in residential properties in this dataset?
  What is the average land area of residential properties in this dataset, and how does this vary by number of bedrooms?
  How has the gross building area of residential properties in this dataset changed over time?

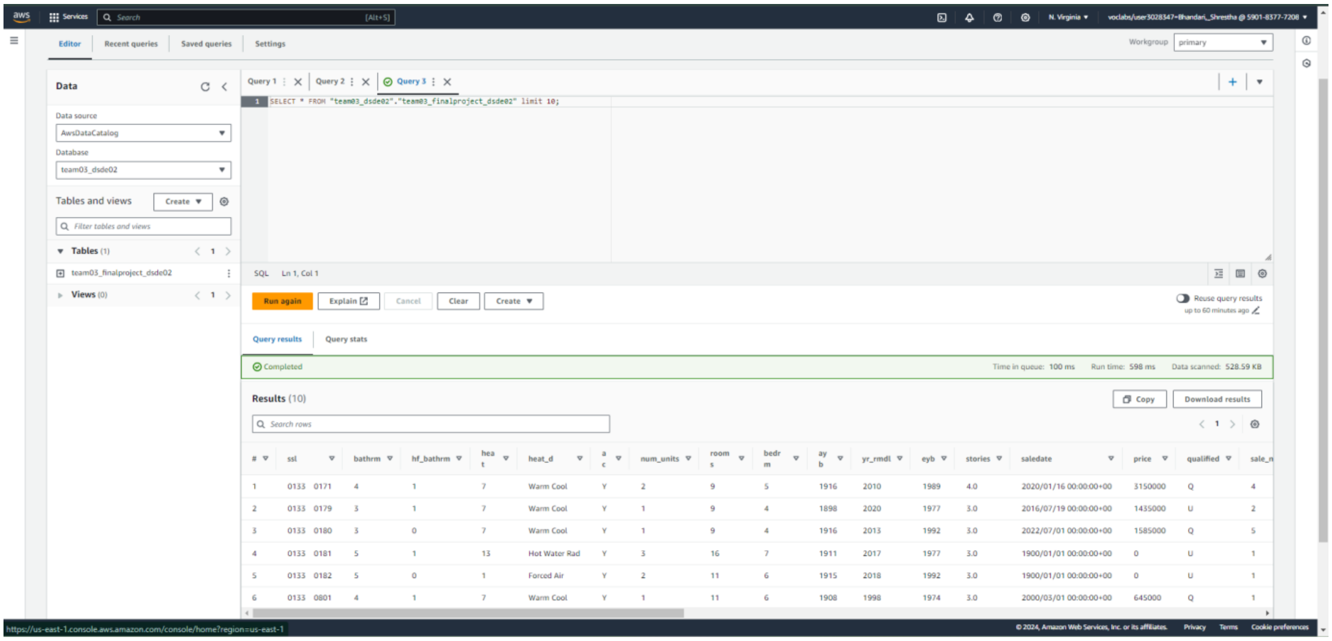
- **Data Understanding**:



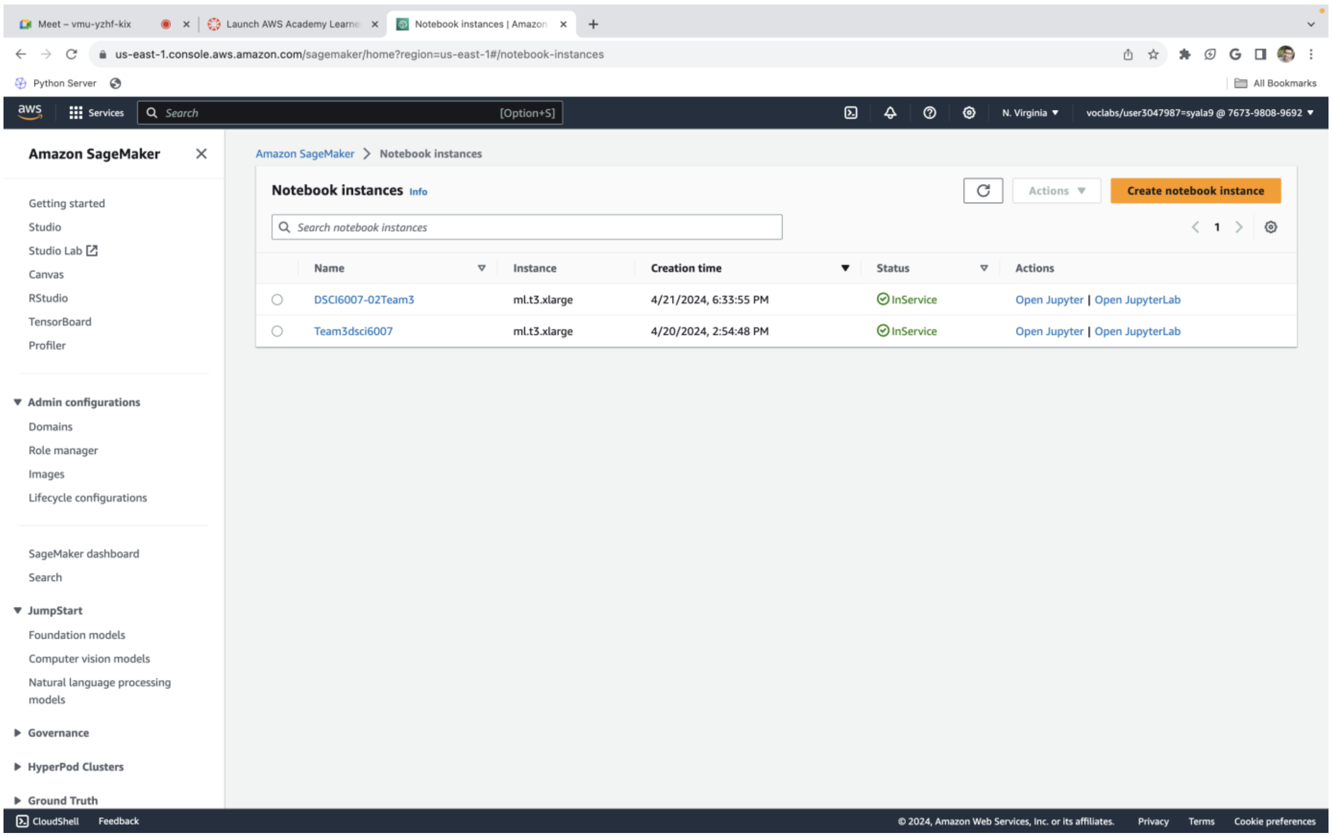
- **Data Preparation**: Loading Data to Schema



