Austin, TX - House Listings

DATA ANALYSIS OF THE HOUSING MARKET

Prepared by: Nina Sysoeva

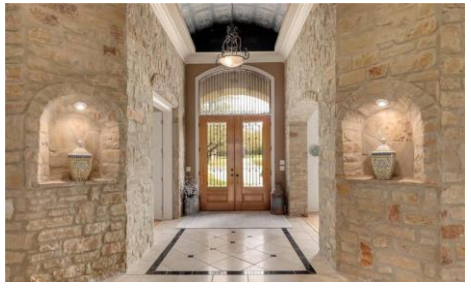




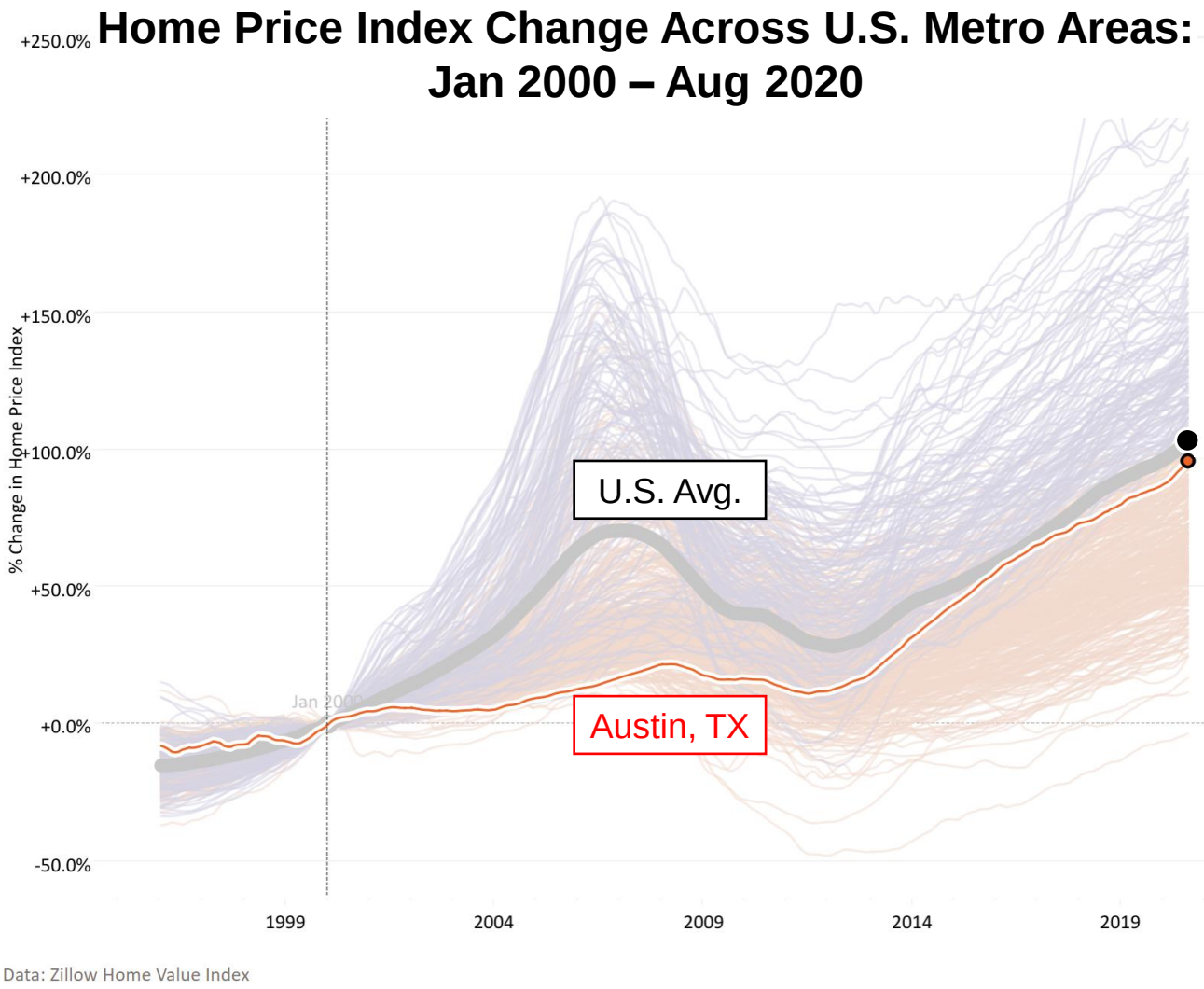









Austin 20-Year Home Price Index Growth Is Comparable to the U.S. Average



	Austin, TX	U.S. Avg.
Difference Jan 2000 – Aug 2020	**+96%**	**+103%**
Home Price Index (Aug 2020)	**$357k**	**$257k**

