Housing Price Analysis

Group 4
Pranav, Priya, Sheel, Sirjana

Reason why we selected this topic?

- Housing market is a hot topic and considered as best long term investment
- House price trend always a concern for majority of population

Question we hope to answer

- Analyse and predict house price based on multiple factors such as Year, Mortgage rate and Immigration
- Provide expected house price based on user input

Data Source

HPI Data: A monthly seasonally adjusted housing prices showing composite HPI and Composite Benchmark was extracted from CMHC (Canada Mortgage and Housing Corporation) from January 2005 to September 2022

Mortgage Data: A dataset showing monthly mortgage rate from 1951 - 2022 was collected from Kaggle.com.

Immigrants Data: Quarterly information on immigrants was extracted from Statistics Canada.

Tools and Technology

Preprocessing Data : Python, Pandas

DataBase : PostgreSQL, SQLAlchemy

Machine Learning: Python - pandas, sklearn, train_test_split, matplotlib

Visualisation: Flask, Plotly, json, Javascript, HTML

Presentation: Google Slides

Quick Database Diagram

Data Set

- Mortgage_rates.csv

3 columns and 858 rows

- house_price.xlsx

13 columns and 214 rows

- immigration.xlsx

30 rows and 130 columns

Summary of Dataframe

	Max	Min	Average
House Price (CAD)	840,000.00	221,100.0	438,980.28
No. of Immigrant	138,190.0	34,070.0	63,574.82
Mortgage Rate (%)	6.81	3.20	4.5

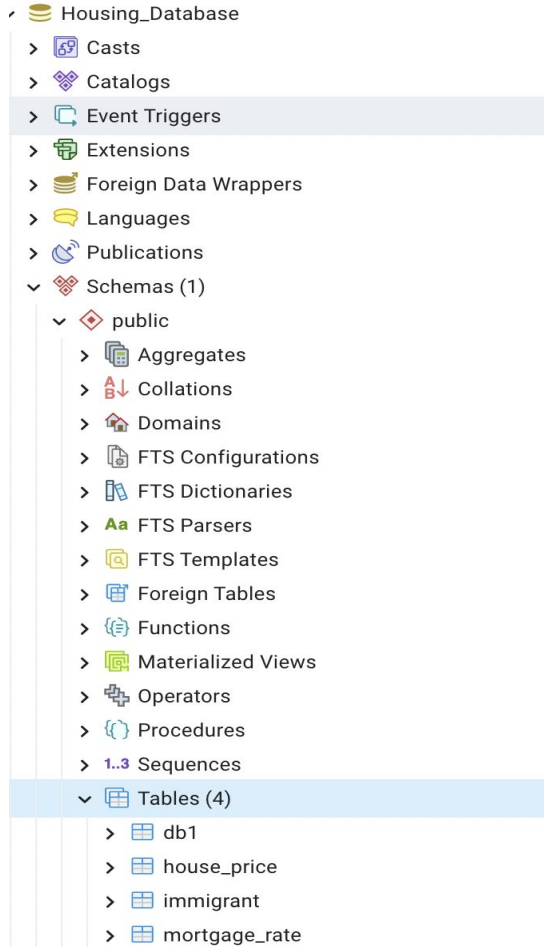
Data Preprocessing

- Cleaned table
  - Date column in mm-dd-yy format
  - Price column as the dependent variable
  - Final tables as csv file
  - Quarterly date changed to monthly format
  - Dropped unnecessary columns
  - Filled missing values for immigrants data
  - Filtered data from 2000-05-01 to 2021-12-01

ERD & Database

- Saved and exported three clean dataframe to PgAdmin using connection string

```
# Save immigrant dataframe(df) to SQL
df3.to_sql(name='immigrant', con=engine,if_exists='replace',index=False)
```



