An Investigation into the Factors Driving Housing Price Growth in the New York Boroughs

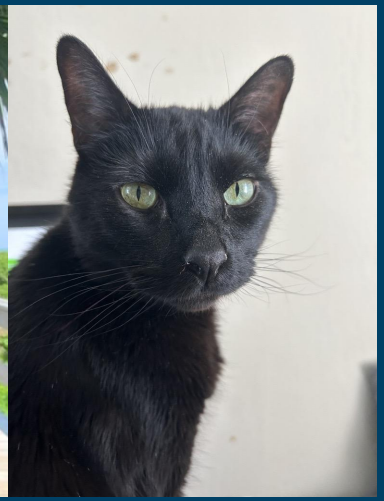
Illidan's Ingenious Infrastructure

Introduction to Team

Will Crocker

Nicholas Fenech

Prateek Bardhan



Illidan

(Will’s cat)

Introduction

- This project focused on investigating the New York City housing market
- Attempted to determine how various factors could impact prices
- Examined trends between sale prices and several different contributing features
- Quantified the extent of the drop in housing affordability

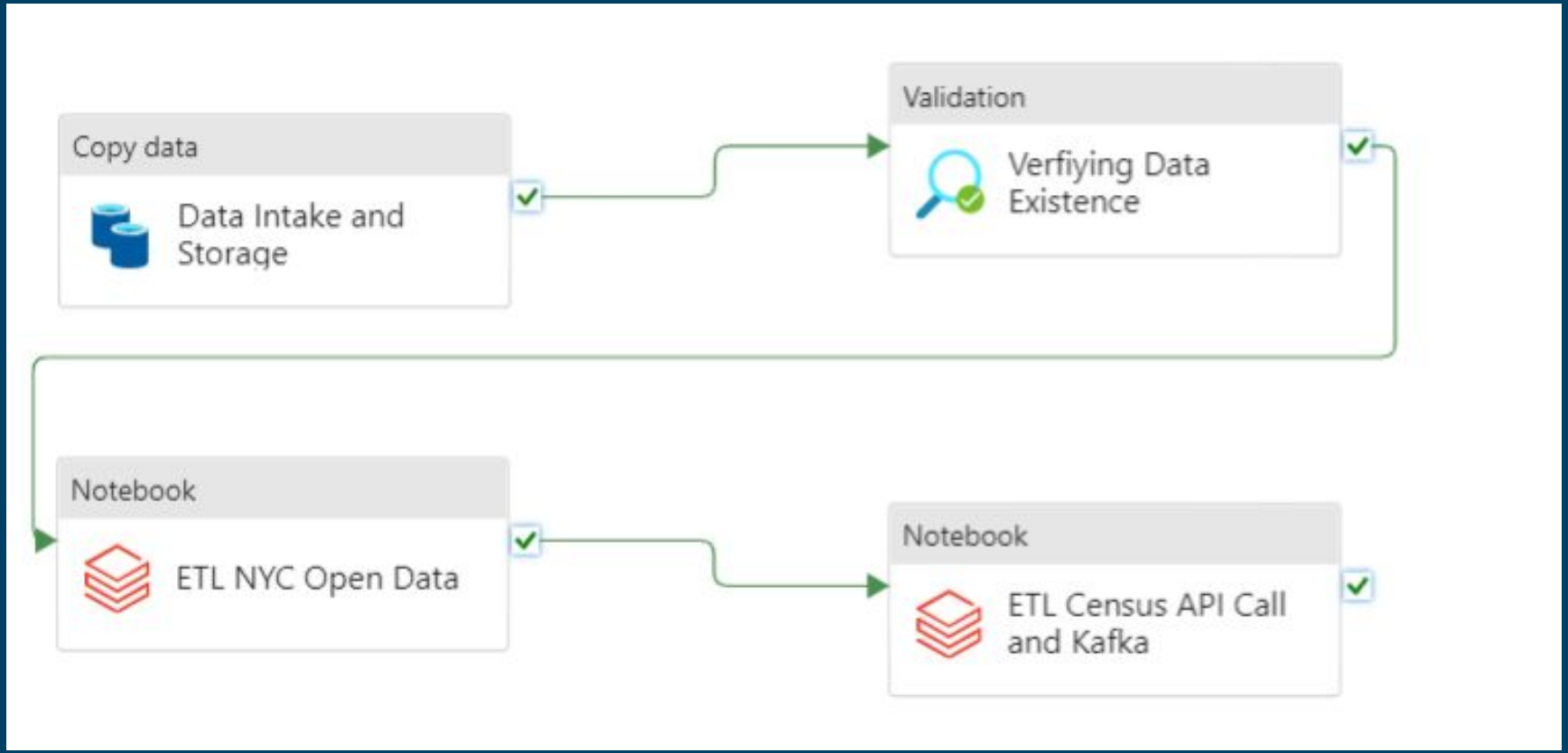
Sources

