Austin Predictions:
Securing a Stronger and Bolder Housing Market

Area 51

Brookfield Residential

Anique Khawar, Chelsea Clinger, Eric Pierce, Jose Sifontes, Owen Nguyen

Table of Contents

Executive Summary

Overview

Brookfield Residential Properties Inc. is a leading North American land developer and homebuilder with operations in 11 major markets. The company entitles and develops land to create master-planned communities and build and sell lots to third-party builders, as well as to their own homebuilding division.

With the use of data analytics, Brookfield Residential Properties has contracted Area 51 Consulting to explore the Austin market and make recommendations to improve business operations.

Goals and Objectives

The goal of our engagement with the client is to create a comprehensive suite of tools that will allow the brokerage's real estate associates to maximize their potential lead conversions, increase understanding of which property features maximize return on investment, predicting an appropriate listing price for the property, and if time permits a recommendation engine that allows the associate to identify potential investments that the seller can make on the property to increase their overall profits.

Customer Segmentation:
Create a clustering model that can help identify the optimal demographic composition of leads that are most likely to buy.

Understand What Makes a Property Valuable:
Create a clustering model that can identify features of a property that increase potential return on investment.

Sale Price Prediction:
Create a model that provides real estate associates an accurate prediction of the sales price of a property based on key features.

Deliverables

Area51 has delivered the following deliverables for the client:

Analysis and Analytical Models

Our team produced a comprehensive descriptive analysis of the data, its makeup, trends, and observations. This analysis enables our team to develop recommendations to our client for better home pricing in the Austin market.

As a part of the analysis, we also produced a model building methodology and the resulting analytical models, which will both be available for updating and training by the client, will be leveraged in our model application.

Insights Dashboard

Our team distilled and visualized the unique and informative insights uncovered during the analysis in the form of an insights dashboard. This will be essential for our client to understand hidden and influential trends in the Austin housing market.

Mobile Pricing Field Agent Application

Our team built a native mobile application for customers who have listed their home but want to see what type of impact updates and home rehab would have on the potential market value. The application uses data pulled from the customer's actual home listing on Zillow as the input to the models the Area 51 team generated and allows customers to see the pricing impact of various updates including the addition of bathrooms, bedrooms and more.

Web-Based Pricing Research Application

Our team built a web-based pricing research and recommendation application for customers who are thinking about listing their properties. This application uses models developed by the team to enable potential listers to input several key features of their home and determine what the recommended price is. This application also enables customers and their real estate agents to determine the optimal advertising strategy to employee across factors such as geographics and demographics.

Modeling Approach

The primary objectives of our engagement are to create two distinct models for both price prediction and customer segmentation for advertisement targeting. Our modeling approaches vary depending on the objective; however, for all modeling approaches we follow the CRISP-DM methodology in which we first aim at understanding the business problem, followed by understanding the data, preparing it for modeling, developing our models, evaluating the results, and then presenting it to the client as a deliverable for future deployment.

Customer Segmentation for Advertisement Targeting

We are using unsupervised learning algorithms to create groups of sold homes which share similar characteristics, both derived from the housing listings themselves, as well as location based demographic metrics.

Price Prediction

We are creating the development of our price prediction models in contrast using a series of regression techniques and a model tournament format to arrive at the most optimal and predictive model.

Following the model development methodologies, the next step in the process after identifying the business problem is understanding the data. Our dataset is sourced from Zillow's Listing API, combined with the U.S. Postal Service Zip code demographics dataset. The Zillow Data contains approximately 15,269 records and 47 variables related to home purchasing sales including lot sizing, square footage, amenities, home features, neighborhood, and school district information. The U.S. Postal Service Zip code data contains demographic information for each zip code within our target area of the Austin, TX metropolitan area. This includes median income, age, and race groups, and education levels for any zip code.