Our dataset was scraped in Jan 2021

https://www.visualcapitalist.com/20-years-of-home-price-changes-in-every-u-s-city/

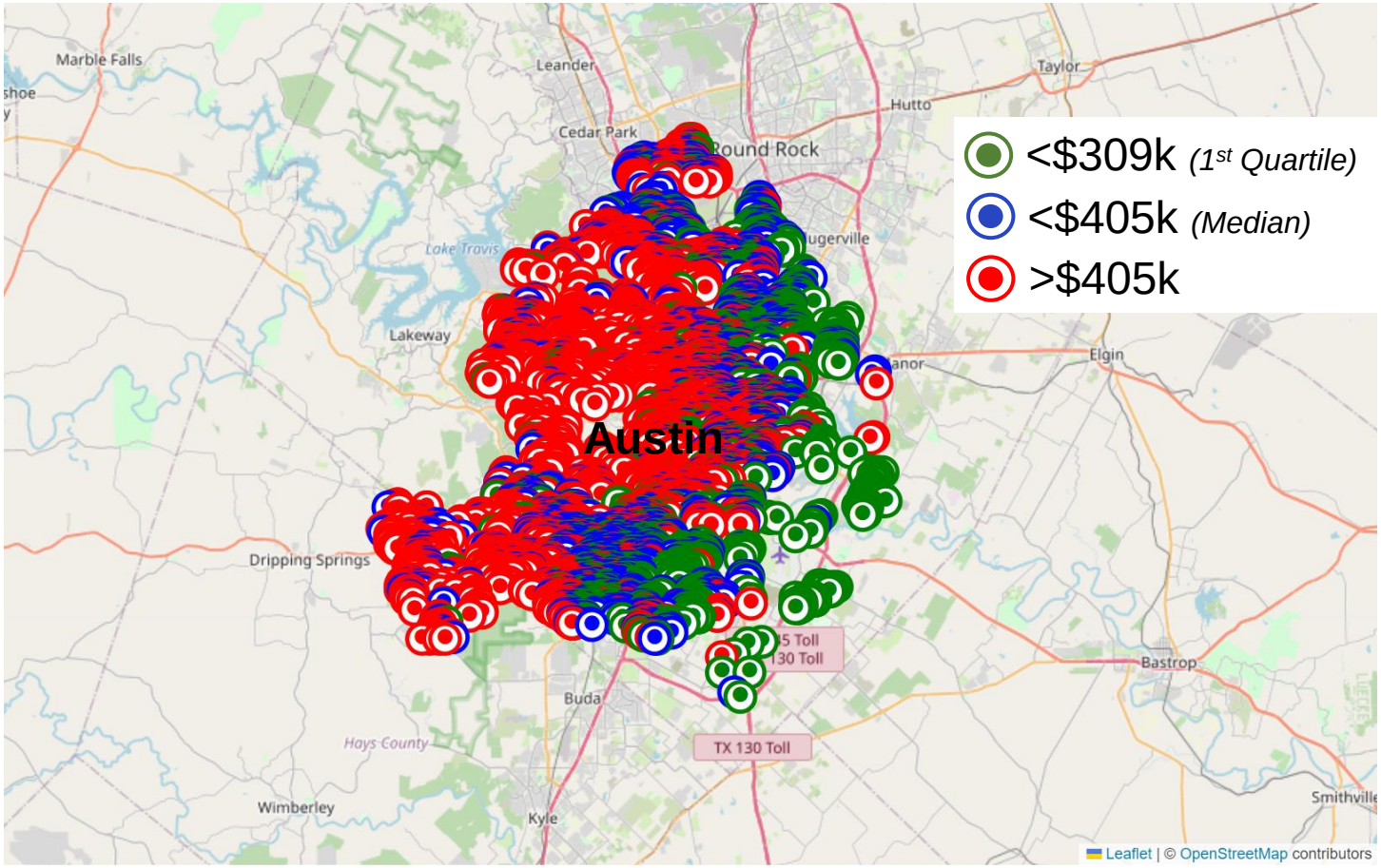
Housing Properties Clustered Around Austin Based on Their Price Bracket

