NYC Crime and Real Estate Data Project

Omkar Kshirsagar
Vikhyat Khare
Carter Noordsij
John Sullivan

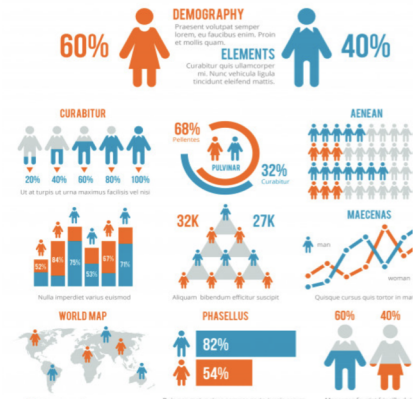
ENGM 182 Data Analytics
6/3/2020

Table of Contents

1. Introduction
2. Data Wrangling
3. Model Building
4. Shiny App UI and Visualization
5. Conclusion

Proposal

For the project, we intend to analyze previous 10 years' data on crimes committed in New York city. This would form the basis of assigning a 'crime score' to each neighborhood. As the second step, we will overlay this analysis with the demographic data for these neighborhoods to develop a model which would help potential house buyers in making purchasing decisions.



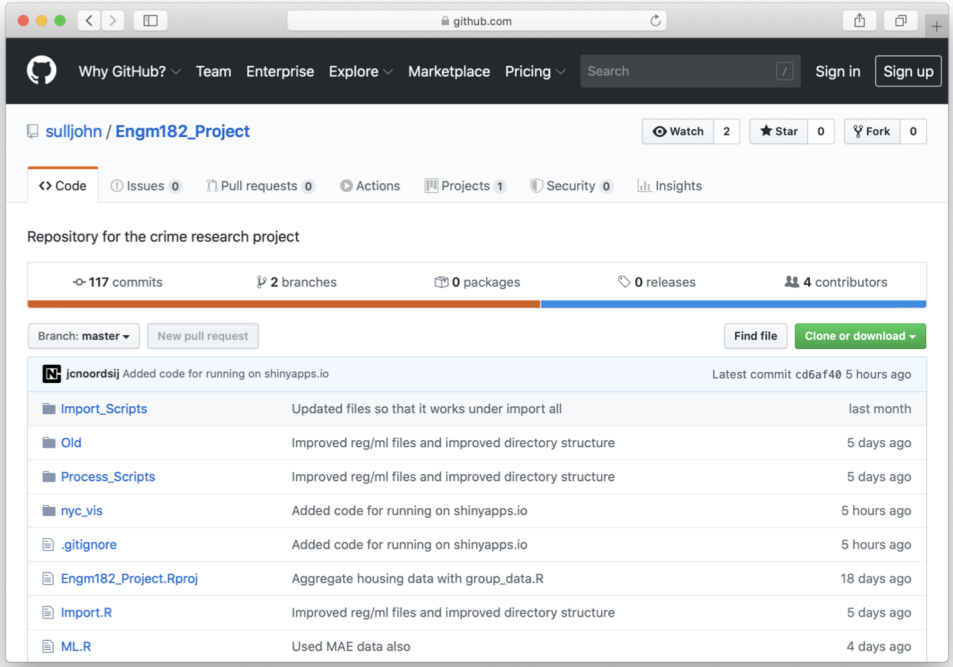
Data Wrangling

1. Technology and Collaboration
2. Data Sources
3. Combining the Data

Technology and Collaboration

1. **GitHub** - hosts our open source code for this project
2. **Google Drive and Dropbox** - for hosting the data
3. **ShinyApps.io** - displays the Shiny app visualization
4. **AWS** - for faster processing of machine learning models

Any user should be able to download the data and run our step-by-step code for the same results



Our GitHub page

Data from NYC OpenData and U.S. Census Bureau

NYC OpenData

NYPD

NYPD Complaints

2006-2017

7.31 million rows

- Crime type
- Date
- Latitude and longitude

NYC Department of Finance

NYC Calendar Sales

2003-2020*

1.42 million rows

- Sale price
- Date
- Square footage
- NYC BBL number

United States Census Bureau

ACS Economic and Demographic Data

2015 estimates for all 2,167 individual NYC census tracts

- Population
- Income per Capita
- Employment breakdown
- Racial breakdown
- Latitude and longitude

