Intro

Brief Explanation of Select Topic
-How does crime and demographic data affect housing prices in Illinois?

Our goal with these data sets was to try and find any correlation of factors that might affect housing value. Using Zillow data we are going off of Zillow Home Value Index (ZHVI) as our way of assigning value to properties.

Factors accounted for:

- Cities in Illinois, not neighborhoods for large cities like Chicago
- Demographic information
- Crime Rates of the Cities

Potential Questions:

- What factors most highly impact ZHVI?
- How much do these uncontrollable factors influence ZHVI?

Sources of Data

- Zillow
- FBI: Uniform Crime Reporting
- United States Census
- OpenDataSoft

Data Exploration

- Our idea for data had gone through several phases:
    - Housing price/demographic data by zip code
    - After some struggles with merging data, decided to add another variable with crime data
    - Zip code was difficult to make work with several sources, so information was merged on city

Data Analysis

- Analysis is determined through Linear Regression model with Machine Learning
- Graphs and maps created in Tableau to help visualize specific cities relative to each other
- Overall Trends determined through ML model, several biases that may not be explained
    - Certain populations are associated with higher housing prices. Is this *due* to the higher amount of that racial population, or is it just a correlation with another variable that is not as heavily weighted (median income vs. crime rates vs. low population)

Sources of Data

- Zillow
  - Market dataset broken down by city and zip code
- US Census
  - Primarily used for demographic and income information
- Department of Justice
  - FBI Uniform Crime Reporting gives state-wide reported crimes by city
- OpenDataSoft
  - Offers latitudinal and longitudinal data for cities to better help with visualizations

Data Exploration

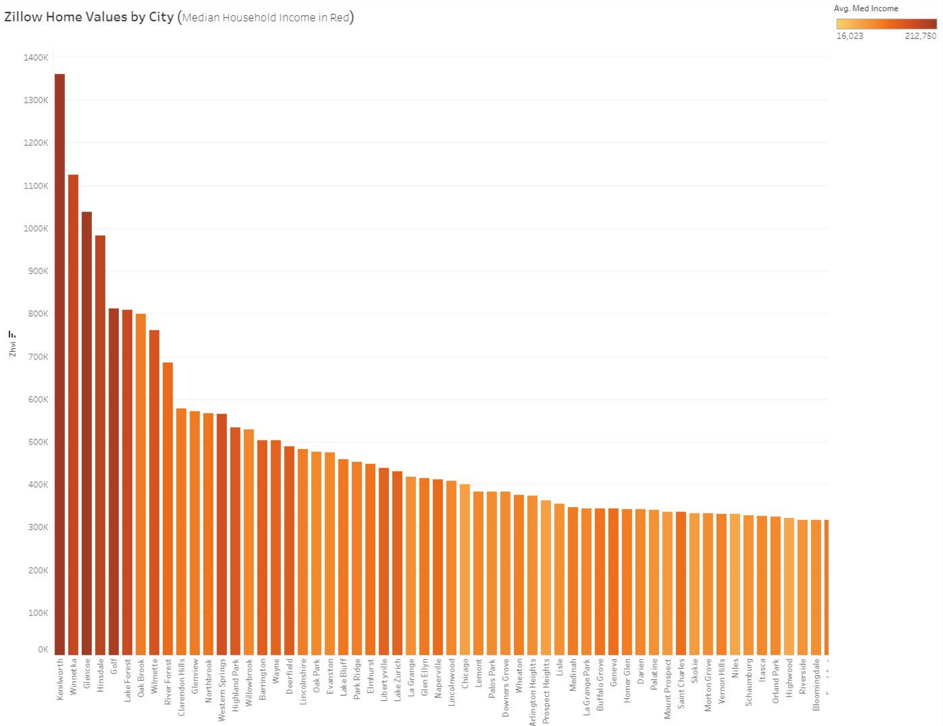
- Initially hoped to be able to determine pricing variations based on demographic data by zip code
  - Determining these factors by city proved to be easier and more effective for the ML model and helping have a better visualization
- Refined tables by dropping unnecessary columns
  - Already aggregated columns, leading to “double dipping” in data when applying to a ML model