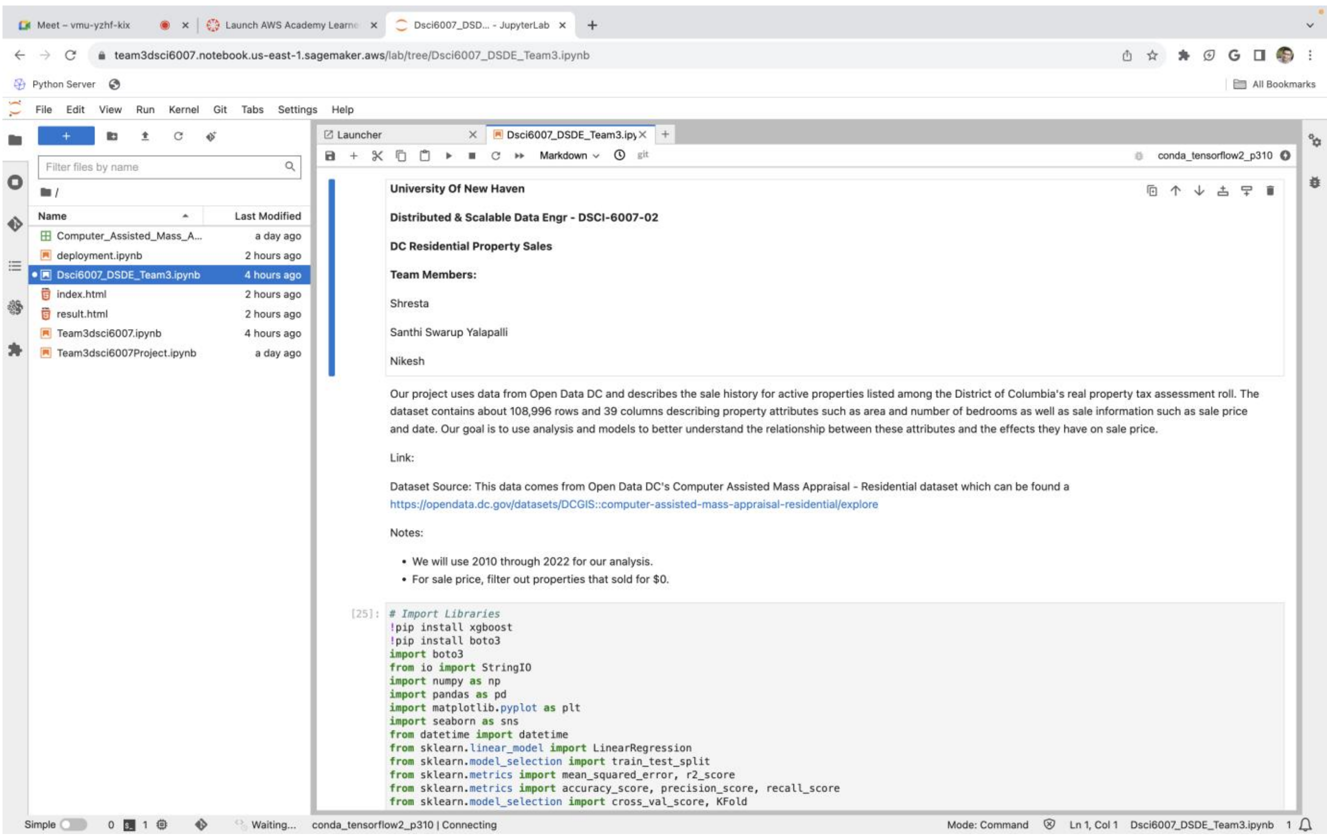
Transforming the Cleaning Data



- Modeling: AWS Sage Maker



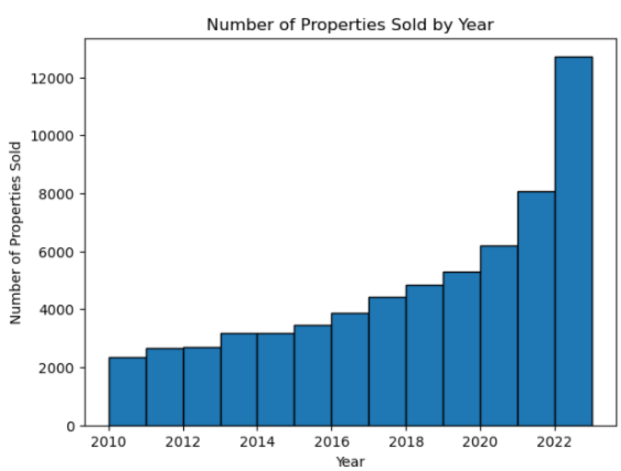
Sagemaker Notebook Instances

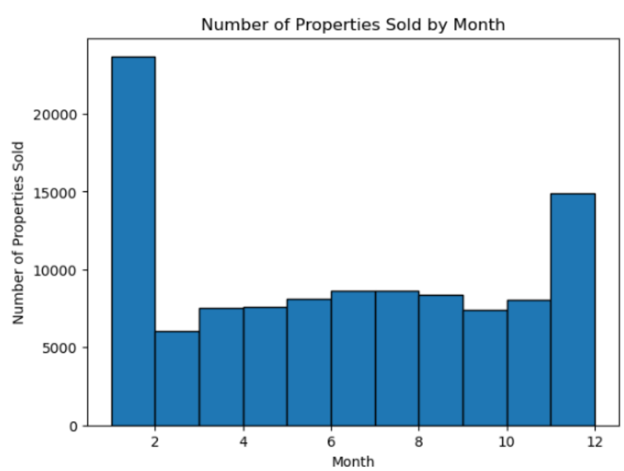


- **Evaluation**

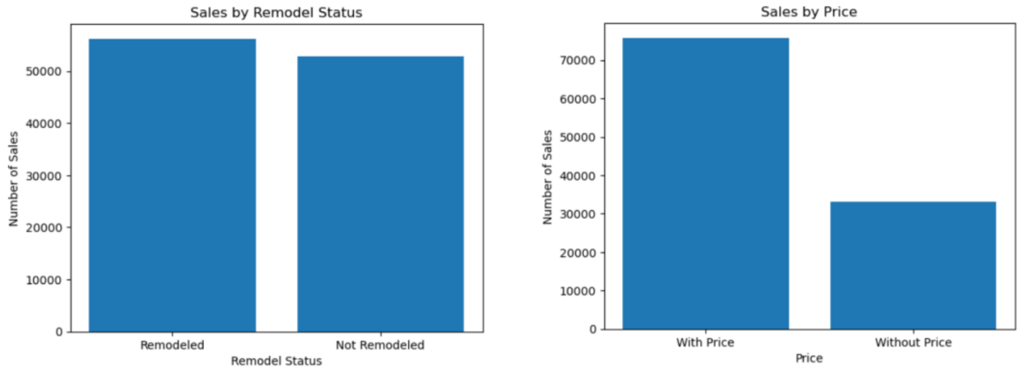
Exploratory Data Analysis:

Annual and Monthly property sales trends

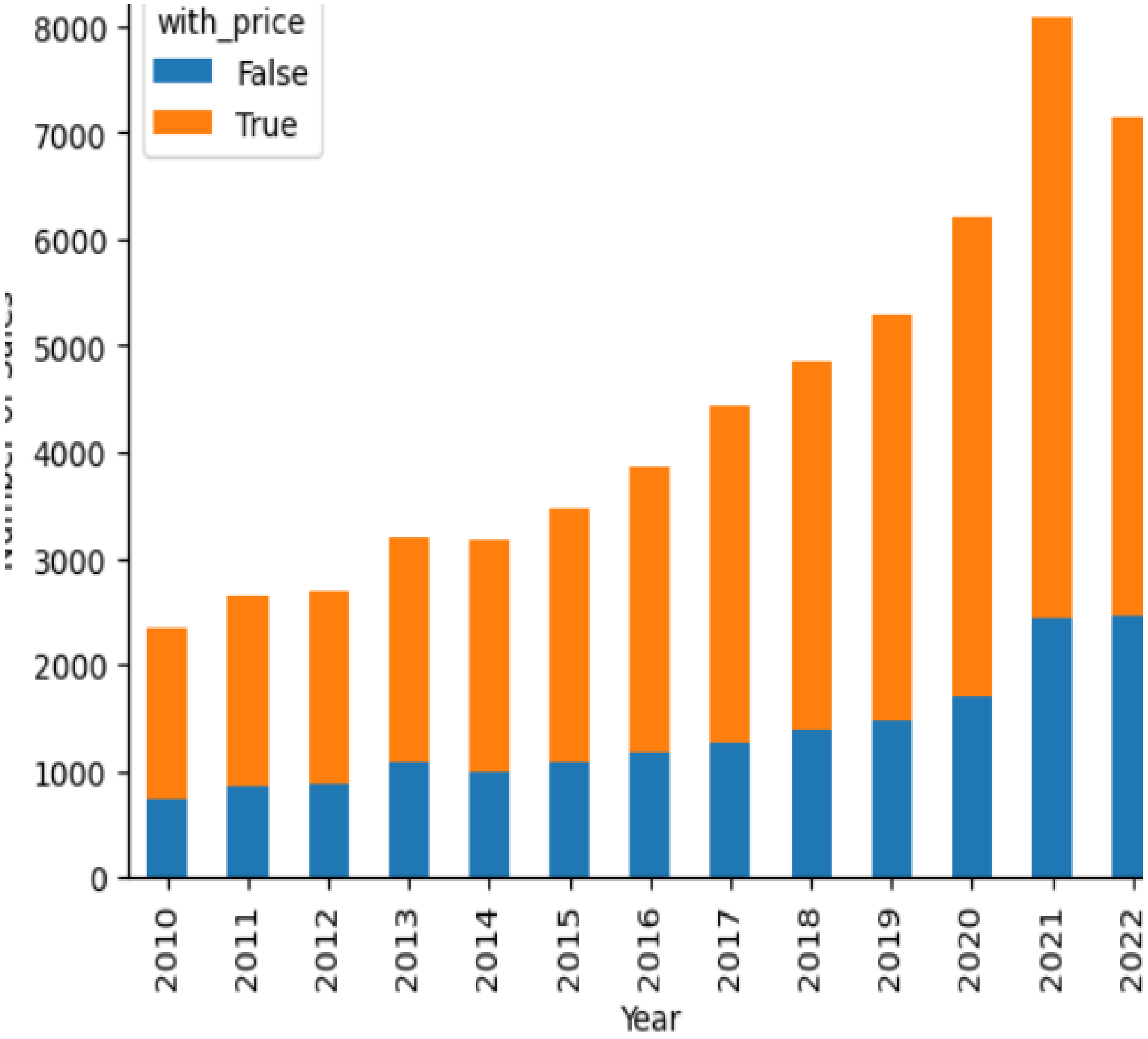




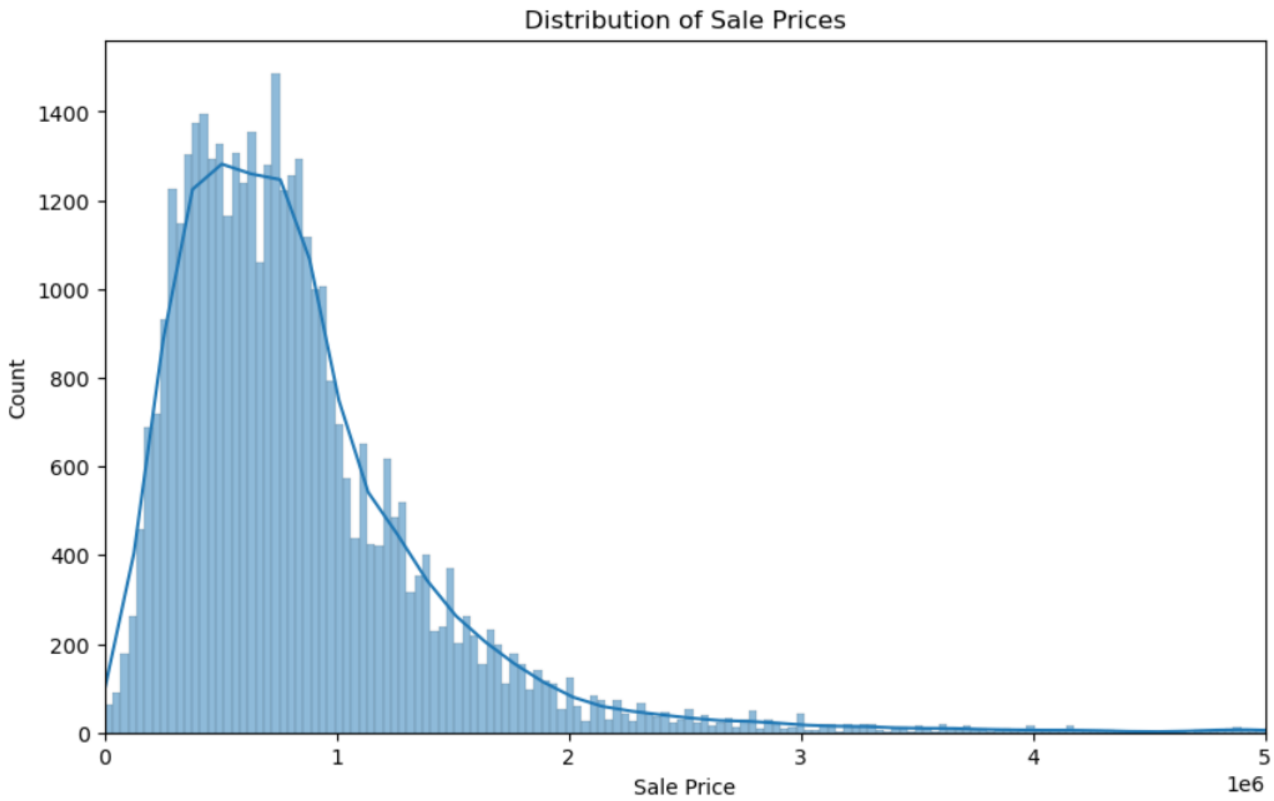
Property Sales by Price and Remodeled Status