Combining the Data

Step 1: Cleaned data on house sales from 2003-2016
Key variables: sale price and date, house size and zip code

Step 2: Converted data of 2015 census from each census tract to zip code and added this to housing data
Key variables: Demography, income, employment, population

Step 3: Created crime scores (explained later) for every year and month post 2000. Linked the yearly data to housing data based on the year of sale and used monthly crime scores for visualization on shiny app

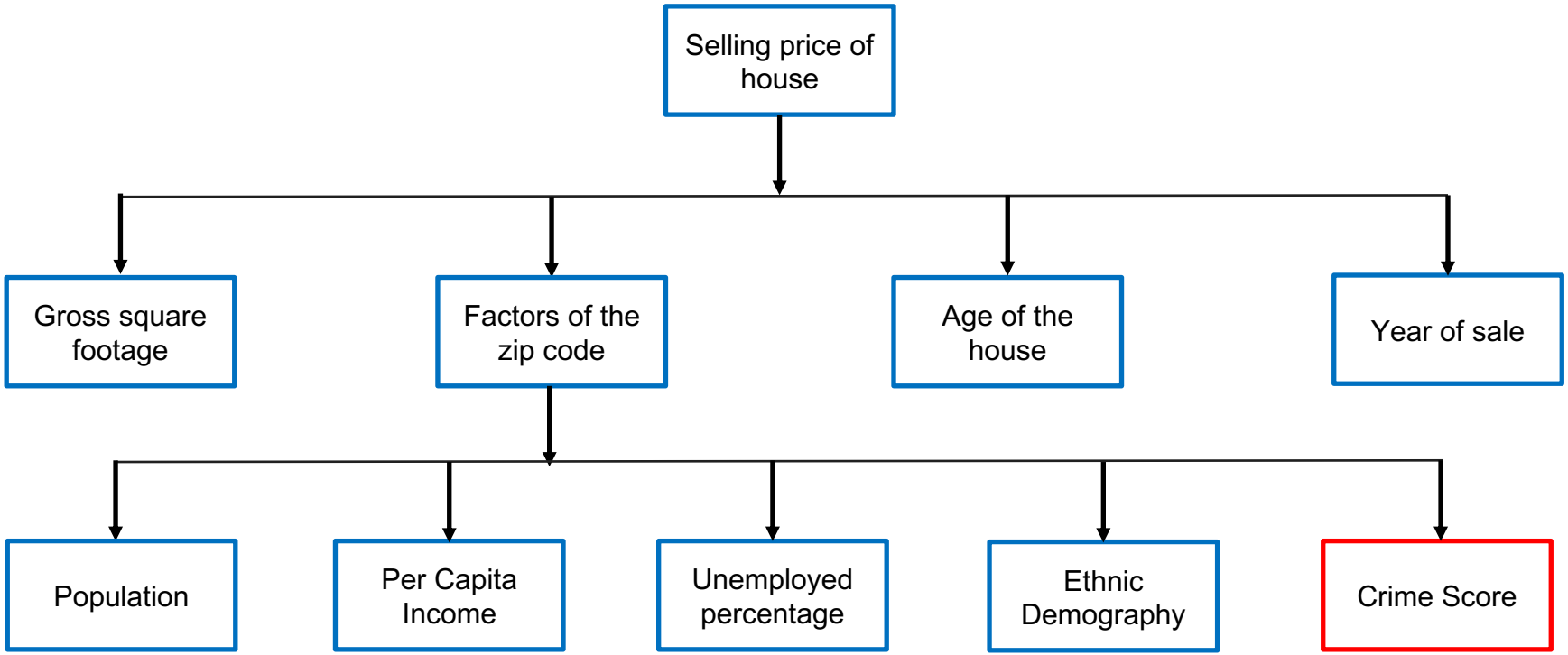
Linked housing data to census data through zip code