Data Analysis

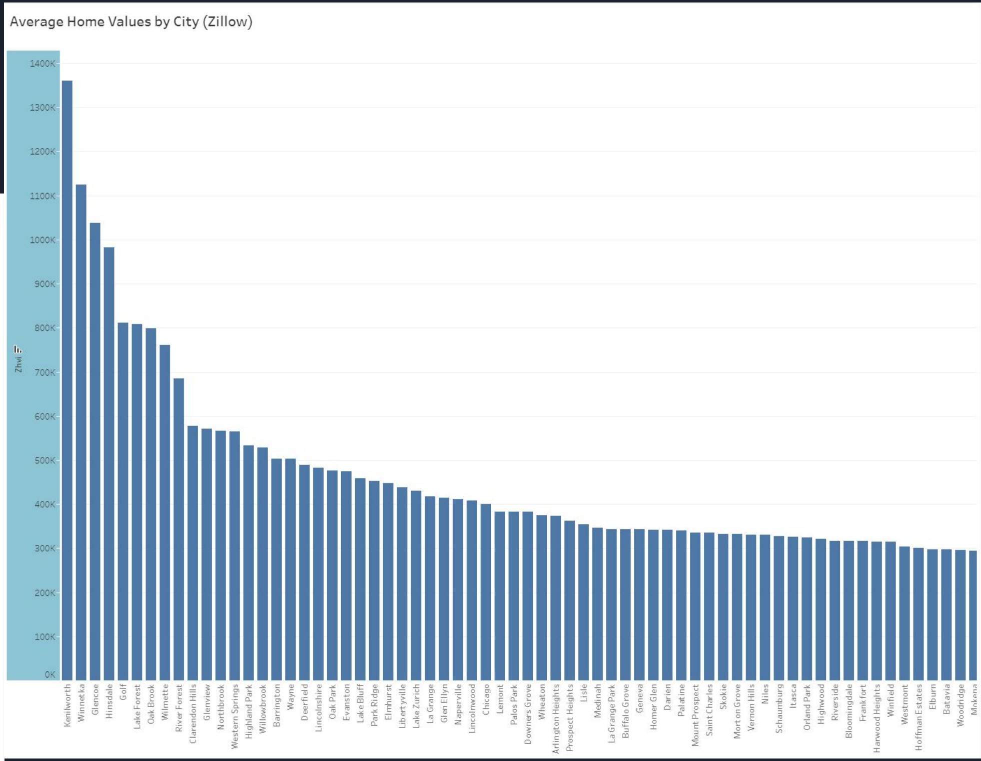
- KNeighborsRegressor for training Linear Regression model
    - Gave highest R-squared value compared to other algorithms
    - Does not require training steps, does not explicitly build any model
- Downsides
    - Sensitive to outliers
    - Works best with small number of input variables

Main Stats

>Import from Tableau,
Just a screenshot in slides presenation



Zillow Home Values by City (Median Household Income in Red)



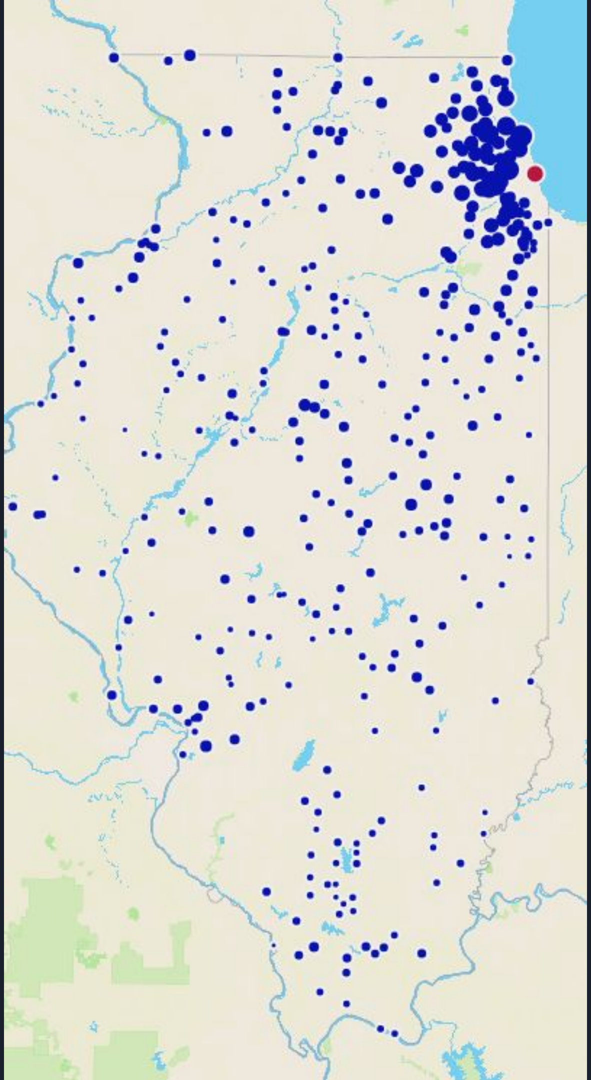
Average Home Values by City (Zillow)

Heatmap

Interactive map that can be scrolled in with, geoJSON

Interactive elements

- Moveable map
- Applicable filters for types of crime
- Mouseover for more specific information



Technologies

- Dashboard will be created through github pages, using HTML5/js to help show and filter information.
- Machine Learning
  - PySpark taking information from tables in PostgreSQL
  - Supervised learning with scikit-learn library
- Tableau
  - Import using iFrame from tableau (embedding within website)
    <iframe src=”public.tableau link” width=”X” height=”X”>
    </iframe>
- geoJSON heatmap for interactive graph