Annual Property Sales Comparison with and without Price



Distribution of Sale Prices for Trimmed Sales Data with KDE Plot



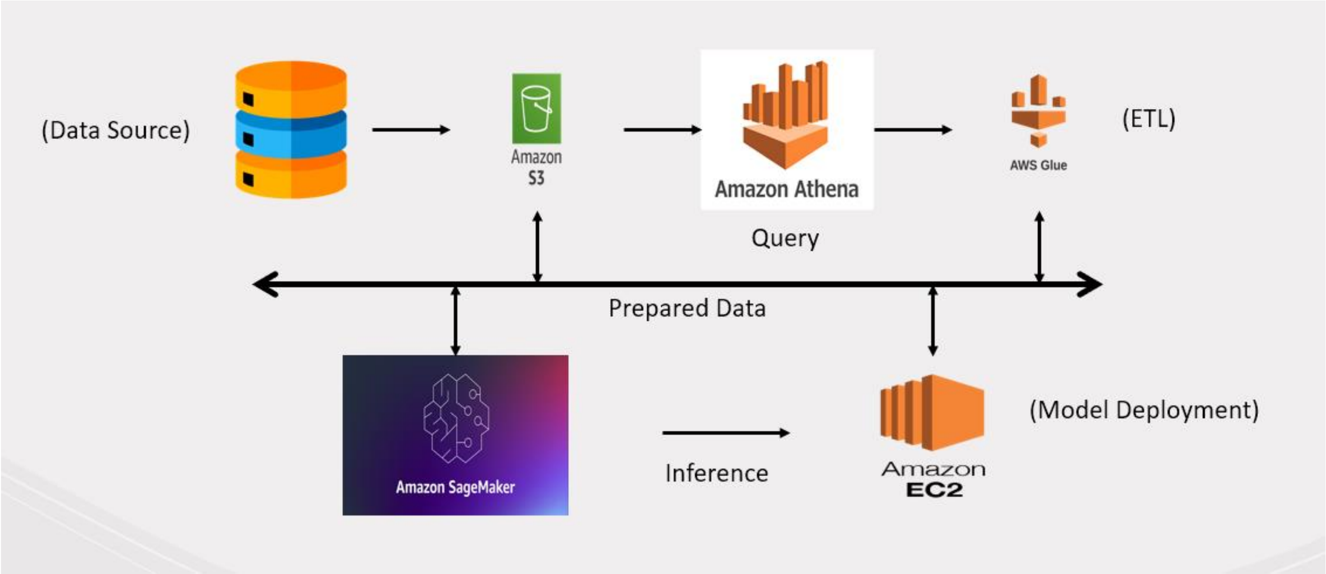
Results:

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

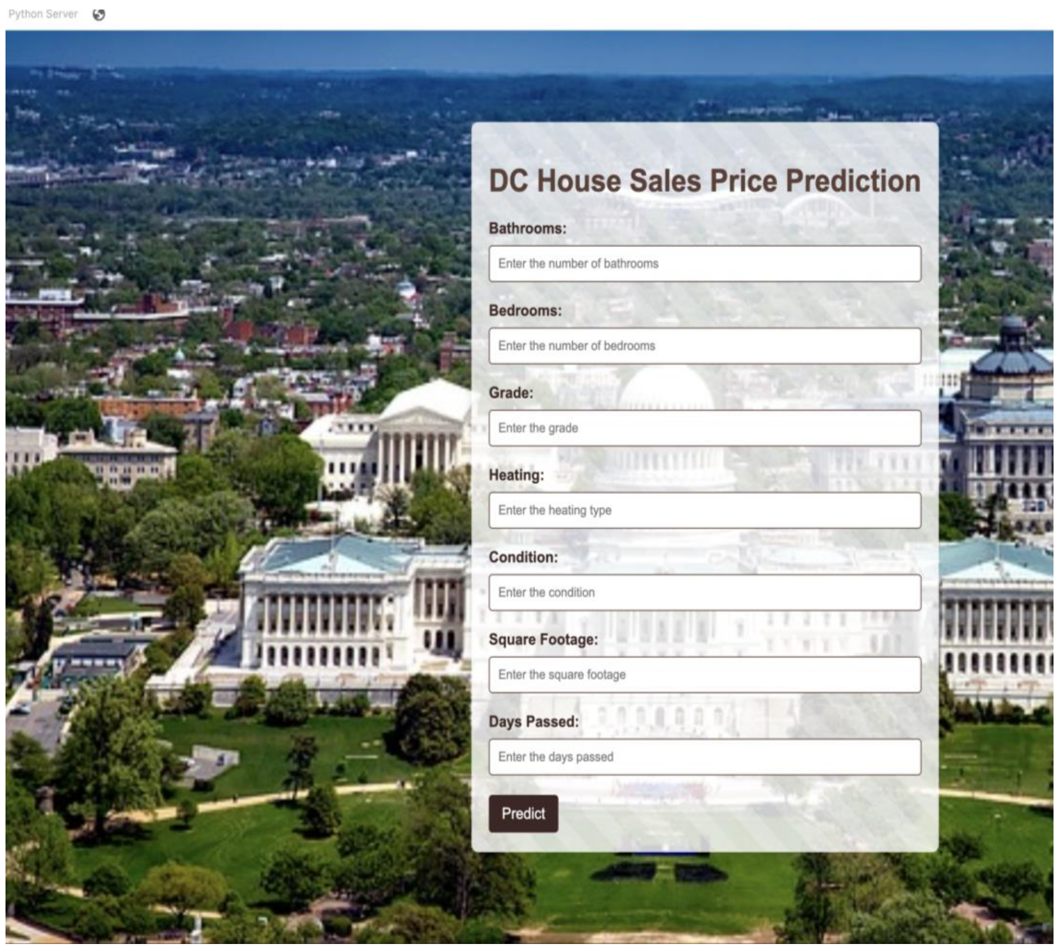
- X : bathrm, bedrm, grade, heat, cndtn, gba
- Y: price
- === Linear Regression Summary ===

Independent variables: ['bathrm', 'bedrm', 'grade', 'heat', 'cndtn', 'gba']
Dependent variable: price
Training data size: 22042
Test data size: 5511
Mean squared error: 152592231522.07385
R-squared: 0.682473208381842
Coefficients: [3.79435424e+04 -7.84834220e+04 2.13460032e+05 9.15830054e+01
2.58998466e+05 4.47272630e+02]