Linked housing data to crime data through zip code and date of sale

Model Building

1. Selecting important factors
2. Crime score
3. Regression model
4. Machine learning models
5. Results (best model)

Selecting important factors



Crime score

[During a period x in a region y]

Analyzed the crime data to classify it into different buckets and assigned them a weight

- Weight = 1, Violation: Speeding ticket, public intoxication etc. X Number of crimes
- Weight = 2, Misdemeanor: Simple assault, driving under influence etc. X Number of crimes
- Weight = 3, Felony 4th degree: burglary, larceny etc. X Number of crimes
- Weight = 4, Felony 3rd degree: robbery, arson etc. X Number of crimes
- Weight = 5, Felony 2nd degree: dangerous weapons, sex crimes etc. X Number of crimes
- Weight = 6, Felony 1st degree: murder, rape etc. X Number of crimes

The sum of these gives an idea about the crime in an area for a period. Dividing that by the population gives the "Crime Score" for our analysis

Sources:

https://patrickparrottalaw.com/differences-between-a-violation-a-misdemeanor-and-a-felony/

https://legaldictionary.net/felony/

Regression model

- We ran a regression model to predict the housing prices
- It helped us to establish a baseline, but was not as accurate as we hoped

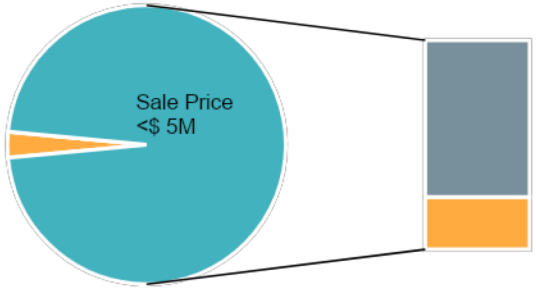
Issues with Regression Model	Solutions
● Large dataset - it took a long time to run the regression	● We used a function called fastLm that removed much of the overhead
● P-values - they go to zero quickly (they all became significant)	● We partitioned the data and observed how accurately its predictions were (MSE)
● Mean-squared error (MSE) - the mean square errors of the predicted values were high, despite the amount of data	● We spoke with Professor Vaze and chose to pursue machine learning models to capture the granularity of our data

Machine learning models

- **Goal** - to predict continuous house prices (not a classification problem)
- **Preprocessing** - before running the models, we performed the following operations:
  1. **Removed N/A**, blank, and rows with improper negative fields
  2. **Separated training data** (2003-2013) and testing (2013-2016)
  3. **Principal component analysis** (PCA) before running the models
  4. **5-fold cross-validation** was used when running the models if possible
- **Caret package (R)** - we used the following regression machine learning models from this package:
  1. Neural Network (one layer)
  2. Deep Neural Network (three layers)
  3. Random Forest
  4. Bagged CART (Bagged Regression Trees)
  5. KNN (K-Nearest Neighbors algorithm)