We have conducted an in-depth analysis of the data using exploratory data analysis techniques. In this analysis, we evaluated the quality of the information to ensure proper detection and remediation of missing data, erroneous data, outliers, and anomalies that might be detrimental to the creation of accurate and useful models for our client. After finalizing and certifying the cleanliness of our dataset, we began the development of new variables and metrics to use as inputs to our models via feature engineering methods and variable transformation calculations. Our feature evaluation also includes several variable reduction methods such as correlation analysis and principal components analysis that help simplify and reduce the size of our datasets. A reduction in size ultimately reduces the computation power needed for our models, limits the data needed to provide a prediction, and increases explainability while still providing us with similar if not better predictive or clustering power.

Key Insights

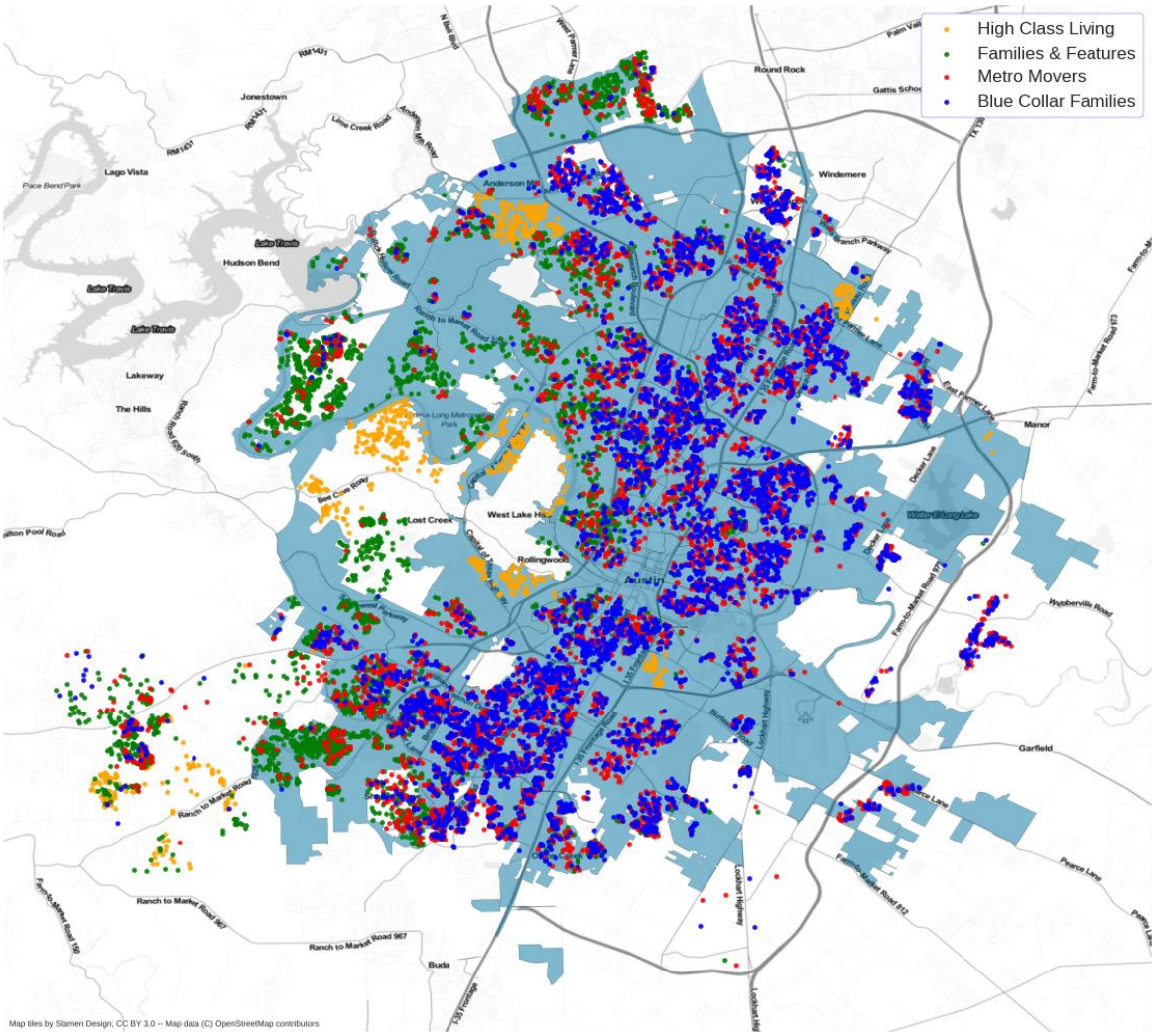
Customer Segmentation and Clustering

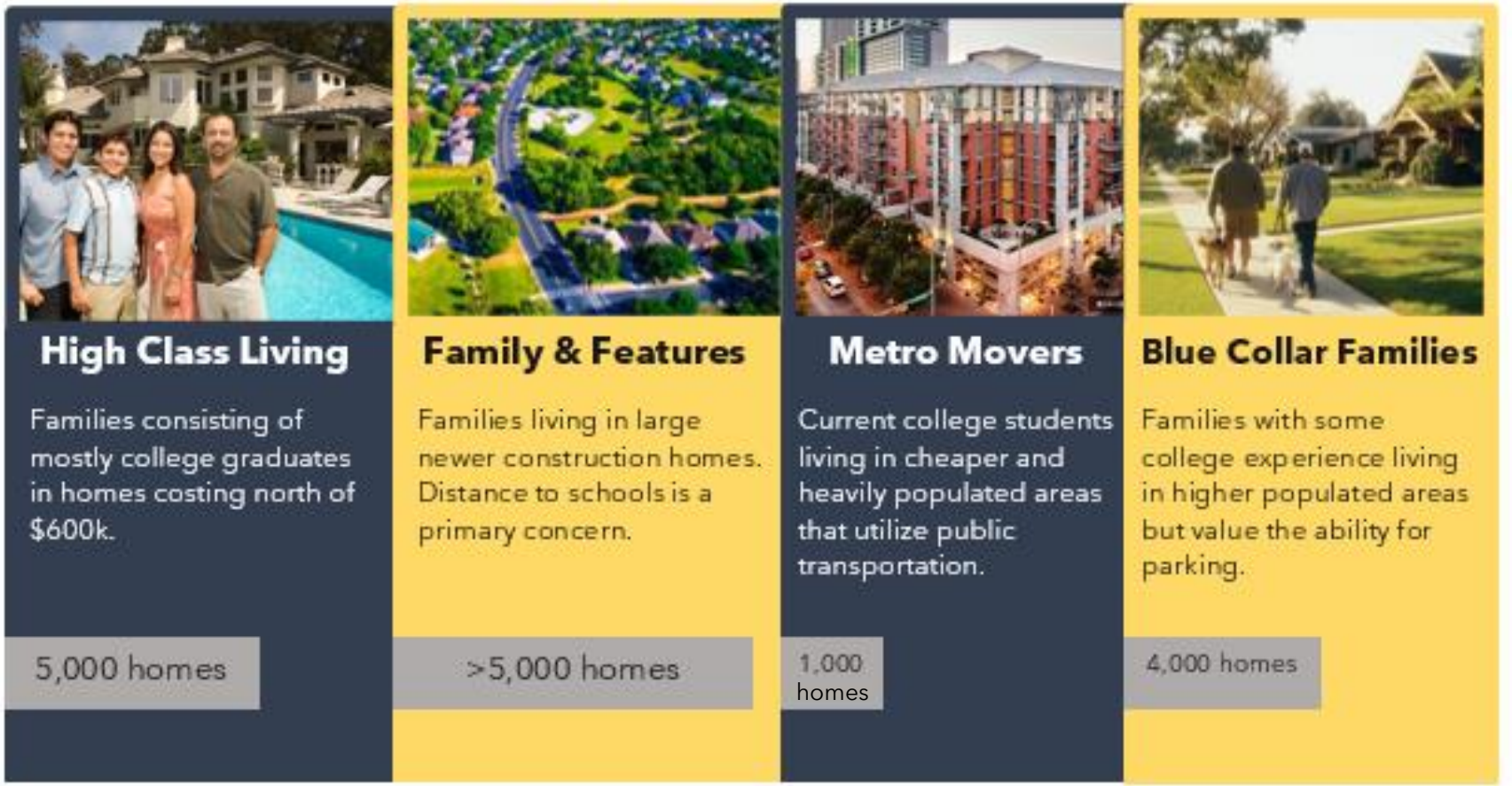
We have identified four distinct clusters in the Austin Housing market using K-Means methods.