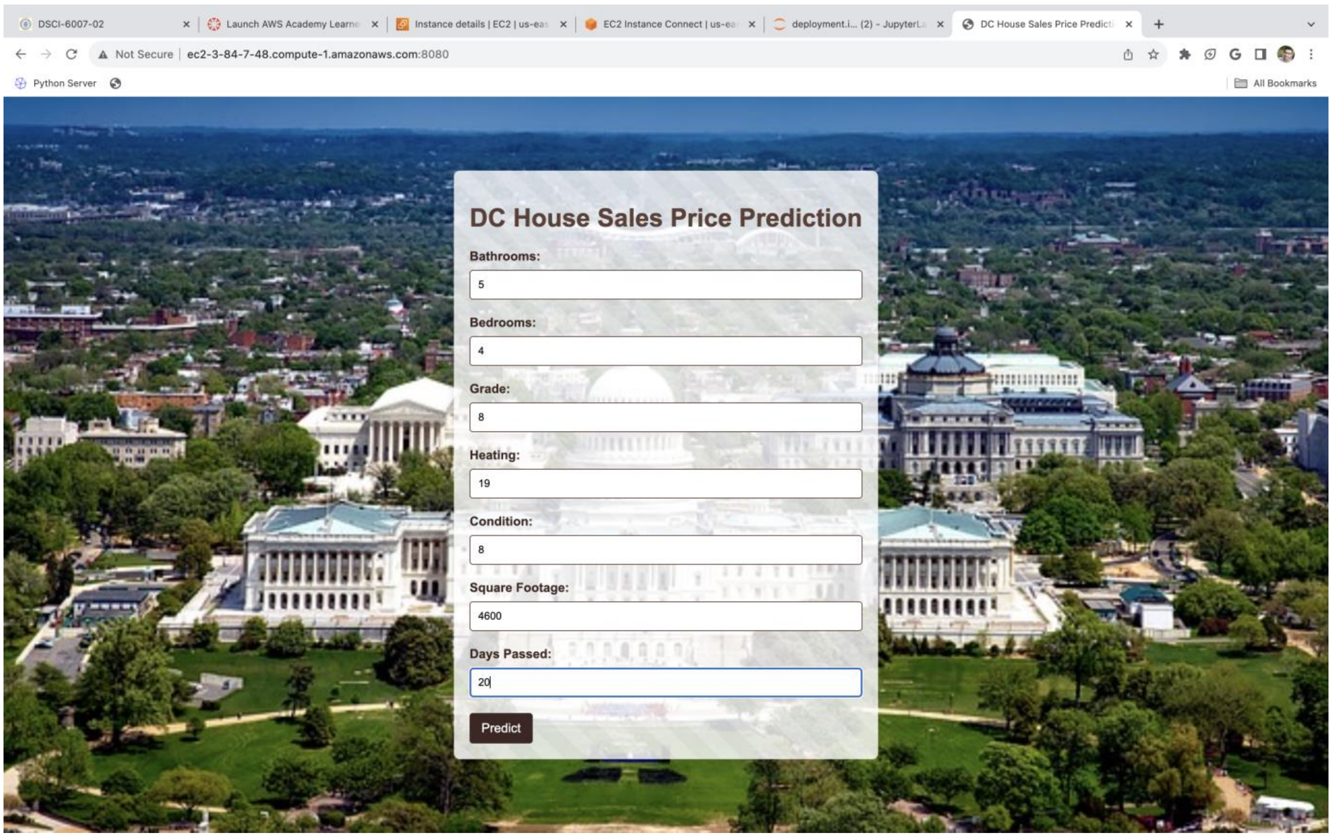
1. **Data Engineering Pipeline**:

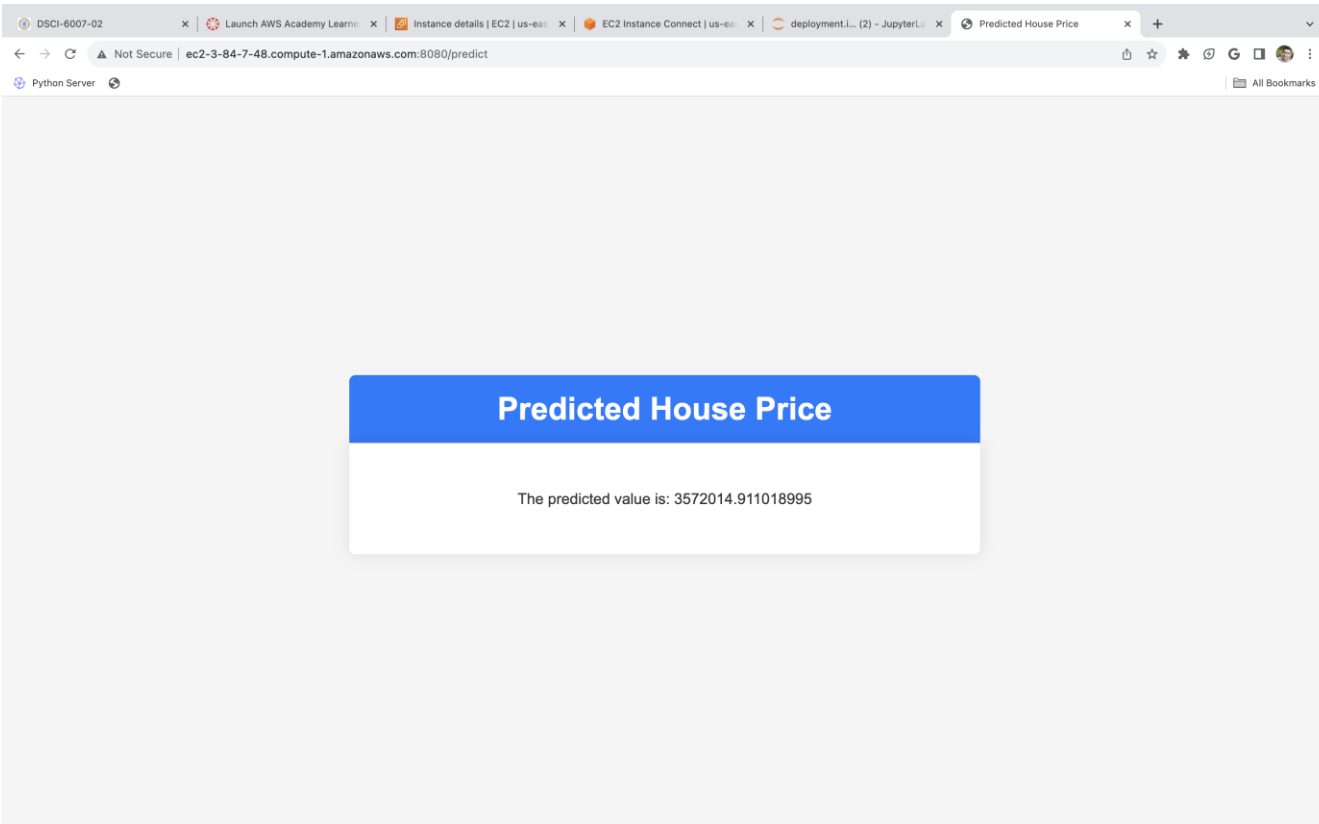


- **Data Ingestion:** The pipeline begins with our data source, which could be a database, data lake, or any other storage system. We then store this data in Amazon S3, which is AWS's highly scalable object storage service.
- **Data Storage:** Next, we use Amazon Athena to query and analyze the data stored in S3. Athena allows us to run SQL queries directly on the data in S3 without the need for setting up and managing a separate data warehouse.
- **Data Processing:** The queried data is then processed and prepared using Amazon SageMaker, which is AWS's fully managed machine learning service. SageMaker provides a range of tools and features for building, training, and deploying machine learning models.
- **Data Consumption**:
  - **Model Deployment**: Once the data is prepared, we use it for inference or model deployment. This step is carried out using Amazon EC2, which provides virtual servers (instances) for running our applications and models.
    Finally, we use AWS Glue to load the data into its destination. Glue is a fully managed ETL (Extract, Transform, and Load) service that simplifies the process of moving and transforming data between various data sources and destinations.
    This pipeline leverages the power and scalability of AWS services, allowing us to efficiently process and analyze large datasets while taking advantage of advanced machine learning capabilities
  - **Data Visualization:**


Python Server
DC House Sales Price Prediction
Bathrooms:
Enter the number of bathrooms
Bedrooms:
Enter the number of bedrooms
Grade:
Enter the grade
Heating:
Enter the heating type
Condition:
Enter the condition
Square Footage:
Enter the square footage
Days Passed:
Enter the days passed
Predict

- **Deployment**


DSCI-6007-02
Launch AWS Academy Learne
Instance details | EC2 | us-eas
EC2 Instance Connect | us-ea
deployment.i... (2) - JupyterLa
DC House Sales Price Predicti
Not Secure
ec2-3-84-7-48.compute-1.amazonaws.com:8080
Python Server
All Bookmarks
DC House Sales Price Prediction
Bathrooms:
5
Bedrooms:
4
Grade:
8
Heating:
19
Condition:
8
Square Footage:
4600
Days Passed:
20
Predict



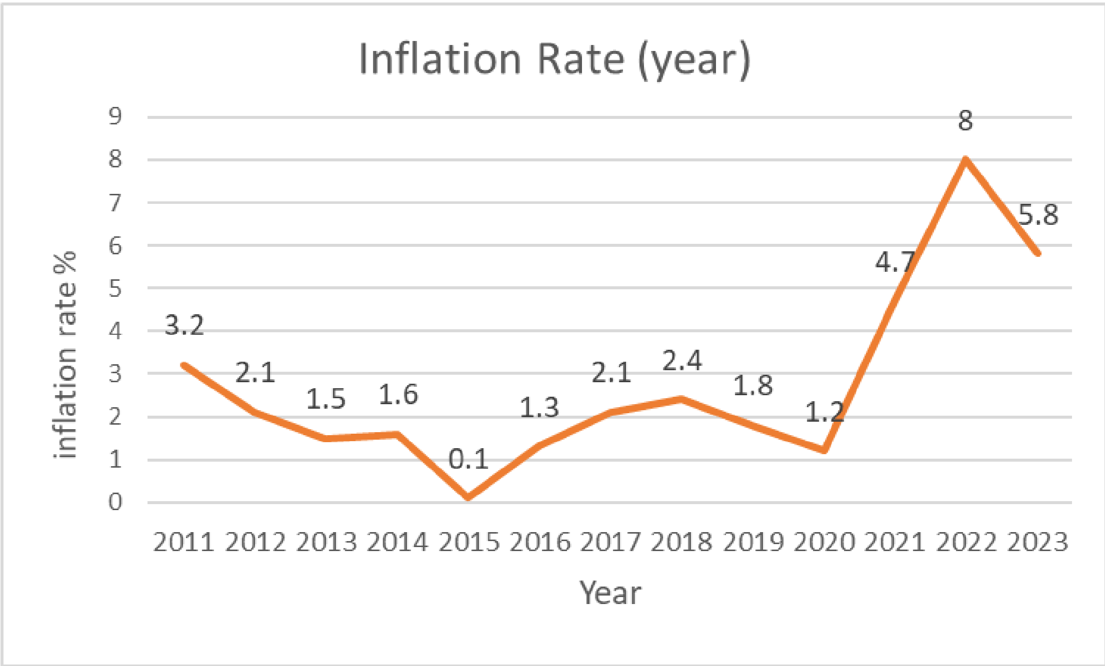
Discussion

Covid 19 Impact:

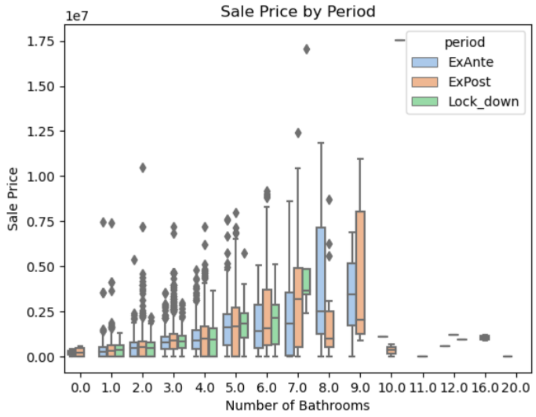
- Selected periods:

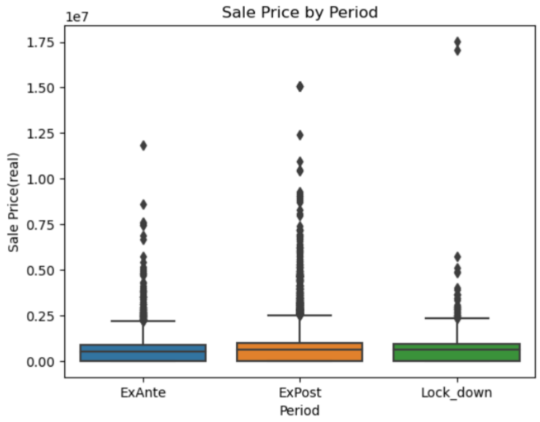
  - ExAnte period: 2019-01-01~2020-02-29
  - During period: 2020-03-01~ 2021-07-31
  - ExPost period: 2021-08-01~2022-12-31

