Brooklyn Housing Analysis

Using Machine Learning to predict future Brooklyn housing prices

Why

To predict future sales prices

Cost of living continues to increase as does the sales price of homes. By identifying trends we can make more educated decisions on things such as investment value, future affordability, and cost of living.

Data Source

Brooklyn Home Sales, 2003 to 2017

https://www.kaggle.com/datasets/tianhwu/brooklynhomes2003to2017

Data used in analysis was Brooklyn property sales from 2003-2017 sourced from kaggle.com by Tommy Wu. Original data was taken from the City of New York and cleaned up to be used for analysis. The original data was gathered to cover all 5 Boroughs of New York.

Top 5 Questions:

Question 1	Can we predict future home prices in Brooklyn based on the month of year the home was bought?
Question 2	Does square footage of Brooklyn homes drastically affect the price of the home?
Question 3	How do the neighborhoods compare in pricing within Brooklyn alone?
Question 4	How have prices fluctuated within a specific range of years?
Question 5	How have prices fluctuated within a specific range of years?

Data Cleaning and Analysis

Technologies Used

- SQL and PostgresSQL
- Excell
- Pandas
- Python and Plotly
- SciKitLearning (Linear Regression Model)
- Tableau

Exploration

- SQL and Pandas: used to clean the Database. Remove excess columns and rows and null values.
- PostgresSQL: Store the Database

Analysis

- Python and Plotly: create Visualizations of data.
- SciKitLearn: ML library used to analyze the data and create future models

Issues and Resolutions of Data Exploration

- Removed columns that would not be useful
- Made list of all unique values in the building_class column
- Explored NYC's building info to identify different building classes and determine differences - which ones were residential and which ones were most "alike".
- This info allowed us to identify "family homes" as the largest population size while being the most similar.

Issues and Resolutions of Data Analysis

- How data was split into training and testing sets: Created Views in Postgres
- Why we chose Linear Regression: We are testing only point of change - there was not enough data to use Random Forest.
- Went from CSV to Postgres. Had to change data types for example SALE_PRICE field was a string and needed to be changed an integer.
- Ran data off the csv and the ML model was inaccurate with .005% however, once we changed the data types the new percentage moved closer with .37%.

Impact