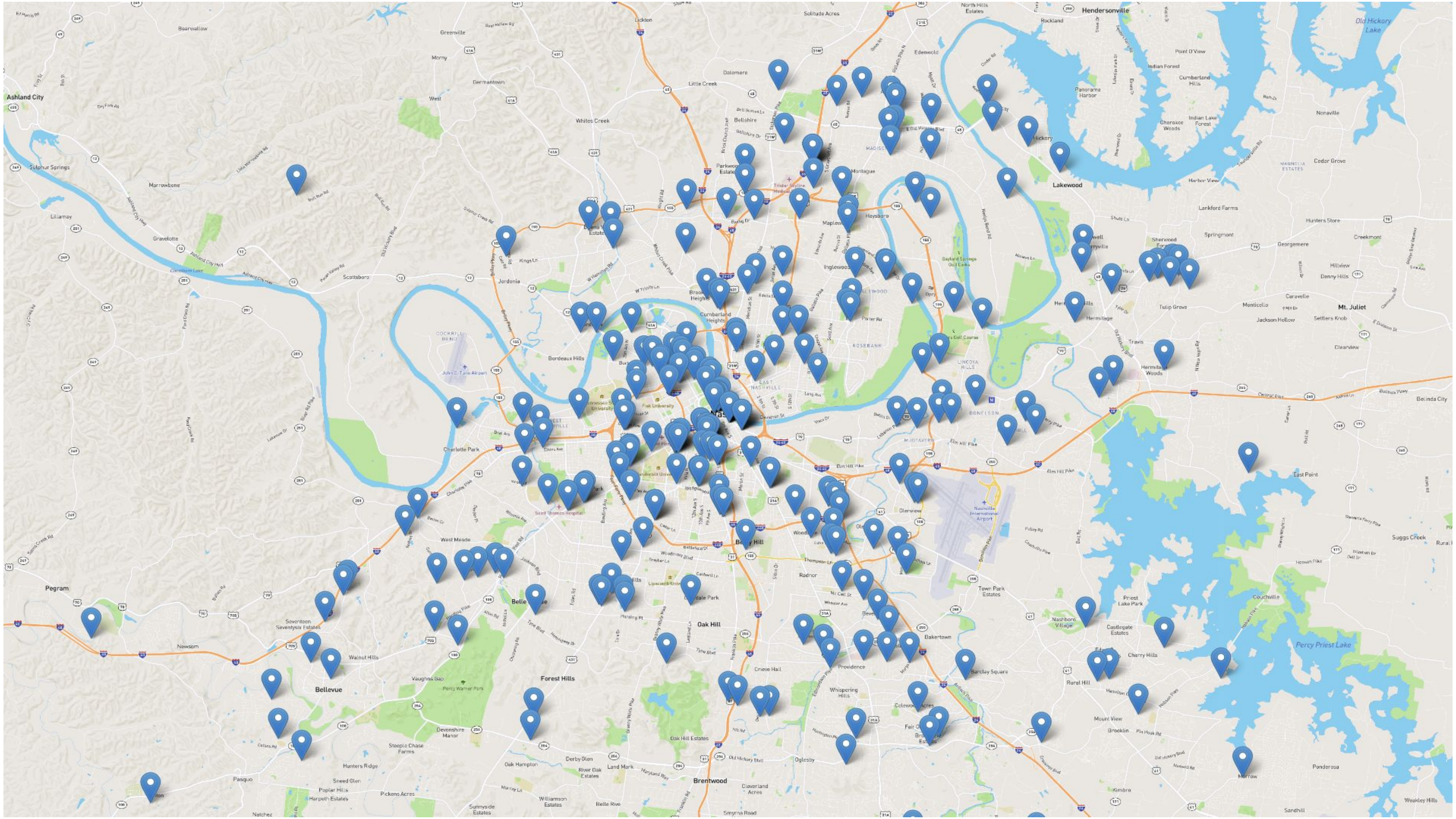
# Machine Learning and Real Estate

Helping buyers find home pricing deals using Machine Learning.