- NYC Open Data
  - Contains data from the past 20 years of housing sales by borough
  - Also shows many useful descriptive factors about each property
  - https://www.nyc.gov/site/finance/taxes/property-annualized-sales-update.page
- Census Bureau
  - American Community Survey 5-Year
  - Survey collects information on housing attributes, transportation, demographics, and economic factors
  - Contains information from federal to local municipality levels
  - https://www.census.gov/data/developers/data-sets/acs-5year.html

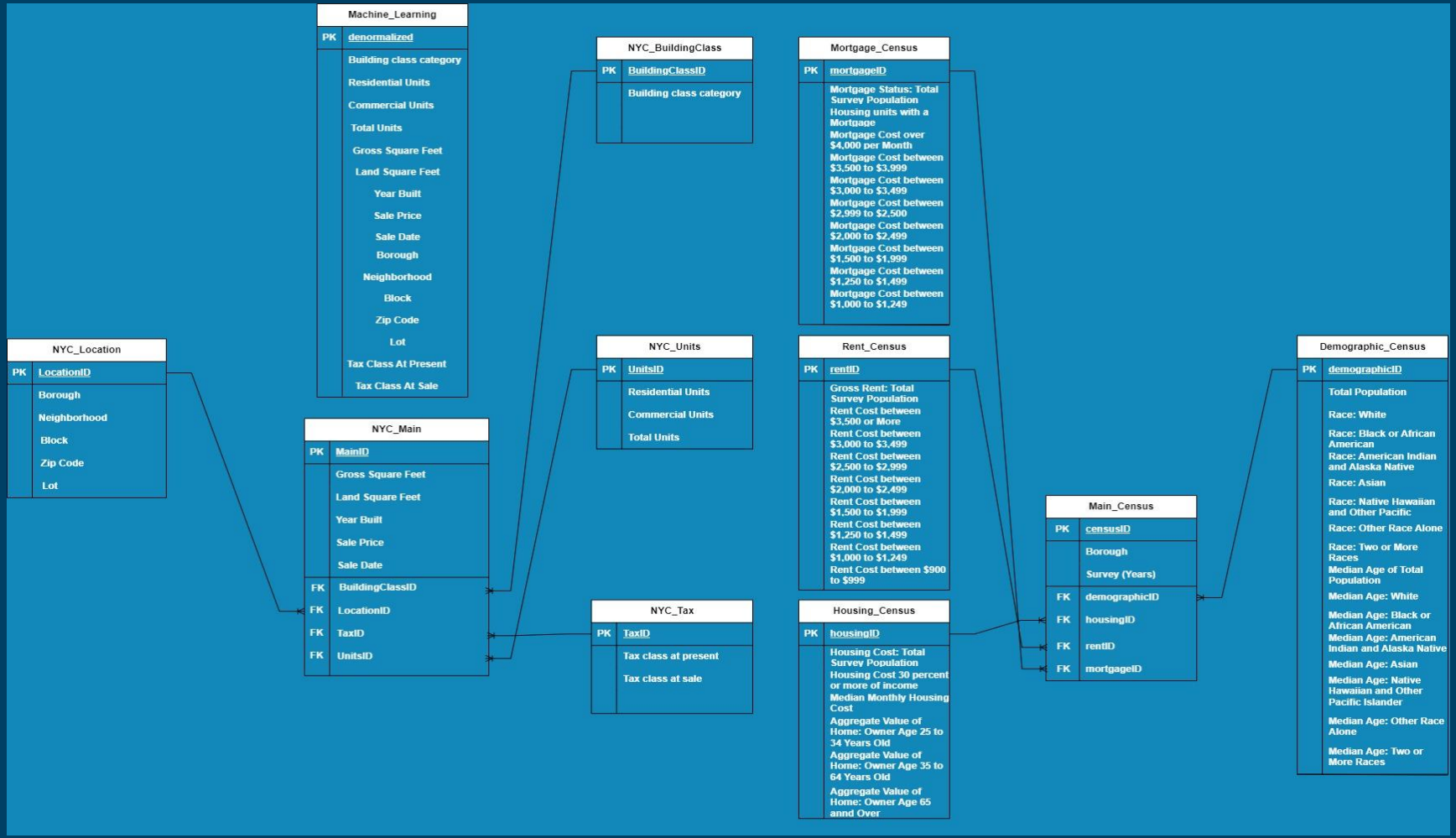
Technologies Utilized to Investigate Housing Prices