12 MB

47 Features

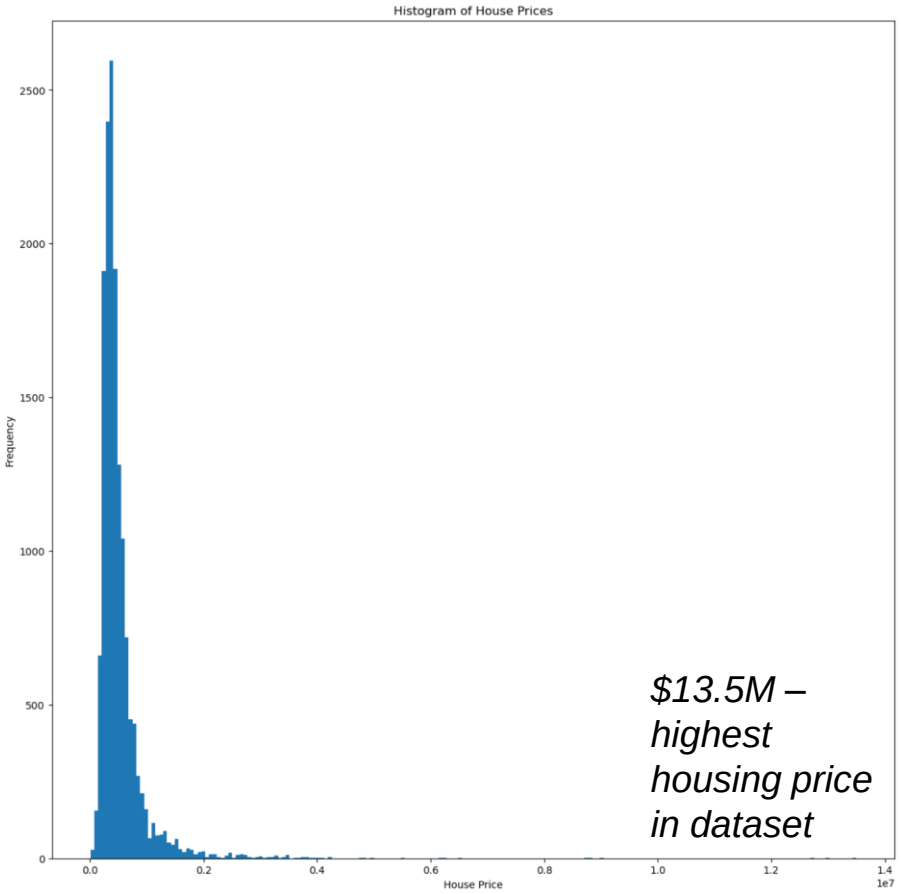
~15,000 Rows

Based on Zillow Listings

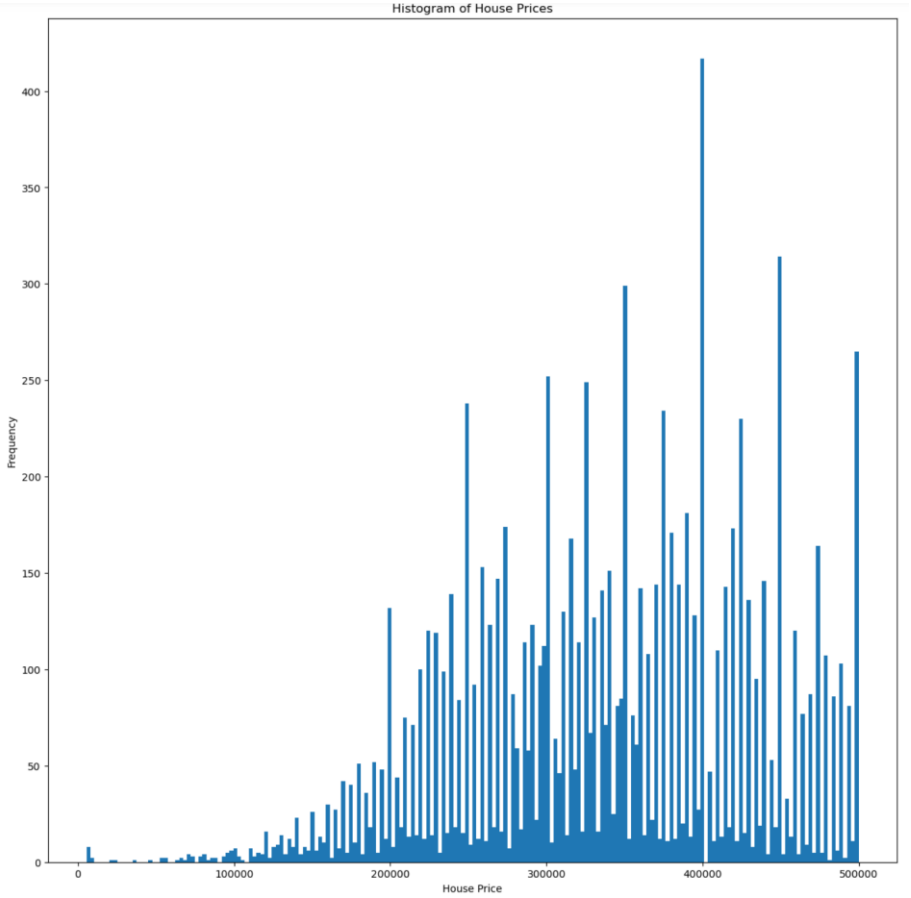


Targeting to Predict Housing Price column ("latestPrice") – dependent variable

★ Removing Outliers Moved Linear Regression Train & Test Scores Closer Together and Adjusted Magnitudes of the Coefficients