# Nashville, ranks in the top 10 of fastest growing metros.

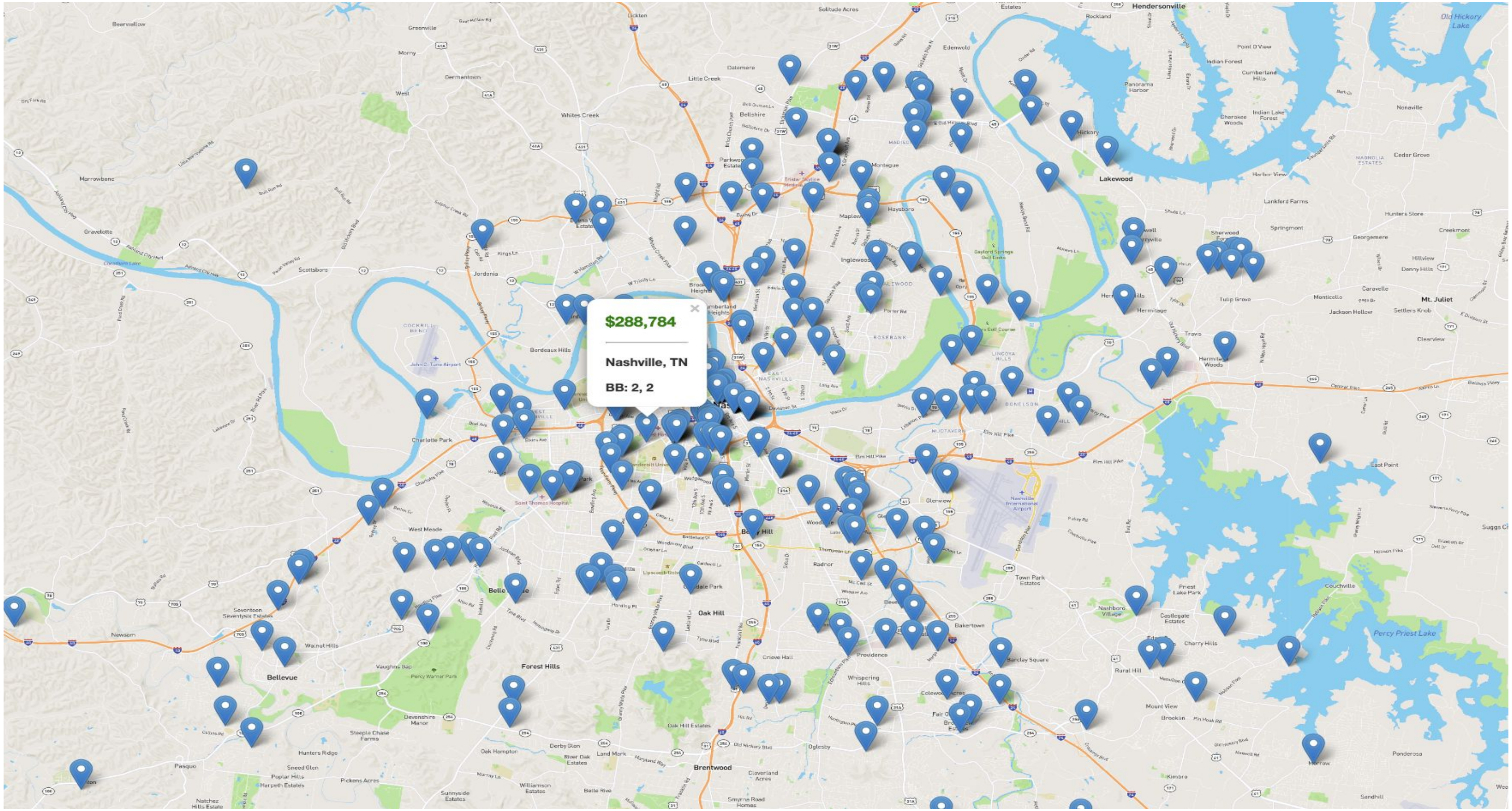
Nashville, TN and and the surrounding boroughs are predicted to be one of the top five housing markets in the country again this year. Historic low and near zero Federal Reserve rates have fueled far more buyers than sellers. With an imbalance of buyers to sellers, home prices can exceed the appraised value of the property. Additionally, many buyers exacerbate the problem by over bidding the value of the home feeding the continuation of over priced homes and non accommodating sellers.