- High Class Living
- Families & Features
- Metro Movers
- Blue Collar families

The identification of these clusters will assist the client in aligning advertisements with the most suitable customer base. Better customer segmentation will prevent unnecessary spending in less desirable markets and allow for increased ROI on advertising to identified clusters.

While each cluster has multiple variables that make it unique, one that is easy to see is location. Below is the location of each home sale in Austin color coded by cluster. It becomes very easy to see that there is a difference between east and west Austin that should be explored further.





Understanding the makeup of the Austin real estate market is vital for targeted advertising.

Using the above four clusters, we know the areas to focus advertisements:

- Larger, more expensive homes
- Homes closer to well-rated schools

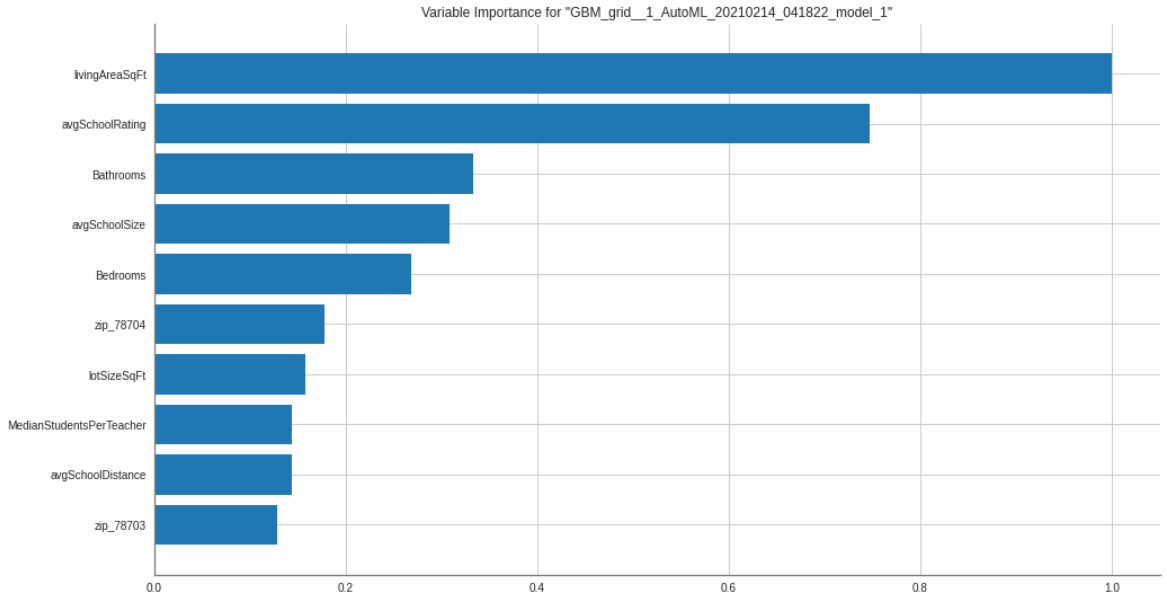
This will ensure that we connect with our desired customer base each time.

Price Prediction

We have identified 9 features as the key predictors of housing price.

It is not surprising that the most impactful variable is the size of the home, this will without a doubt drive up the price of a home.

Austin has some of the best middle and high schools in Texas, and people will pay to be near them. According to the U.S. News and World Report, 22 Austin-area high schools are nationally ranked. The desire to be in these school districts puts school ranking as the second most impactful variable.



To successfully utilize the most sufficient model for our analysis we implemented a tournament system to score and rank several models. The output of these models are listed on the table to the right.