Results 1

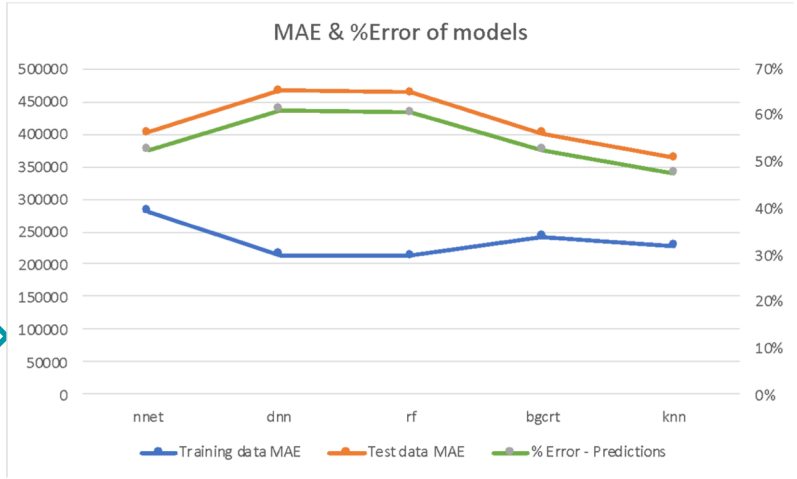
1% of Data



Training data: sale year < 2013

Test data: sale year > 2013

5 models each with 10-fold cross validation

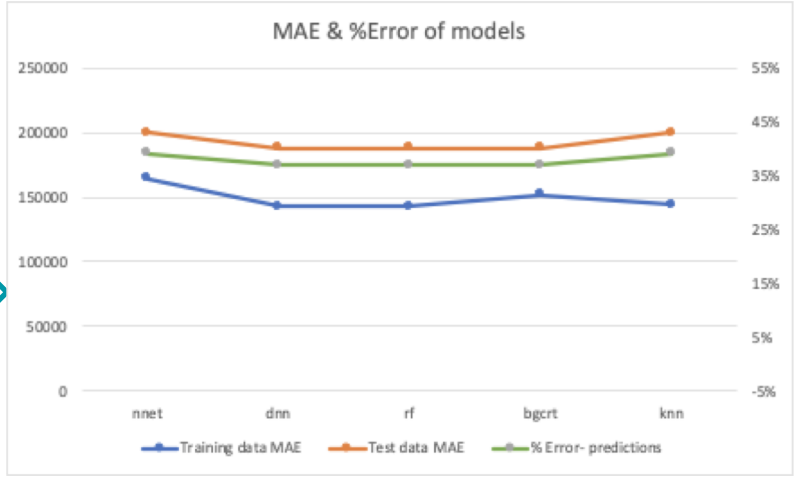


1% of Data

Training data: sale year < 2013

Test data: sale year > 2013

5 models each with 10-fold cross validation



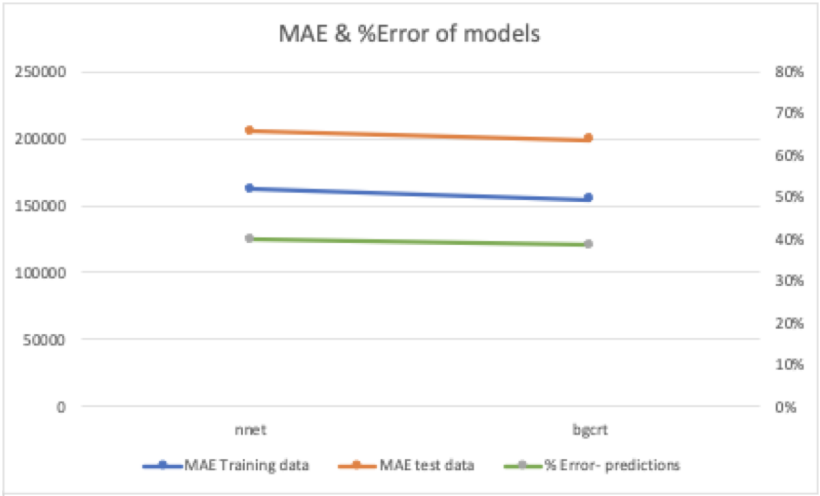
Results 2

100% of Data

Training data:
sale year < 2013

Test data: sale
year > 2013

2 models each with
5-fold cross validation

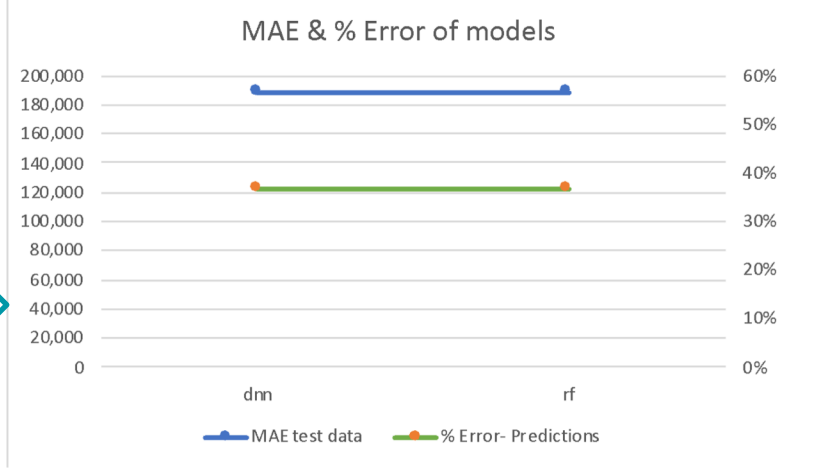