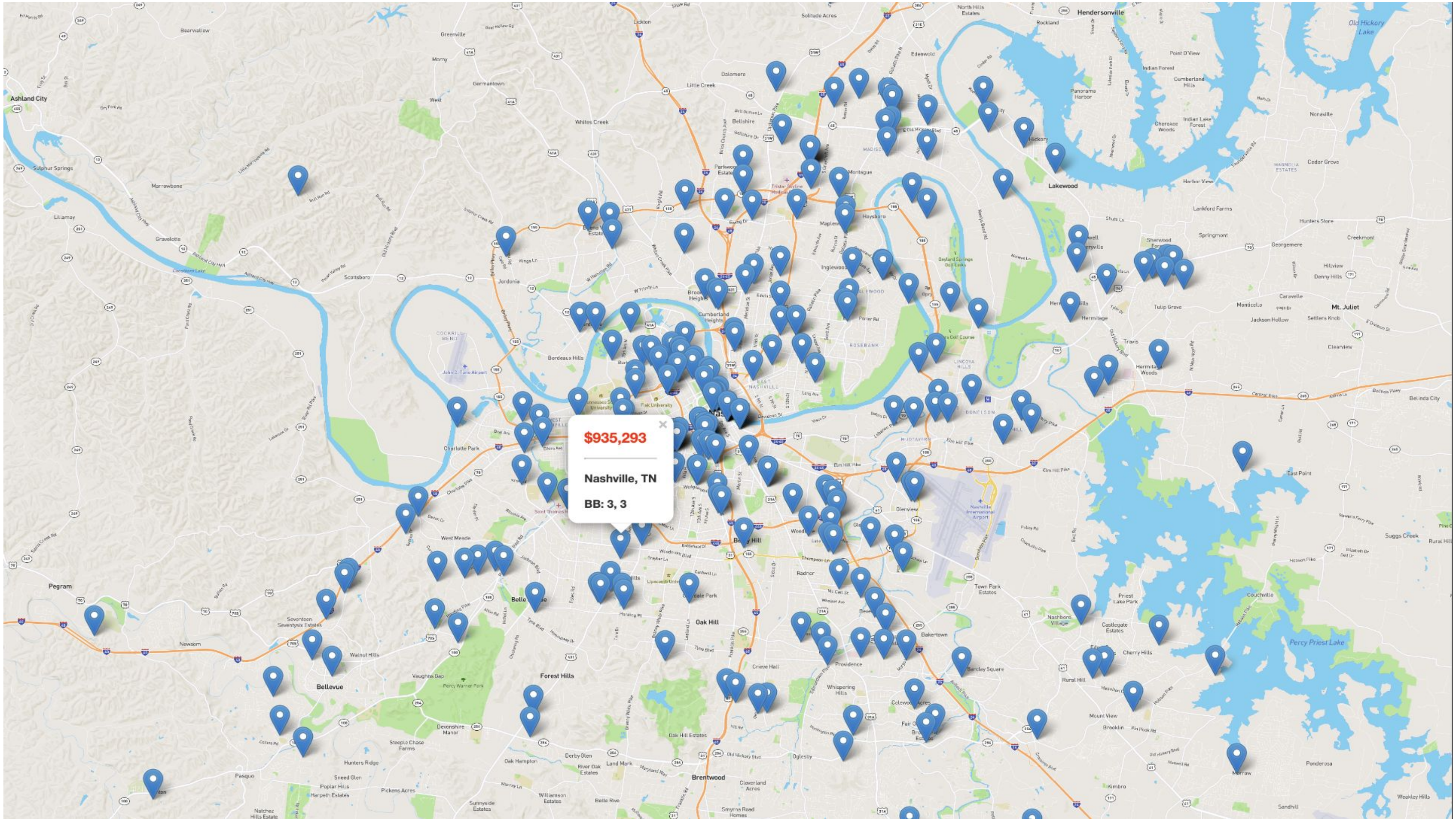
The Grey team will use Machine Learning to examine homes for sale and compare the current asking price, along with other predefined features, to previous sold homes to determine if the listing price is fair, above, or below market value.