- Microsoft Azure
- Databricks
- Kafka
- SQL Database
- PowerBI
- Machine Learning Algorithm
- Python and PySpark

ETL Diagram


Copy data
Data Intake and Storage
Validation
Verfiying Data Existence
Notebook
ETL NYC Open Data
Notebook
ETL Census API Call and Kafka

ERD for SQL Database

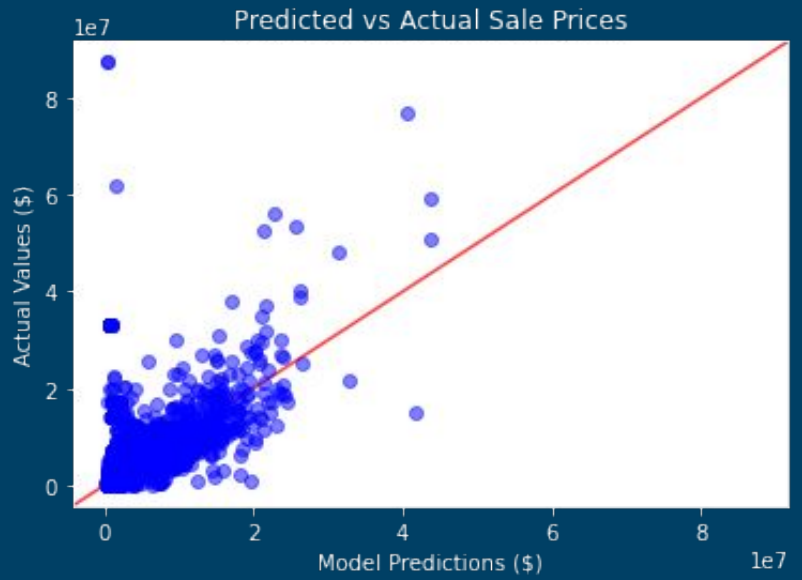


Machine Learning

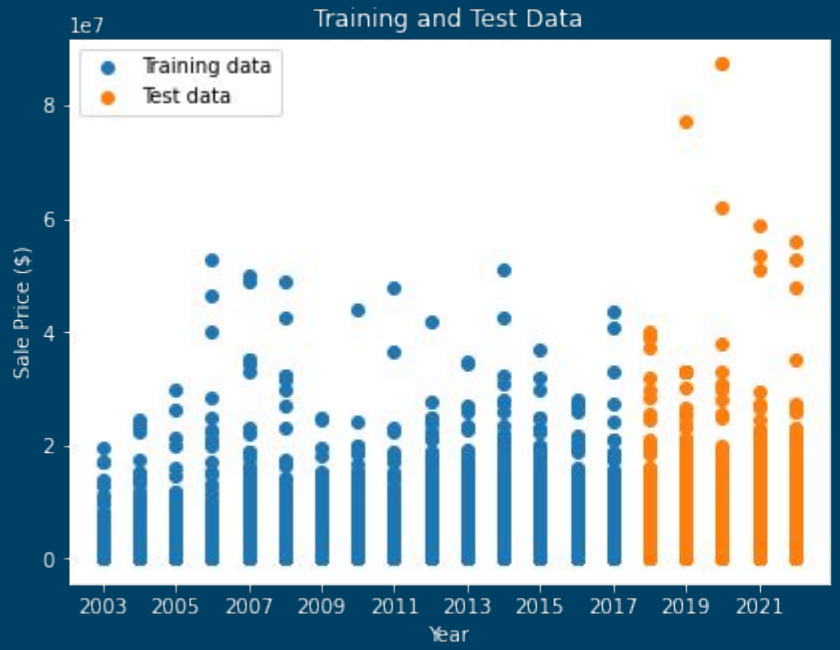
- Goal: use machine learning techniques to predict sale prices
- XGBoost is an open-source software library which uses gradient-boosted decision trees to perform regression and classification
- We used the following factors in our model:
  - Year of sale
  - Land square feet
  - Gross square feet
  - Year built
  - Total units
  - Borough
  - Tax class
  - Neighborhood
  - Building class category