20% of Data

Training data:
sale year < 2013

Test data: sale
year > 2013

2 models each with
no cross validation

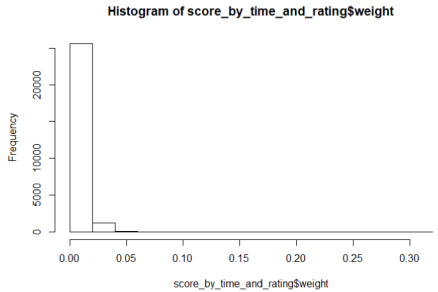


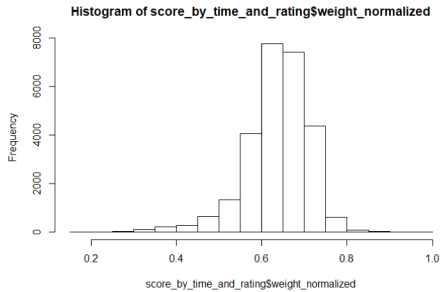
Shiny App UI and Visualization

1. Normalizing Crime Score
2. Predictive Model Integration
3. Merging Data with Polygons
4. App Demo

Filtering and Normalizing Crime Score

- Crime data merged with housing data and census data
- For crime data, removed NAs
  Excluded crime data before 2007
  (As it was reported differently)