Ashland City
Bearwallow
Greenville
Germantown
Whites Creek
Lickton
Little Creek
Dolomere
Bellshire
Parkwood Estates
Montague
Haysboro
Edenwold
Rockland
Hendersonville
North Hills Estates
Old Hickory Lake
Point O'View
Indian Forest
Cumberland Hills
Panorama Harbor
Cherokee Woods
Indian Lake Forest
Nonaville
Cedar Grove
Harbor View
Lakewood
Lankford Farms
Hunters Store
Springmont
Georgemere
Creekmont
Sherwood Forest
Hillview
Denny Hills
Caraville
Monticello
Jackson Hollow
Settlers Knob
Mt. Juliet
Tulip Grove
Hermitage
Travis
Hermitage Woods
Clearview
Belinda City
East Point
Suggs Creek
Couchville
Percy Priest Lake
Priest Lake Park
Castlegate Estates
Cherry Hills
Nashboro Village
Rural Hill
Mount View
Brooklin
Kimbro
Ponderosa
Sandhill
Weakley Hills
Marrowbone
Sulphur Springs
Lillamay
Gravelotte
Scottsboro
Jordonia
Cockrill Bend
John C. Tune Airport
Bordeaux Hills
Charlotte Park
Inglewood
Maplewood
Madison
Rosebank
Lincoya Hills
Donelson
Midtavern
Nashville International Airport
Glenview
Town Park Estates
Woodbine
Radnor
Berry Hill
Belmont
Fisk University
Tennessee State University
Vanderbilt University
Saint Thomas Hospital
West Meade
Belle Meade
Pegram
Newsom
Seventeen Seventysix Estates
Walnut Hills
Bellevue
Vaughns Gap
Percy Warner Park
Forest Hills
Oak Hill
Crieve Hall
Providence
Whispering Hills
Bakertown
Barclay Square
Coleman Acres
Fair Oaks
Oak Hill Estates
Devonshire Manor
Steeple Chase Farms
Hunters Ridge
Sneed Glen
Poplar Hills
Harpeth Estates
Pasquo
Natchez
Pickens Acres
Oak Hampton
Derby Glen
River Oak Estates
Land Mark
Williamson Estates
Sunnyside Estates
Belle Rive
Brentwood
Oglesby
Cloverland Acres
Smyrna Road

# The Problem

Many buyers are finding that the market is moving faster than they can process whether the home's asking price is a good deal (fair market value) or a bad deal (above market value).  With Machine Learning, the Grey Team will process the current listings of homes for sale and develop a map of homes that are defined as good or bad value. Home buyers can click on a marker on the map and the color code of the price will alert the buyer if the home is above market value, by showing the home price in Red or fair value listing the home price in Green. Using the map provided by the Machine Learning algorithm, prospective home buyers can concentrate their efforts and resources on homes in the fair market value to make competitive bids.