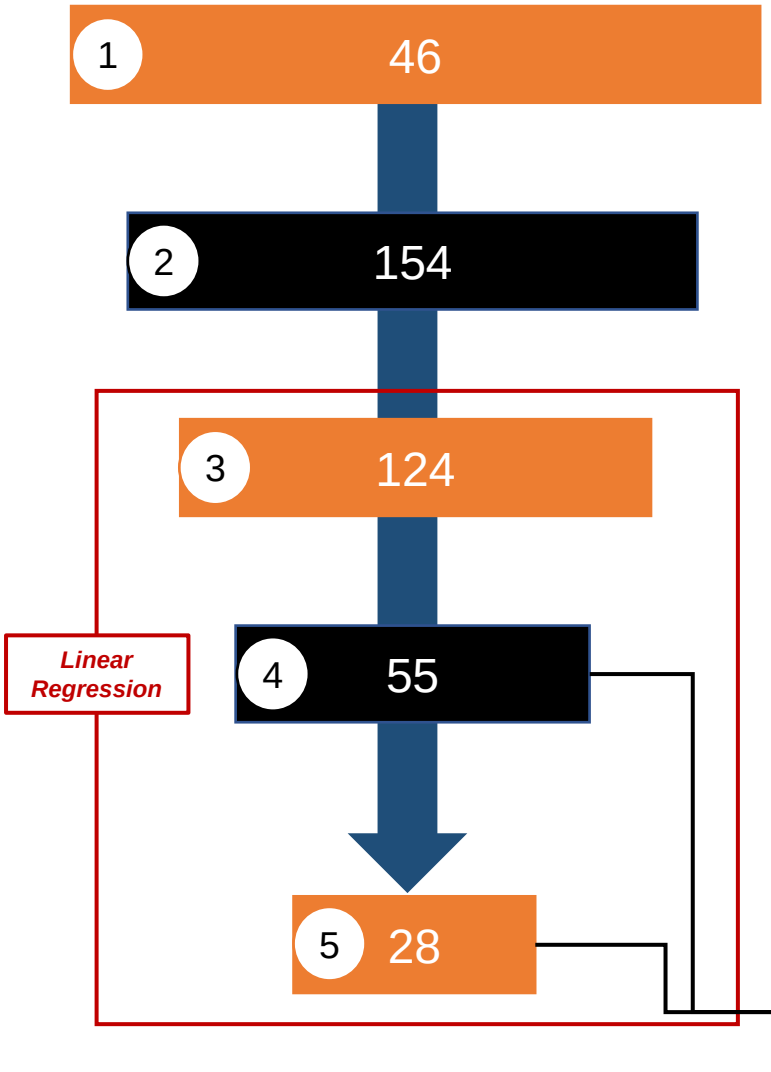


$13.5M – highest housing price in dataset



Removed outliers: where House Price > $500,000

Two Linear Regression Iterations Resulted in Accuracy of ~30%



PROCESS DESCRIPTION

- *"description" column – processed using NLP*
- *"city" and "homeType" columns → dummy variables*

- *Sequentially dropped features with correlation >0.5*
- *No multicollinearity identified*

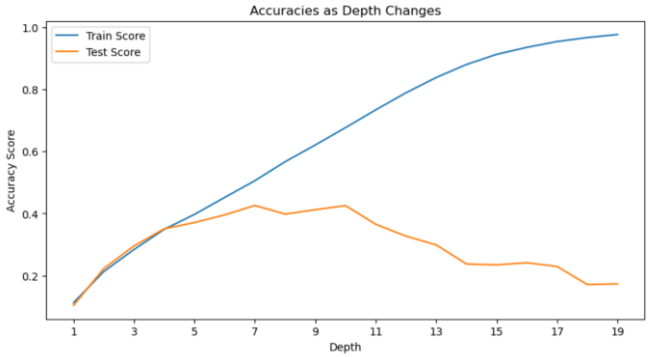
- *Dropped features that had p-values >0.05*

Initial Linear Regression → *55 features; large coefficients* → ★ *Removed outliers* → *Linear Regression* (*Dropped features that had p-values >0.05*)

Train Score: ~31%
Test Score: ~33% ✗

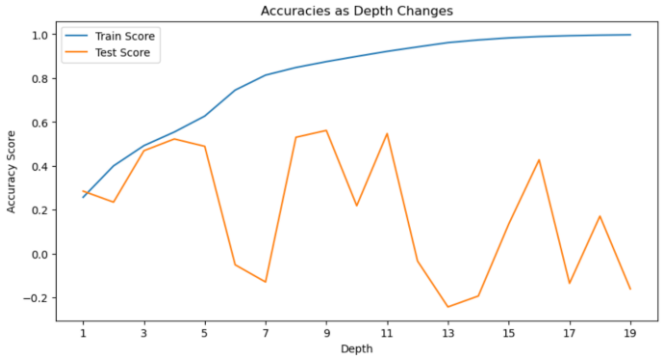