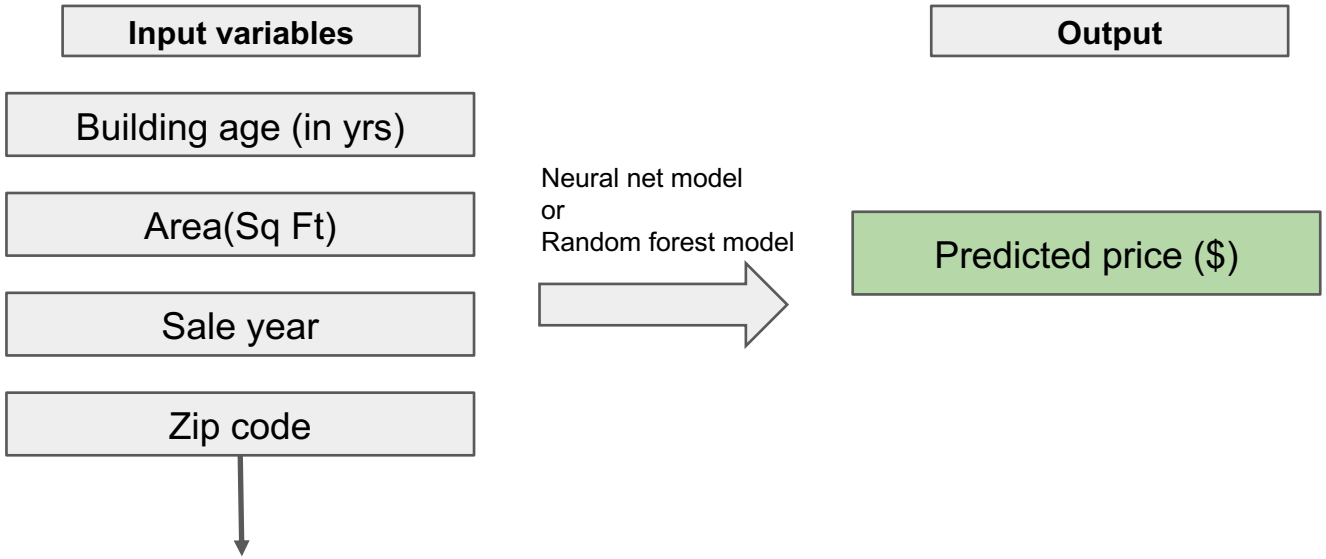


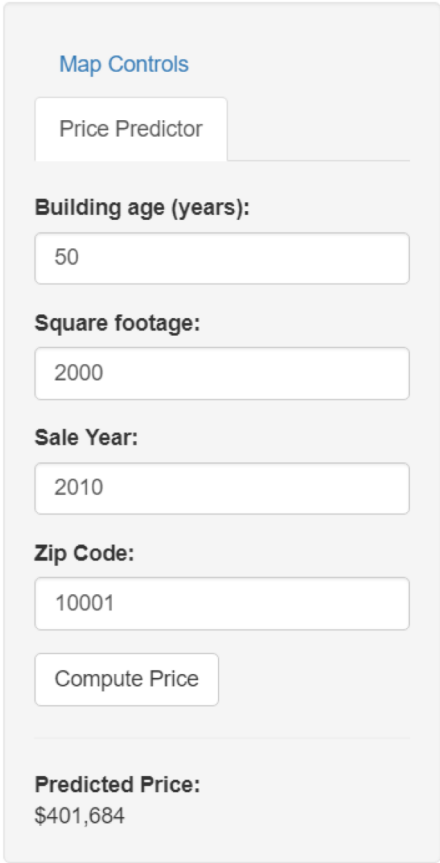


Normalized weight =

[Log((weight)*1000)] ÷ [max(log((weight)*1000))]