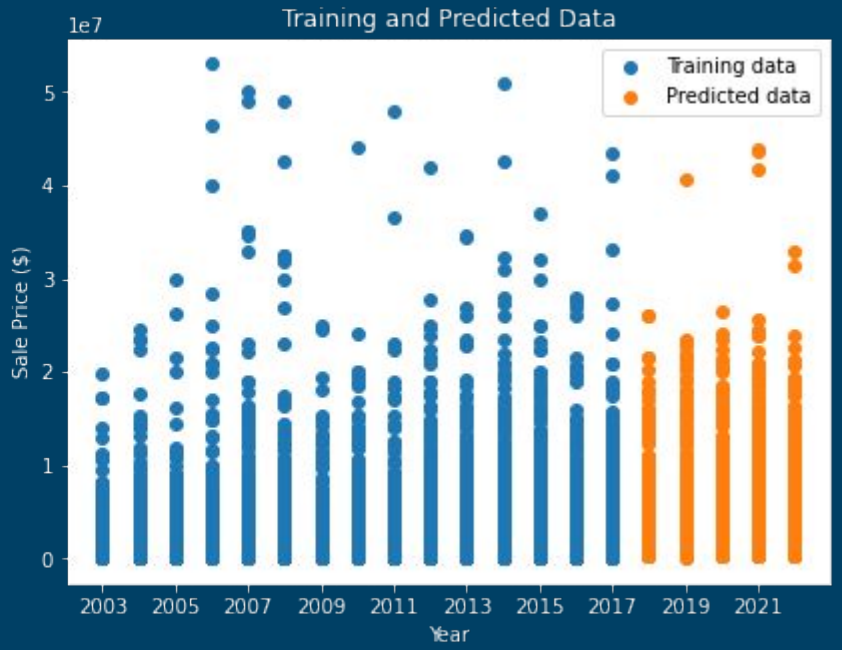
Machine Learning

- We trained our model on data from 2003 to 2017 and used it to attempt to predict sale prices from 2018 to 2022
- Training set accuracy score: **0.80**
- Test set accuracy score: **0.50**
- This is disappointing!
- Let's see if we can find out why...