The winner of this tournament was overwhelmingly the Automated Machine Learning (AutoML) model. AutoML has many abilities such as automated hyperparameter tuning functionality and its own secondary tournament that gives it a leg up against the manually created and tuned models.

Despite all the advantages used for the AutoML model, the random forest and manual ensemble models followed closely behind.

Model_Name	MSE	RMSE	R2	MAE
AutoML Regression	0.02	0.13	0.71	0.07
Tuned Random Forest Regression	0.02	0.13	0.70	0.08
Ensemble Regression	0.02	0.13	0.70	0.08
Random Forest Regression	0.02	0.13	0.70	0.08
Bagging Regression	0.02	0.14	0.68	0.08
XGBoost Regression	0.02	0.14	0.68	0.08
Gradient Boosting Regression	0.02	0.14	0.68	0.08
Linear Regression without PCA	0.02	0.14	0.65	0.09
Lasso Regression without PCA	0.02	0.14	0.65	0.09
ElasticNet Regression without PCA	0.02	0.14	0.65	0.09
Neural Network Regression	0.02	0.14	0.65	0.09
Linear Regression with PCA	0.02	0.15	0.63	0.09
Lasso Regression with PCA	0.02	0.15	0.59	0.10
ElasticNet Regression with PCA	0.02	0.15	0.59	0.10
ADABoost Regression	0.03	0.18	0.42	0.13
Decision Tree Regression	0.03	0.19	0.41	0.11

Computer Vision

The data we pulled from Zillow included links to the images uploaded to the posting. As image data can be highly valuable and potentially predictive, we leveraged computer vision modeling methodologies to determine how useful the image is in terms of price prediction.

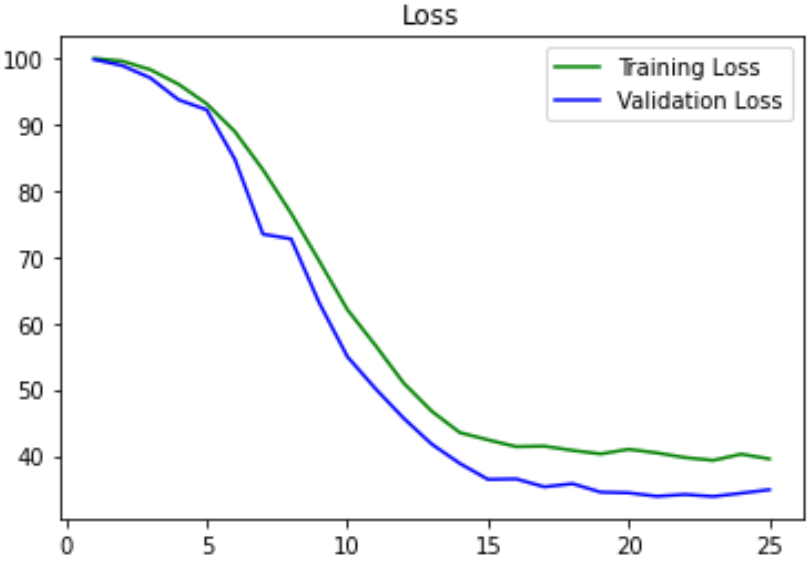
For model architecture we built a Convolutional Neural Network. This model uses a fully connected linear activation function, enabling us to use image data as input for a regression based deep neural network.

We use mean absolute percentage error as the loss evaluator and trained the model over 25 epochs with a batch size of 8.

Ultimately, our model had a mean absolute percentage error of 34.87%, indicating that by using the image data alone, on average our predicted price is around 35% off of the actual predicted price. While this surprisingly accurate for just image data alone, we felt it was not accurate enough for the standards and expectation of the Area51 team and our clients.

We believe that there are several reasons for our model results, including the limited amount of information that can be captured in a single image about a home listing, such as location, features, bedroom count, bathroom count, and so on. The only thing that the images will be able to base prices off of is "curb appeal", which is a rather subjective and non-specific thing to measure.

In the future the Area51 team may evaluate using image based pricing predictions as an input feature to our production models.