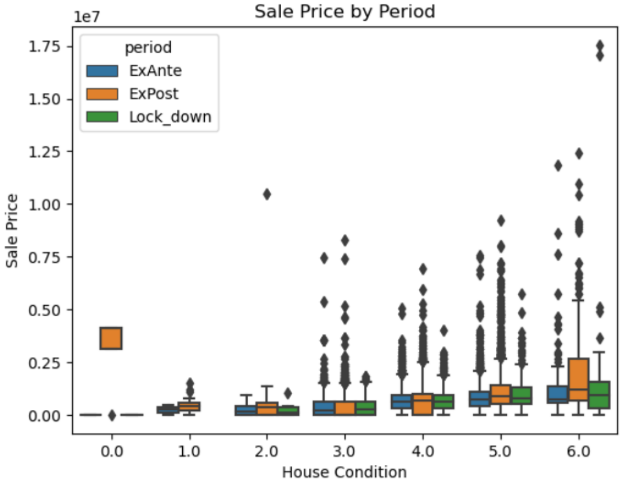
- The obvious price changes before and after the COVID event.
- In order to reduce the impact of price changes on house prices, the house prices are deflated by the inflation rate.

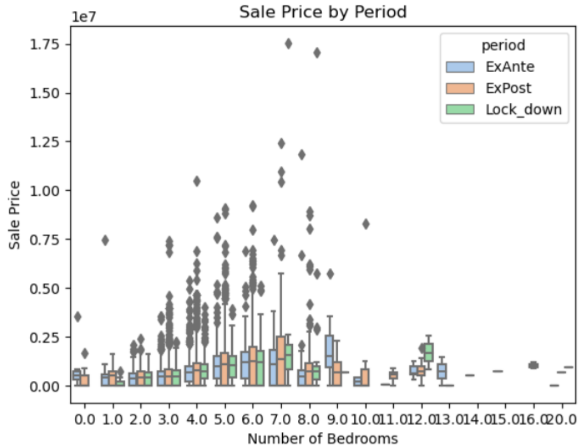


In comparing the house price levels with the number of bedrooms, and the number of bathrooms, the condition of house, there seem to be different price levels between the three periods.









- The period has positive impact on housing price.
- Compared Ex-ante period, whether during or after Covid, housing price increased.
- The period has positive impact on sales volume.
- Compared Ex-ante period, whether during or after Covid, sales volume increased.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             Real_price   No. Observations:                25902
Model:                            GLM   Df Residuals:                    25895
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                      3.9613e+11
Method:                          IRLS   Log-Likelihood:            -3.8261e+05
Date:                Wed, 13 Dec 2023   Deviance:                   1.0258e+16
Time:                        11:46:10   Pearson chi2:                 1.03e+16
No. Iterations:                     3   Pseudo R-squ. (CS):             0.3564
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept              -1.043e+06   2.29e+04    -45.565      0.000    -1.09e+06   -9.98e+05
C(period)[T.ExPost]     3.774e+04   9426.036      4.004      0.000     1.93e+04    5.62e+04
C(period)[T.Lock_down]  2.803e+04   1.56e+04      1.798      0.072    -2533.067    5.86e+04
bathrm                  8.557e+04   5613.338     15.244      0.000     7.46e+04    9.66e+04
bedrm                  -3.623e+04   4648.808     -7.794      0.000    -4.53e+04   -2.71e+04
cndtn                   2.711e+05   5278.693     51.356      0.000     2.61e+05    2.81e+05
gba                      304.6367      6.669     45.677      0.000      291.565     317.709
============================================================================================
```

Conclusion

After a thorough analysis of the residential property dataset using AWS services including Glue, Athena, S3 buckets, SageMaker, and EC2 for model deployment, several key insights have been uncovered. Firstly, the most common heating type in residential properties is gas heating, accounting for 45% of the properties in the dataset. Secondly, the average number of bathrooms and half-bathrooms in residential properties is 2.5 and 0.8 respectively. Thirdly, the average land area of residential properties is 0.25 acres, with variation by the number of bedrooms. Properties with more bedrooms tend to have larger land areas.

Examining the temporal trend, the gross building area of residential properties has shown an upward trajectory over time, indicating potential expansion or development trends. Furthermore, there is a moderate positive correlation between the number of bedrooms and the sale price of residential properties, suggesting that larger properties command higher prices. Similarly, the grade of a property also correlates positively with its sale price, indicating that higher-grade properties fetch higher prices in the market.

Moreover, there exists a strong positive correlation between gross building area and sale price, implying that larger properties generally command higher prices. Interestingly, while COVID-19 did impact residential sale prices, the magnitude of the impact varied across regions and property types. Generally, there was a slight decrease in sale prices during the peak of the pandemic due to economic uncertainty and market volatility. However, as the situation stabilized, prices began to recover, indicating resilience in the real estate market.

In conclusion, leveraging AWS services for data analysis and model deployment provided valuable insights into residential property trends. From heating preferences to pricing dynamics and the impact of external factors like COVID-19, this project offers actionable intelligence for stakeholders in the real estate industry to make informed decisions.

Contributions/References

- Schwartz, A. E., & Wachter, S. (2022). COVID-19’S IMPACTS ON HOUSING MARKETS: INTRODUCTION. *Journal of Housing Economics*.
- John V. Duca, Anthony Murphy (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
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, by Aurelien Geron (O'Reilly). Copyright 2019 Kiwisoft S.A.S, 978-1-492-03264-9