**Purpose**

The purpose of this project is to conduct an analysis on Real Estate trends on the Dallas real estate market and find correlations between housing prices and a variety of housing characteristic variables from Redfin’s open-source data.

**Context**

First time buyers and families looking to purchase homes often are overwhelmed by the countless variables that go into buying a home. Companies like Trulia, Zillow, Realtor.com, Estately, Homes .com, and Redfin seek to make available homes easier to find by future homeowners, but can easily be overwhelmed by the 50+ variables on Zillow or 75+ variables on Redfin. Since the first thing you see when looking at a home is the listed price, the square footage of the home, and the number of days the home was listed, we wanted to examine if these variables, even matter when considering a home offer price. We also look into Home Owner Associations (HOA), because there are varying viewpoints on the pros/cons of buying a home with HOA fees.

**Data**

We downloaded data from RedFin specifically in the Dallas region, but we had to clean the data up to get a final data set. Please refer to the project\_2\_data\_collection document for the full data collection process. We also examined homes sold within the past three months.

Data Source: <https://www.redfin.com/city/30794/TX/Dallas/filter/max-price=10M>

Data Collection: Redfin only allows you to download up to 350 listings at a time. As of 30 March 2019, there were 4679 homes listed on Redfin.com for the city of Dallas. Therefore, we downloaded the data in increments up to 350 and compile it together to create one master file.

Process to acquire master data set of Dallas Real Estate Market from Redfin Data:

1. https://www.redfin.com/city/30794/TX/Dallas/
2. Select a filter from 0-75k
3. Download file at bottom of the listings.
4. Repeat steps 2 – 3 for ranges: 75k-150k, 150-200k, 200-250k, 250-300k, 300-350k, 350-375k, 375-400k, 400-450k, 450-500k, 500-550k, 550-600k, 600-700k, 700-900k, 900k-1.25m, 1.25-2m, 2-10m
5. Combine all downloaded files into one master file
6. Delete all duplicate listings set at 75k, 150k, 200k, 250k, 300k, 350k, 375k, 400k, 450k, 500k, 550k, 600k, 700k, 900k, 1.25m, 2m. Duplicates are caused by filter at these thresholds.
7. Final result was 4,679 homes.

The original Data set consisted of 18 downloads from Redfin as shown below by,

csv: number of homes in the data set, price range of homes drawn from redfin

redfin\_2019-03-30-11-50-48.csv: 134 homes, 0k – 75k

redfin\_2019-03-30-11-50-20.csv: 292 homes, 75k – 150k

redfin\_2019-03-30-11-49-22.csv: 324 homes, 150k – 200k

redfin\_2019-03-30-11-48-52.csv: 318 homes, 200k – 250k

redfin\_2019-03-30-11-48-17.csv: 335 homes, 250k – 300k

redfin\_2019-03-30-11-47-41.csv: 344 homes, 300k – 350k

redfin\_2019-03-30-11-46-43.csv: 188 homes, 350k – 375k

redfin\_2019-03-30-11-46-18.csv: 210 homes, 375k – 400k

redfin\_2019-03-30-11-45-32.csv: 325 homes, 400k – 450k

redfin\_2019-03-30-11-44-43.csv: 315 homes, 450k – 500k

redfin\_2019-03-30-11-43-49.csv: 249 homes, 500k – 550k

redfin\_2019-03-30-11-43-10.csv: 221 homes, 550k – 600k

redfin\_2019-03-30-11-42-51.csv: 290 homes, 600k – 700k

redfin\_2019-03-30-11-40-59.csv: 336 homes, 700k – 900k

redfin\_2019-03-30-11-38-49.csv: 315 homes, 900k – 1.25M

redfin\_2019-03-30-11-35-09.csv: 341 homes, 1.25M – 2M

redfin\_2019-03-30-11-34-48.csv: 346 homes, 2M – 10M

redfin\_2019-03-30-12-16-03.csv: 15 homes, 10M – no Max

From there, we have our final data set labeled dallas\_available\_real\_estate.csv.

We then delete several columns from the data set that are not relevant to what we are looking at (SOLD DATE, STATE, NEXT OPEN HOUSE START TIME, NEXT OPEN HOUSE END TIME, URL, MLS#, FAVORITE, INTERESTED, LATITUDE, LONGITUDE). We then change all the variables to lower case and underscores in spaces.

Our final codebook after this process is:

|  |  |
| --- | --- |
| Variable | Description |
| sale\_type | how the home is being sold |
| property\_type | type of property (ie. Multi-family, single-family, vacant land) |
| address | property address |
| city | City |
| zip\_or\_postal\_code | zip code |
| price | listed price |
| beds | number of beds |
| baths | number of baths |
| location | city district of the home |
| square\_feet | total square feet of living space |
| lot\_size | total size of the property in square feet |
| year\_built | what year the home was built |
| days\_on\_market | number of days on the market |
| price\_per\_square\_feet | price variable divided by square\_feet variable |
| hoa\_per\_month | amount of monthly HOA fees due |
| status | property’s status as active or not |
| source | where the information for the listing was found |

We conducted the same process with the listed data with the following sold data with about 2,682 homes:

redfin\_2019-04-09-10-14-17.csv: 263 homes, 0k – 100k

redfin\_2019-04-09-10-13-47.csv: 315 homes, 100k – 150k

redfin\_2019-04-09-10-13-24.csv: 266 homes, 150k – 175k

redfin\_2019-04-09-10-12-59.csv: 251 homes, 175 – 200k

redfin\_2019-04-09-10-12-34.csv: 173 homes, 200k – 225k

redfin\_2019-04-09-10-11-49.csv: 278 homes, 225k – 275k

redfin\_2019-04-09-10-10-46.csv: 259 homes, 275k – 325k

redfin\_2019-04-09-10-10-13.csv: 350 homes, 325k – 400k

redfin\_2019-04-09-10-09-44.csv: 285 homes, 400k – 500k

redfin\_2019-04-09-10-09-07.csv: 308 homes, 500k – 700k

redfin\_2019-04-09-10-08-27.csv: 346 homes, 700k – no Max

For the sold data, we also removed all sold homes from March 31 to April 9 so that our sold data would represent the same time slice as the listed data on March 30, 2019. This sold data was compiled and saved as ‘dallas\_real\_estate\_sold\_filtered.csv’.