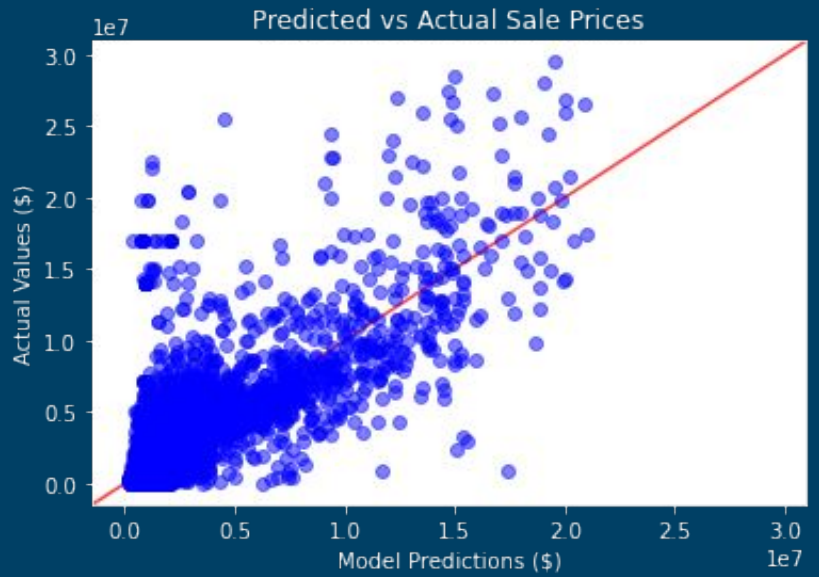
Machine Learning



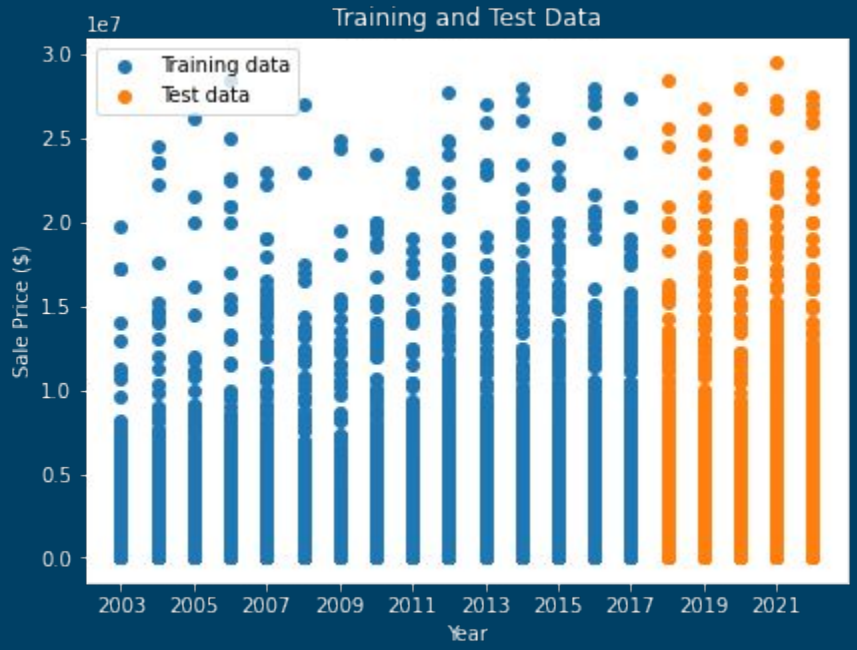


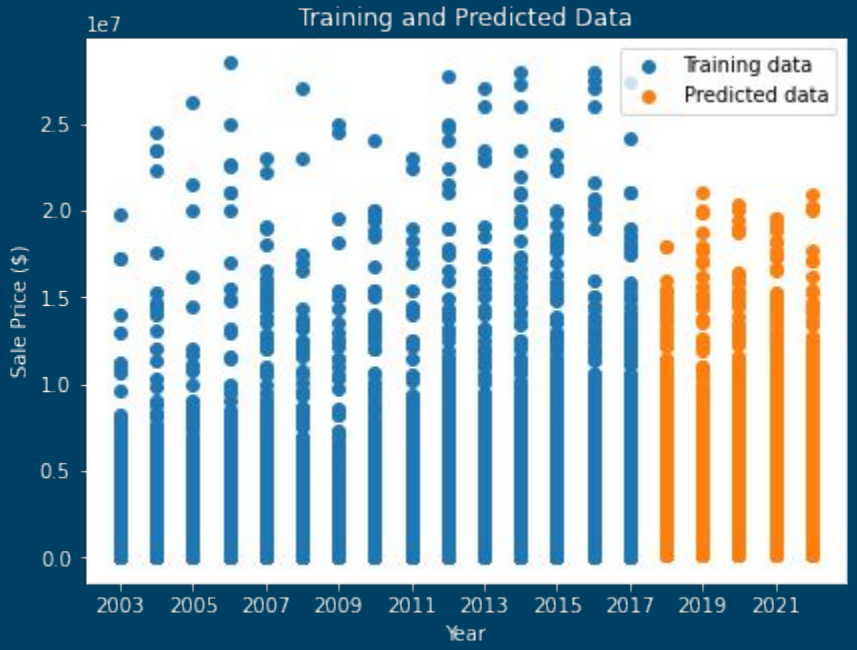
Machine Learning

- Removed sale prices greater than $30,000,000
- New training set accuracy score: **0.78**
- New test set accuracy score: **0.64**
- Not perfect, but much better!