$288,784
Nashville, TN
BB: 2, 2
Hendersonville
Old Hickory Lake
Lakewood
Hickory
Madison
Parkwood Estates
Bellshire
Whites Creek
Little Creek
Dalemere
Edenwold
Rockland
Goodlettsville
Solitude Acres
Lickton
Greenville
Morny
Germantown
West
Bearwallow
Marrowbone
Gravelotte
Scottsboro
Jordonia
Bordeaux Hills
Cockrill Bend
John C. Tune Airport
Charlotte Park
West Meade
Belle Meade
Bellevue
Newsom
Seventeen Seventysix Estates
Walnut Hills
Vaughns Gap
Percy Warner Park
Devonshire Manor
Steeple Chase Farms
Hunters Ridge
Pasquo
Sneed Glen
Poplar Hills
Harpeth Estates
Pickens Acres
Natchez Hills Estate
Forest Hills
Oak Hill
Oak Hampton
Derby Glen
River Oak Estates
Land Mark
Williamson Estates
Belle Rive
Sunnyside Estates
Brentwood
Cloverland Acres
Smyrna Road Homes
Oglesby
Crieve Hall
Whispering Hills
Providence
Radnor
Berry Hill
Woodbine
Glenview
Inglewood
Maplewood
Haysboro
Rosebank
East Nashville
Cumberland Heights
Brooklyn Heights
Nashville International Airport
Donelson
Lincoya Hills
Hermitage
Hermitage Woods
Travis
Tulip Grove
Sherwood Forest
Springmont
Lankford Farms
Georgemere
Hunters Store
Creekmont
Cedar Grove
Nonaville
Harbor View
Mt. Juliet
Monticello
Jackson Hollow
Settlers Knob
Clearview
Caravelle
Hillview
Danny Hills
Percy Priest Lake
Couchville
Priest Lake Park
Castlegate Estates
Nashboro Village
Charry Hills
Rural Hill
Mount View
Brooklin
Kimbro
Ponderosa
Sandhill
Morrow
East Point
Suggs Cr
Bakertown
Barclay Square
Town Park Estates
Colewood Acres
Fair Oaks
Saint Thomas Hospital
Vanderbilt University
Lipscomb University
Panorama Harbor
Indian Forest
Cumberland Hills
Cherokee Woods
Indian Lake Forest
Point O'View
Magnolia Estates
Gaylord Springs Golf Links



$935,293

Nashville, TN

BB: 3, 3

REDFIN

# Finding the best dataset

Two data sets were needed to create our predictive model on housing price. One including recently sold houses in the Nashville metropolitan area (train and test data) and one for houses to be sold.

Finding a working API to pull sold housing data proved harder to find than initially thought. Realtors try their best to keep this data private, in an attempt to keep themselves relevant for buyers.

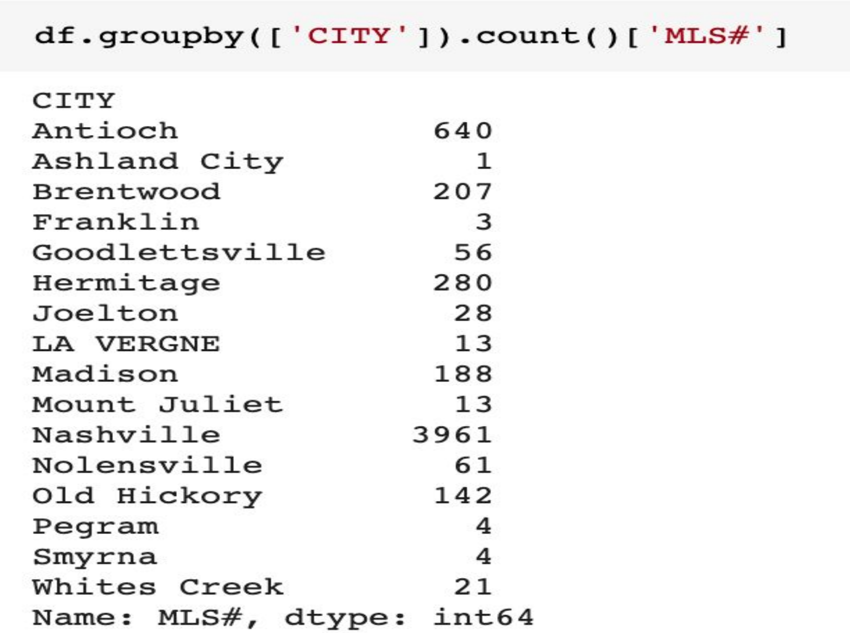
Without being able to find a working API, we manually exported data from the Redfin Realty website.

# Cleaning the Data

- The two datasets were combined and EDA was completed.

-Unnecessary data and columns with a large number of nulls were dropped.