Decision Tree Regressor (with 4 Nodes) Provides 52% Accuracy in 2nd Scenario

28 Independent Features



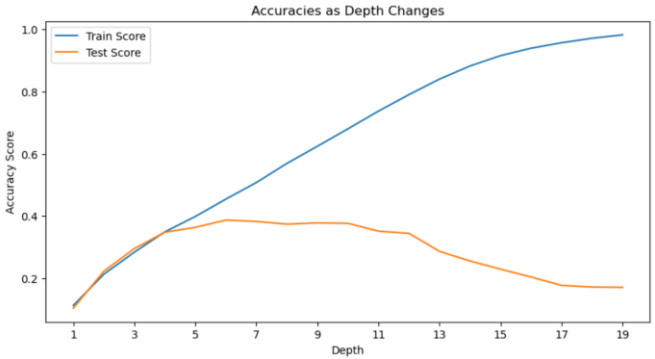
Train Score: 35%
Test Score: 35%

124 Independent Features



Train Score: 55%
Test Score: 52%

124 Independent Features &
★ Removed Outliers



Train Score: 35%
Test Score: 35%

Conclusions & Next Steps

Decision tree regressor (before removing outliers) gives the best accuracy result of 52%

However, I would generally expect higher accuracy (70-80%) given the good amount and relevancy of the independent features available to us

---

Next steps – explore other available regressor models attempting to identify one with a higher test score

---

Optimize hyperparameters of other available regressor models to tune up their performance