Machine Learning



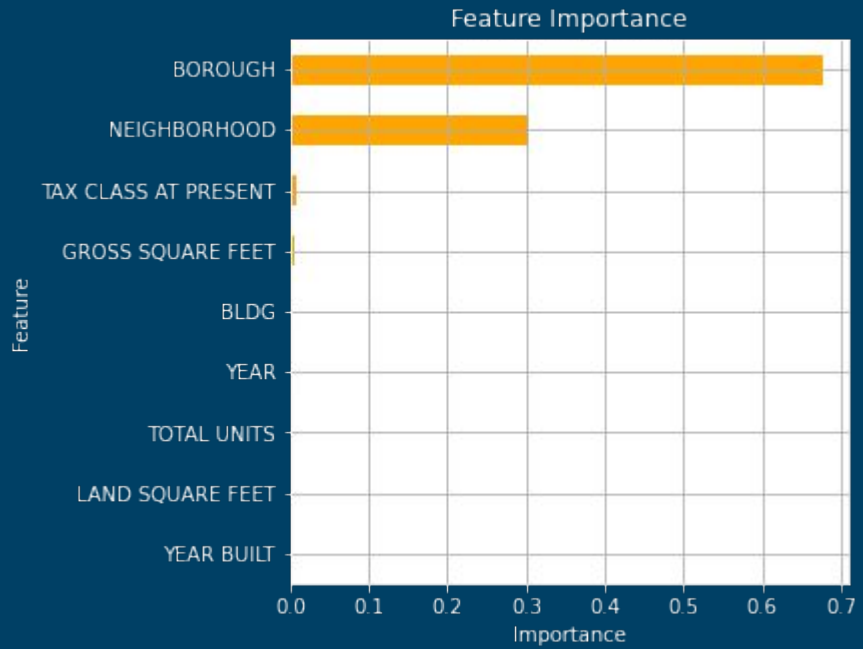


Machine Learning

- Split the test data to see the accuracy by individual year
- 2018 test set accuracy score: **0.69**
- 2019 test set accuracy score: **0.65**
- 2020 test set accuracy score: **0.59**
- 2021 test set accuracy score: **0.64**
- 2022 test set accuracy score: **0.64**
- These results match with our expectations!
    - 2018 is the highest since it directly follows the training set
    - 2020 is the lowest, with COVID-19 the housing market would be harder to predict

Machine Learning

- Feature importance shows the biggest contributing factors
- Borough and neighborhood



Dashboard

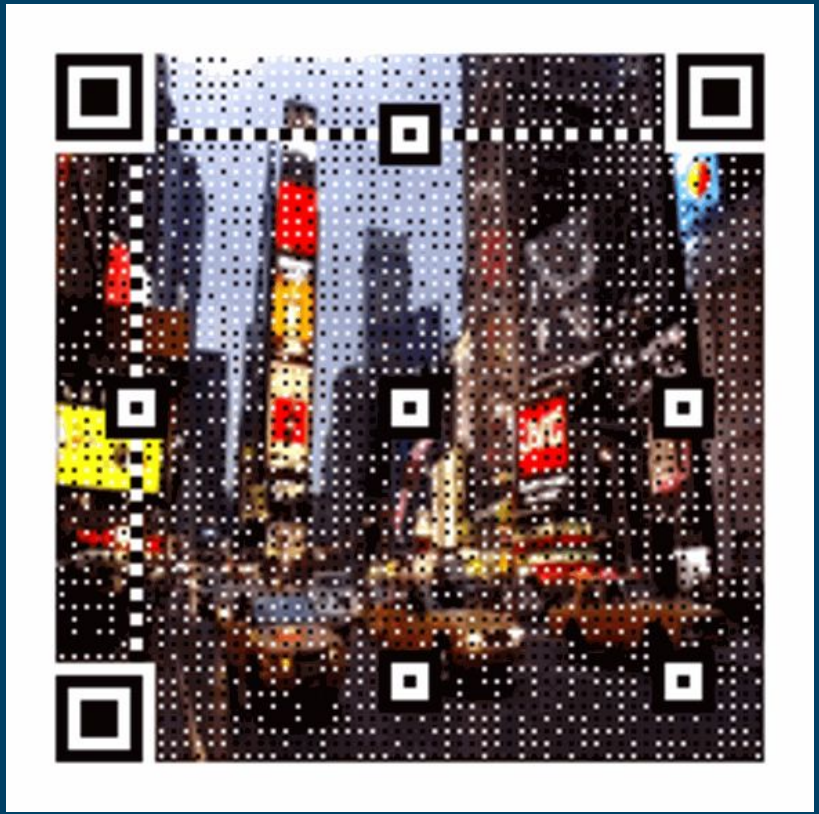
Onto Power BI Online

Conclusions

- Housing Prices have almost tripled across the last 20 years
- Neighborhood is the one of the biggest contributing factors to price
- More people are paying higher rent
- The majority of home value is owned by middle aged individuals

Questions?

Please feel free to provide feedback by scanning the QR Code to the right!



Thank you for listening!