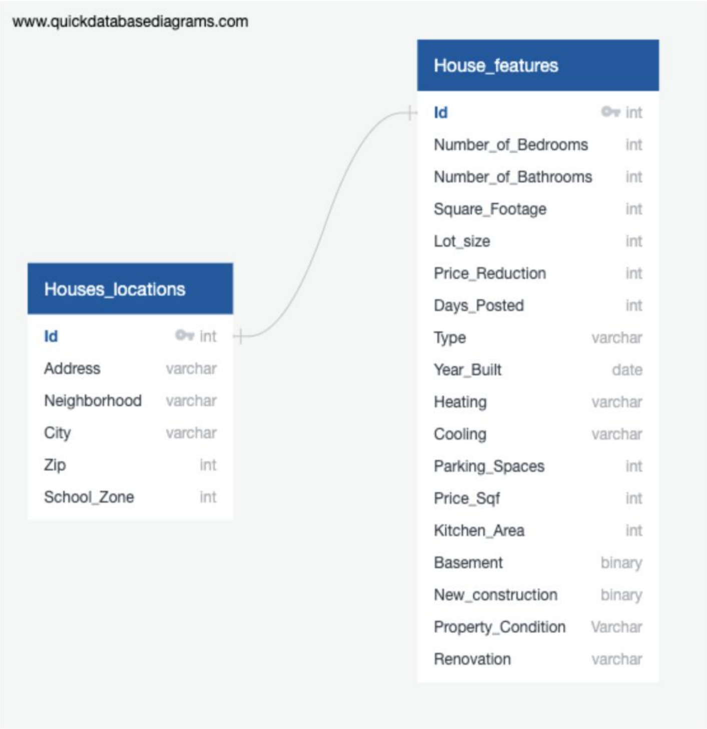
- Data was pushed to the The RDS Database using PostgresSQL.



```
df.groupby(['CITY']).count()['MLS#']
```

```
CITY
Antioch              640
Ashland City           1
Brentwood            207
Franklin               3
Goodlettsville        56
Hermitage            280
Joelton               28
LA VERGNE             13
Madison              188
Mount Juliet          13
Nashville           3961
Nolensville           61
Old Hickory          142
Pegram                 4
Smyrna                 4
Whites Creek          21
Name: MLS#, dtype: int64
```

# ERD



www.quickdatabasediagrams.com

| Houses_locations | |
|---|---|
| Id | int |
| Address | varchar |
| Neighborhood | varchar |
| City | varchar |
| Zip | int |
| School_Zone | int |

| House_features | |
|---|---|
| Id | int |
| Number_of_Bedrooms | int |
| Number_of_Bathrooms | int |
| Square_Footage | int |
| Lot_size | int |
| Price_Reduction | int |
| Days_Posted | int |
| Type | varchar |
| Year_Built | date |
| Heating | varchar |
| Cooling | varchar |
| Parking_Spaces | int |
| Price_Sqf | int |
| Kitchen_Area | int |
| Basement | binary |
| New_construction | binary |
| Property_Condition | Varchar |
| Renovation | varchar |

# ETL & EDA Methods Used

- Request function - used this to send a request to Redfin.com and it returned  a response with our specific parameters (city, home size, etc.)
- Json_normalize - used to flatten our json into a pandas DataFrame
- For Loops - used to append the responses
- Df.describe - used for summary statistics and distribution

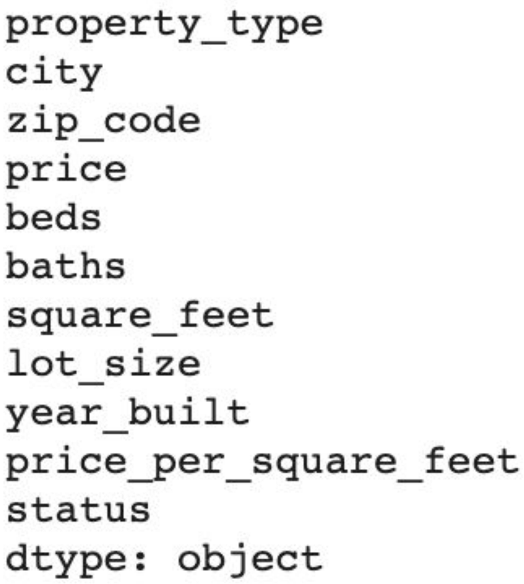
# EDA & Database Challenges

- Working around data point limit of 200
- Importing dataframe into PostgresSQL database
- Wanted to include neighborhoods to be more granular but we had over 2,000 different neighborhoods listed in the data set for under 6,000 rows