Predictive Model Integration



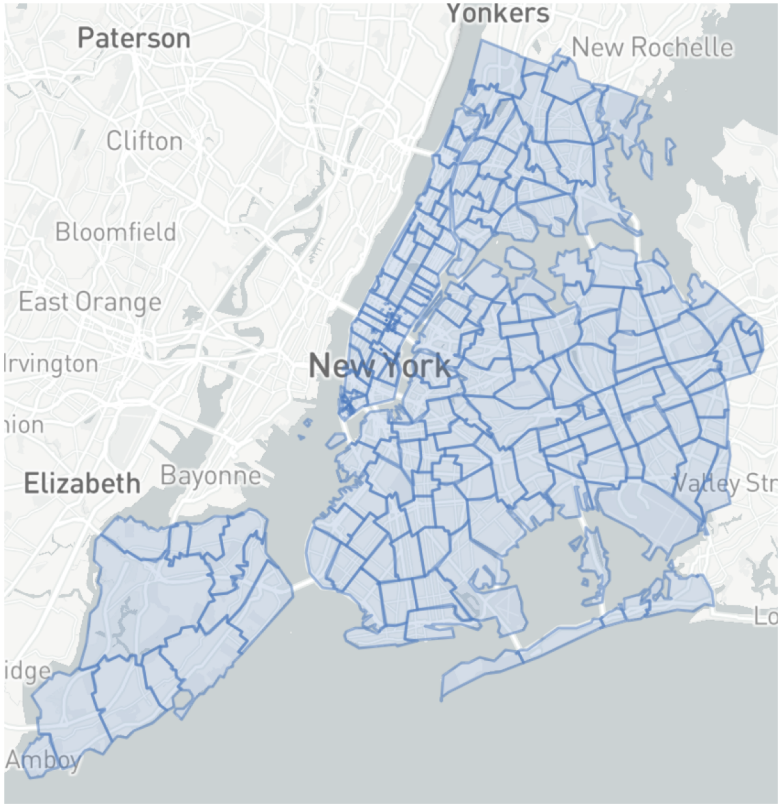


Grouping and merging the data with a shapefile

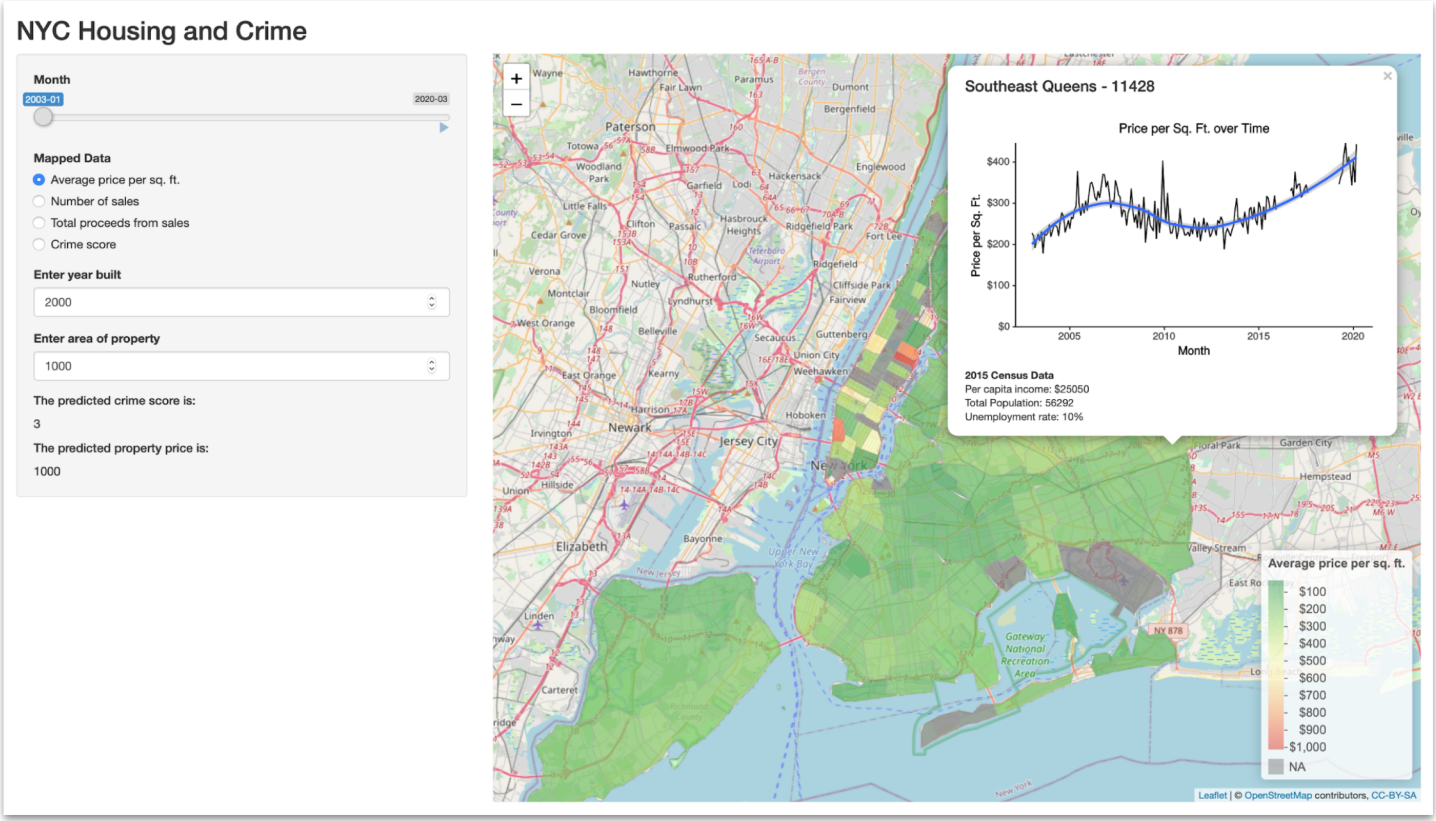
We chose to group the data by zip code and month with hopes of improving app performance and showing meaningful trends in each neighborhood.

This required reverse geocoding the housing data by NYC BBL (borough, block, lot) numbers, and the crime and census data by latitude and longitude coordinates.

NYC Developer Portal
Geoclient API



Shiny App Demo



Conclusion

Challenges

- Many factors go into house prices (difficult to predict)
- Cleaning raw data
- Missing data
- Computing power limitations
- Selection of relevant data

What we learned

- Cleaning the data takes more time than processing
- Need higher computing, or API for reducing computing time
- Limitations of insights due to missing data

Future Work

What could be next...