Final Database

- Created final database by joining three tables on date column

Date	HPI	Price	Immigrants	Mortgage Rate
2005-01	100.0	221100	18812.666666666700	5.6
2005-02	100.6	222500	18812.666666666700	5.59
2005-03	101.4	224200	18812.666666666700	5.6
2005-04	102.2	225900	24823.666666666700	5.67
2005-05	102.8	227400	24823.666666666700	5.55
2005-06	103.8	229600	24823.666666666700	5.31
2005-07	105.1	232400	25315.0	5.26
2005-08	106.5	235400	25315.0	5.32
2005-09	107.9	238500	25315.0	5.3
2005-10	109.4	241900	18462.0	5.39
2005-11	111.0	245400	18462.0	5.56
2005-12	112.4	248500	18462.0	5.6
2006-01	114.1	252200	18378.0	5.65
2006-02	115.6	255600	18378.0	5.75
2006-03	117.2	259100	18378.0	5.78
2006-04	119.0	263000	22636.33333333330	5.88

Final Database

Imported to Pandas using connection string to connect it to machine learning model and visualization dashboard

```
connect_string = f"postgresql://postgres:zunu1900@127.0.0.1:5432/Housing_Database"
```

```
engine=create_engine(connect_string)
data = pd.read_sql("SELECT * FROM db1", engine)
print(f"Got dataframe with {len(data)} entries")
```

Machine Learning Model

Preliminary data processing

- Date column was changed to ordinal
- HPI column was dropped

Feature Selection

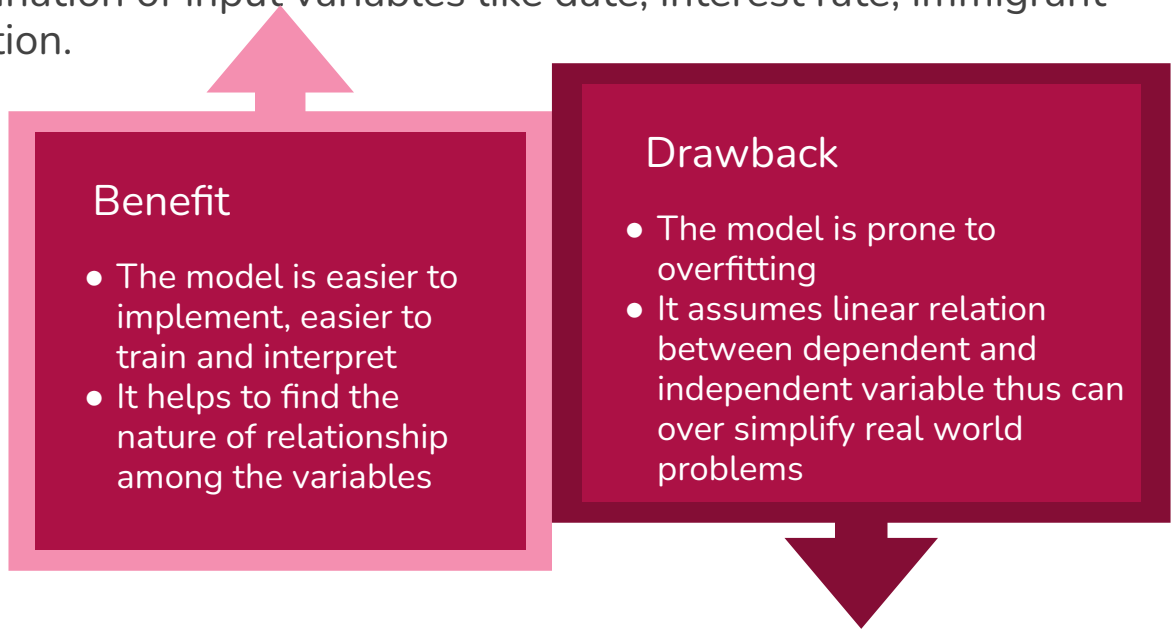
- Target variable: Price
- Features: rate, immigrants, date