# Preparing data For Machine Learning

- One Hot Encoder and Min Max Scaler were used to scale and encode the features to prep them for the machine learning code

- SKlearn’s train, test split was used to split our Sold Housing Data into our training and testing



```
property_type
city
zip_code
price
beds
baths
square_feet
lot_size
year_built
price_per_square_feet
status
dtype: object
```

# Machine Learning Testing

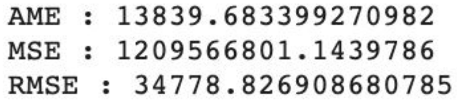
- Multiple models were tested (Multiple Linear Regression, Xg Boost Regression, and Neural Network). Absolute Mean Error (AME), Mean Squared Error(MSE) and Root Mean Squared Error(RMSE) were used to measure loss and predicted accuracy on the test data.
- Neural network performed the best in all metrics and was selected. Different variations of features were used to find the best measures.
- Hyper parameter tuning helped select kernel initializer and activation function using Sklearn’s GridSearchCV.
- Code was run against multiple different numbers of epochs to confirm the best model accuracy.

# Loss and accuracy against testing data

- Across all models, a high AME, MSE, and RMSE was recorded for both the predicted value and the loss.

- However, all of these values are a measure of how far off the regression is from the data points, which means it is completely relative to the dataset itself

- The values being predicted ranges from the hundreds of thousands to millions.  Thus the high error is expected when put in perspective of our dataset



```
AME : 13839.683399270982
MSE : 1209566801.1439786
RMSE : 34778.826908680785
```

# Predicting Housing Value

- The For Sale data was run through the machine learning model to predict the prices of each house

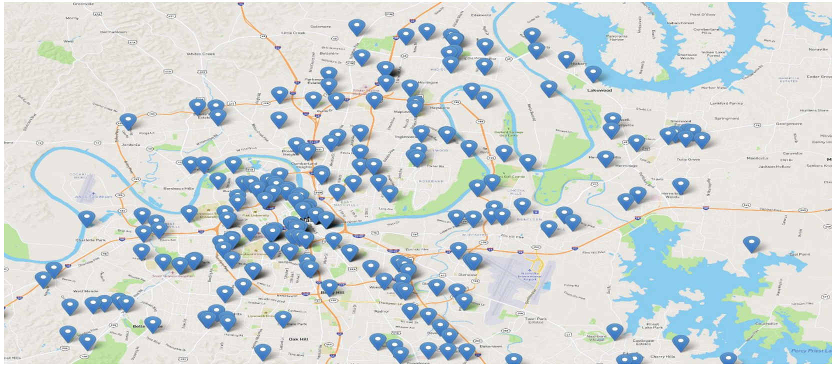
- We compared our predicted pricing against the the list price on the house. From this, we indicated whether the house was overvalued or undervalued and marked each house as such.

- The predicted housing data was then loaded back into the RDS Database

# HTML and Leaflet

**HTML**: was used to create and structure our vision using Leaflet

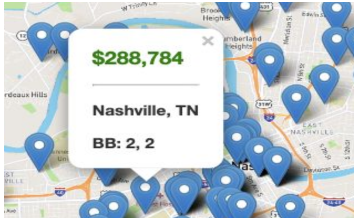
- We used a url within our HTML to pull in our map and stylesheet from Leaflet
- Path to our data using D3 to read our Json (more from Matt)
- Path to our PopUp to maximize visualization.
-



# CSS

**What is CSS**? **CSS** is the language for describing the presentation of Web pages, including colors, layout, and fonts: Css makes our App presentable.

- We used CSS to create the colors, Fonts. (Good Value or Bad value)
- Layout of our app.



# Hosting the Data

- A flask app is used to query the RDS Database that houses our machine learning data. It is then routed to our javascript file and connected to our formatted.html file through flask as well

- Heroku is used to host the the flask app and present our fully functioning dashboard

Link to our app: https://mtjb-real-estate-final.herokuapp.com/

# URL to Working App

https://mtjb-real-estate-final.herokuapp.com/