- Use of alternative spatial visualization methods (heat mapping, hexagonal binning, etc.)
- Examination of seasonal trends
- Further improvement of the housing price model with integration of more data sources (building amenities, last date of renovation, etc.)
- Run additional machine learning models in another language, like Python

Acknowledgements

Thank you for your help!

- Professor Parker
- Professor Vaze
- Holly Buker
- Teaching assistants

Questions

Data Sources and References

Our GitHub repository: https://github.com/sulljohn/Engm182_Project

NYC Crimes 2014-2015: https://www.kaggle.com/adamschroeder/crimes-new-york-city

NYC Crimes 2006-2017: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

NYC Real Estate Data 2003-2015: https://data.cityofnewyork.us/Housing-Development/NYC-Calendar-Sales-Archive-/uzf5-f8n2

NYC Real Estate Data 2016-2017: https://www.kaggle.com/new-york-city/nyc-property-sales

NYC Real Estate Data 2019-2020: https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

NYC Zip Code Shapefile: https://jsspina.carto.com/tables/nyc_zip_code_tabulation_areas_polygons/public/map

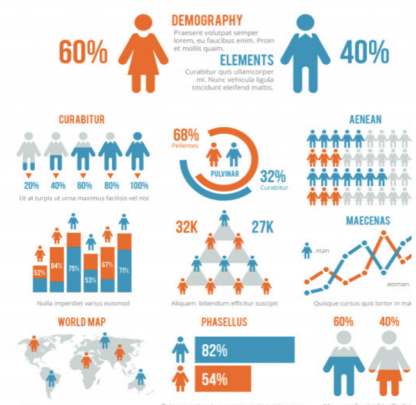
NYC Geoclient API: https://developer.cityofnewyork.us/api/geoclient-api

Question mark image: https://www.psychologytoday.com/nz/blog/how-do-life/201407/asking-the-right-questions

Appendix

Problem Statement

Crime information is not currently used as a factor in helping predict housing prices or displayed visually with housing data for demographic information for the boroughs of New York City.



Regression Results

- Using the independent variables discussed previously (we had large dataset issues):
  - **P-values** - many were significant (approached 0)
  - **Adjusted R-squared value (R^2)** = 27.16%
- We tested the models ability to make predictions:
  - **Root Mean Squared (RMS)** = 197,718.8
  - **Root Mean Squared Error (RMSE)** = 43.40%

```
> # RMS =
> rms = sqrt(mean(results$diff_squared))
> rms
[1] 197718.8
> # RMS Error =
> rms_error = rms/mean(results$y_test)
> rms_error
[1] 0.4339573
```

```
> fit1 <- fastLm(x_train, y_train)
> summary(fit1)

Call:
fastLm.default(X = x_train, y = y_train)

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-950020.0  -99223.0    2296.7  110620.0 1064100.0

                              Estimate      StdErr     t.value   p.value
land_square_feet            1.3001e-02  7.1805e-03  1.8105e+00 0.0702140 .
PerCapitaIncome             1.5723e+00  5.5897e-02  2.8128e+01 < 2.2e-16 ***
Unemployed                 -1.2609e+04  1.6815e+02 -7.4989e+01 < 2.2e-16 ***
TotalPop.x                  1.0917e+10  2.2631e-01  4.8242e+10 < 2.2e-16 ***
Men                        -1.0917e+10  1.7174e-01 -6.3571e+10 < 2.2e-16 ***
Women                      -1.0917e+10  2.8541e-01 -3.8252e+10 < 2.2e-16 ***
                                    •
                                    •
                                    •
year_built1839             -2.2311e-01  0.0000e+00        -Inf < 2.2e-16 ***
year_built1840              3.5509e-01  0.0000e+00         Inf < 2.2e-16 ***
 [ reached getOption("max.print") -- omitted 155 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.964e+05 on 349122 degrees of freedom
Multiple R-squared: 0.2723,     Adjusted R-squared: 0.2716
```

P-values Going to 0