Training and Testing split

- Data were split into train and test size in the ratio of 80:20

Model Choice

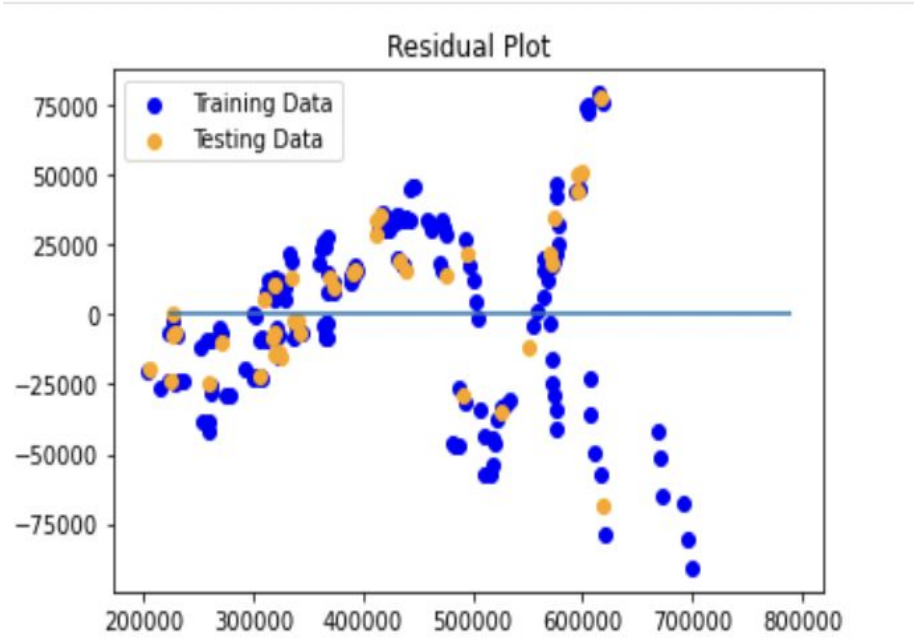
Linear Regression Model is used as we are predicting house price based on a combination of input variables like date, interest rate, immigrant population.



Main Analysis-Findings

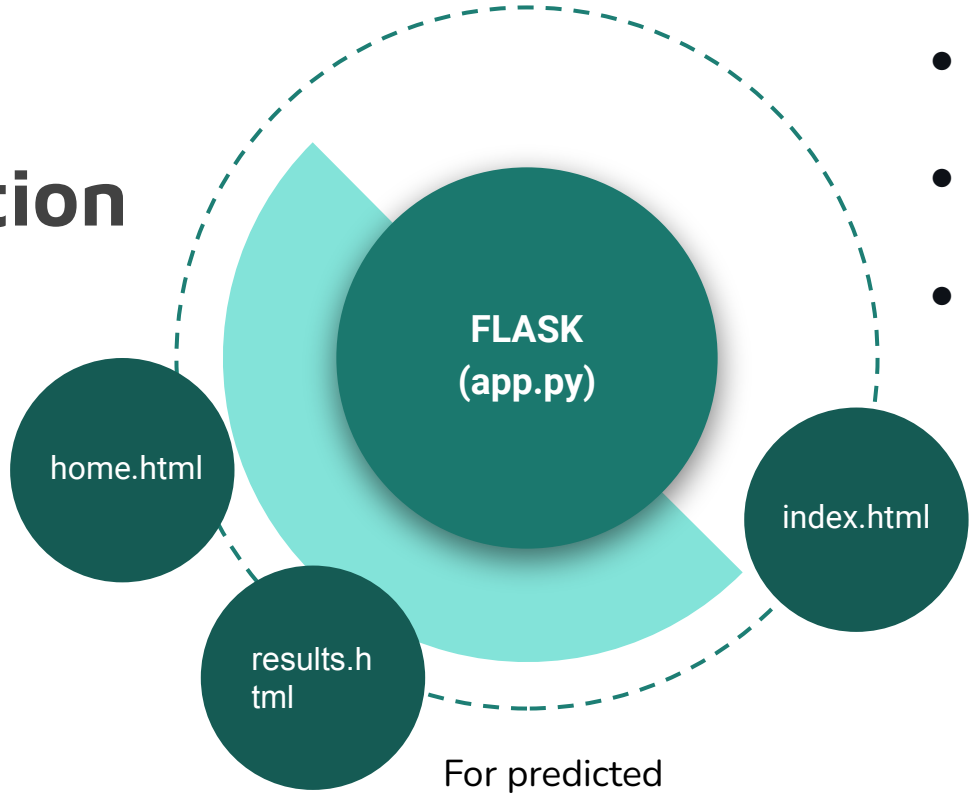
Linear Regression Model was run and following are the results from the Model

- R-squared: 0.94358



Visualization

- Flask; to create route for web browser
- HTML; to build webpage
- Plotly; to build interactive chart
- CSS; to style html

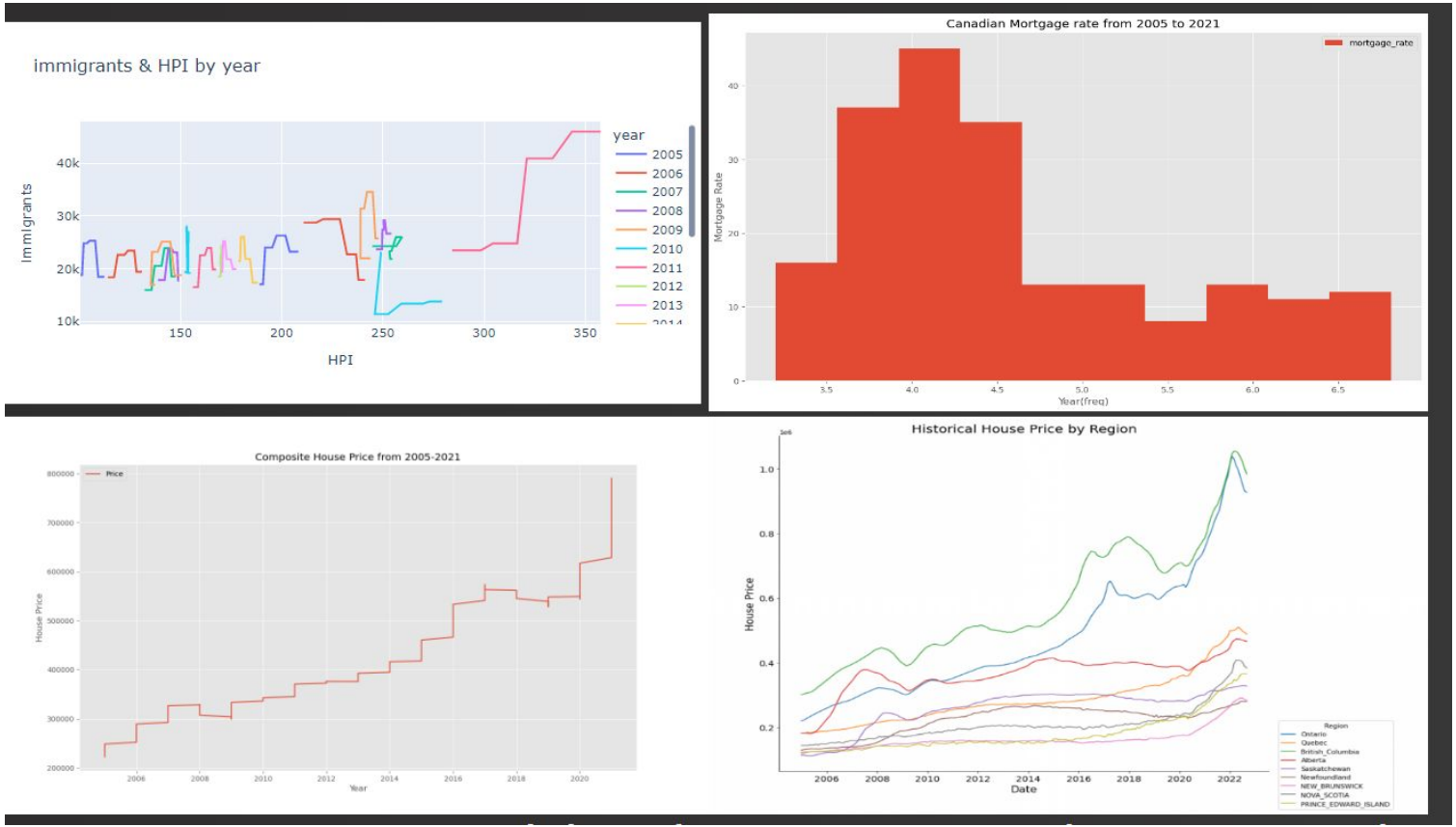


For machine learning prediction webpage

For main webpage

For predicted value

Exploratory Data Analysis



Final Dashboard

Machine Learning Prediction

Please enter the parameters

Date 2025-07-23

Mortgage_rate 7.5

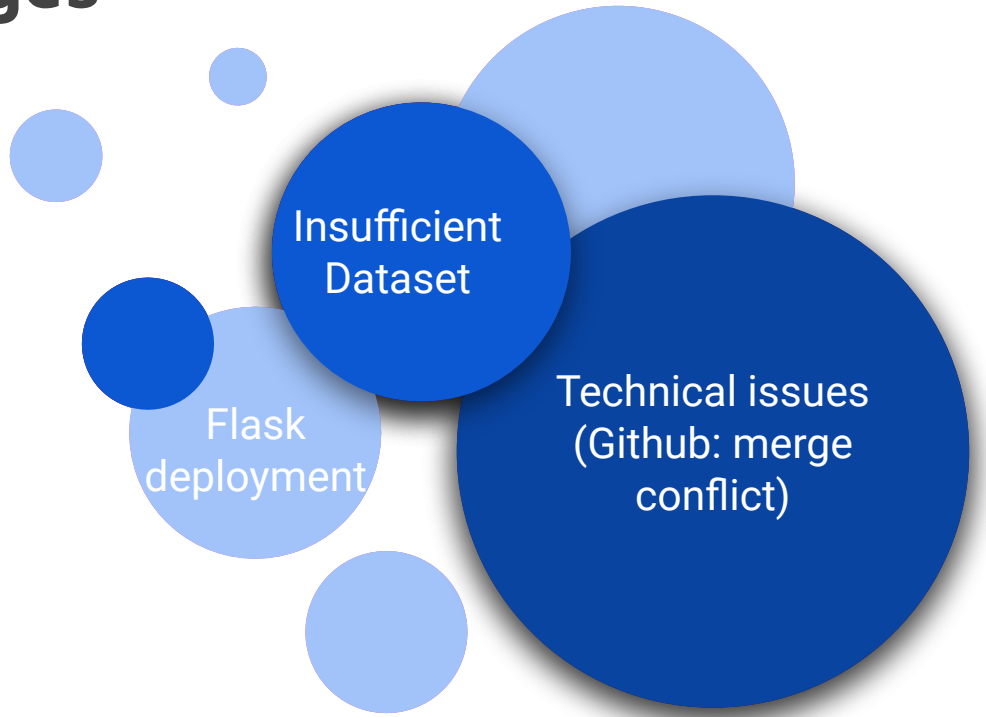
Immigrants

Predict

Results Sheet

Predicited value :798615 CAD with 94 % confidence

Challenges


Insufficient Dataset
Technical issues (Github: merge conflict)
Flask deployment

Recommendations for Future Analysis

Time series Analysis

Include daily data to conduct time series analysis

Include more features in the dataset

Include more features like region, income level, house types in the analysis

THANK YOU