Mean Absolute Error Percentage

Web Application

Our Home Pricing Prediction web application is built on a responsive web framework that scales not only to all PC browsers but also all mobile and tablet browsers. The application is custom hosted at the following url: http://app.area51austin.com/ and can be accessed by our client and their customers alike at anytime from anywhere.

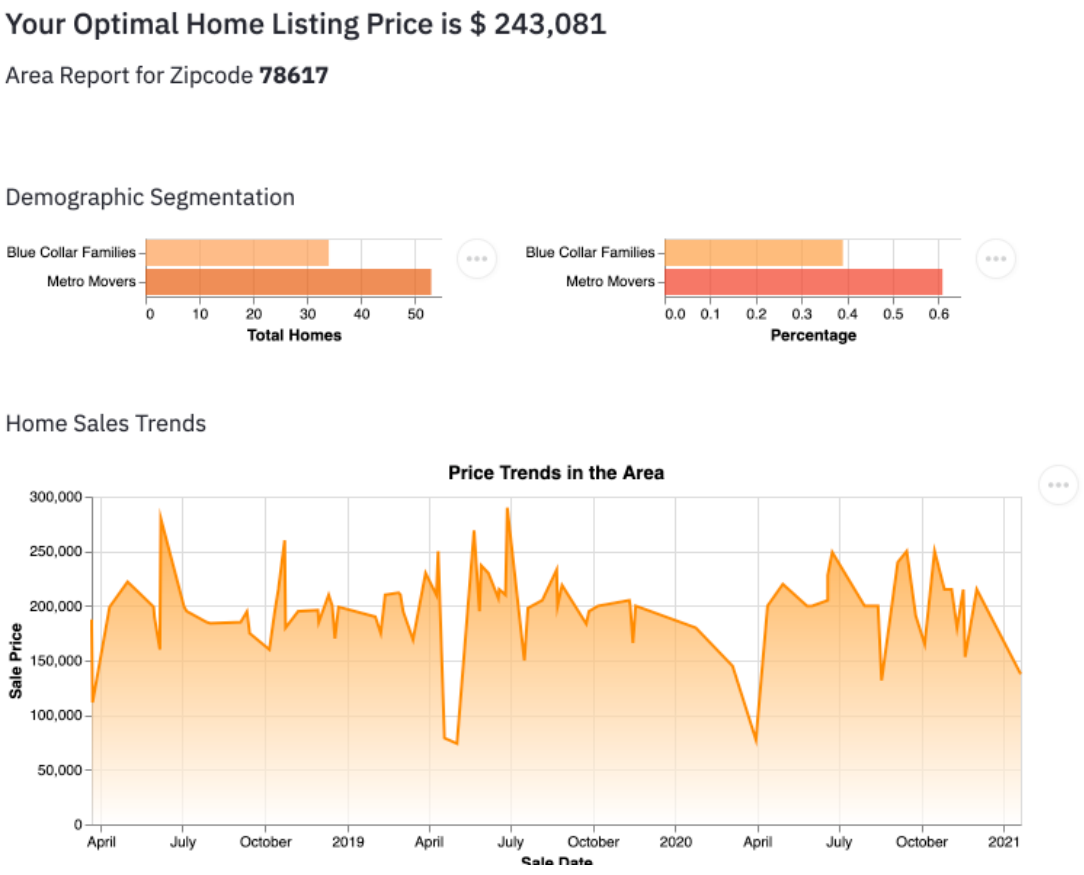
The intuitive and interactive interface allows users to input all relevant information pertaining to the home for sale, such as zip code and lot size. This information is then fed into the model to be priced. This result is based on a model that considers not only location and size of the property but also its features and the importance of the many factors that are part of the inputs. This result is based on thousands of listings spanning several years.

Price Prediction

The initial output for the prediction application will be an Optimal Home Listing Price. This is the optimal price at which the seller should consider listing the home.

Population Clusters

In addition to the price, the user is provided detailed information on the types of customers that live in that area. This will allow the agent to better target their offers.

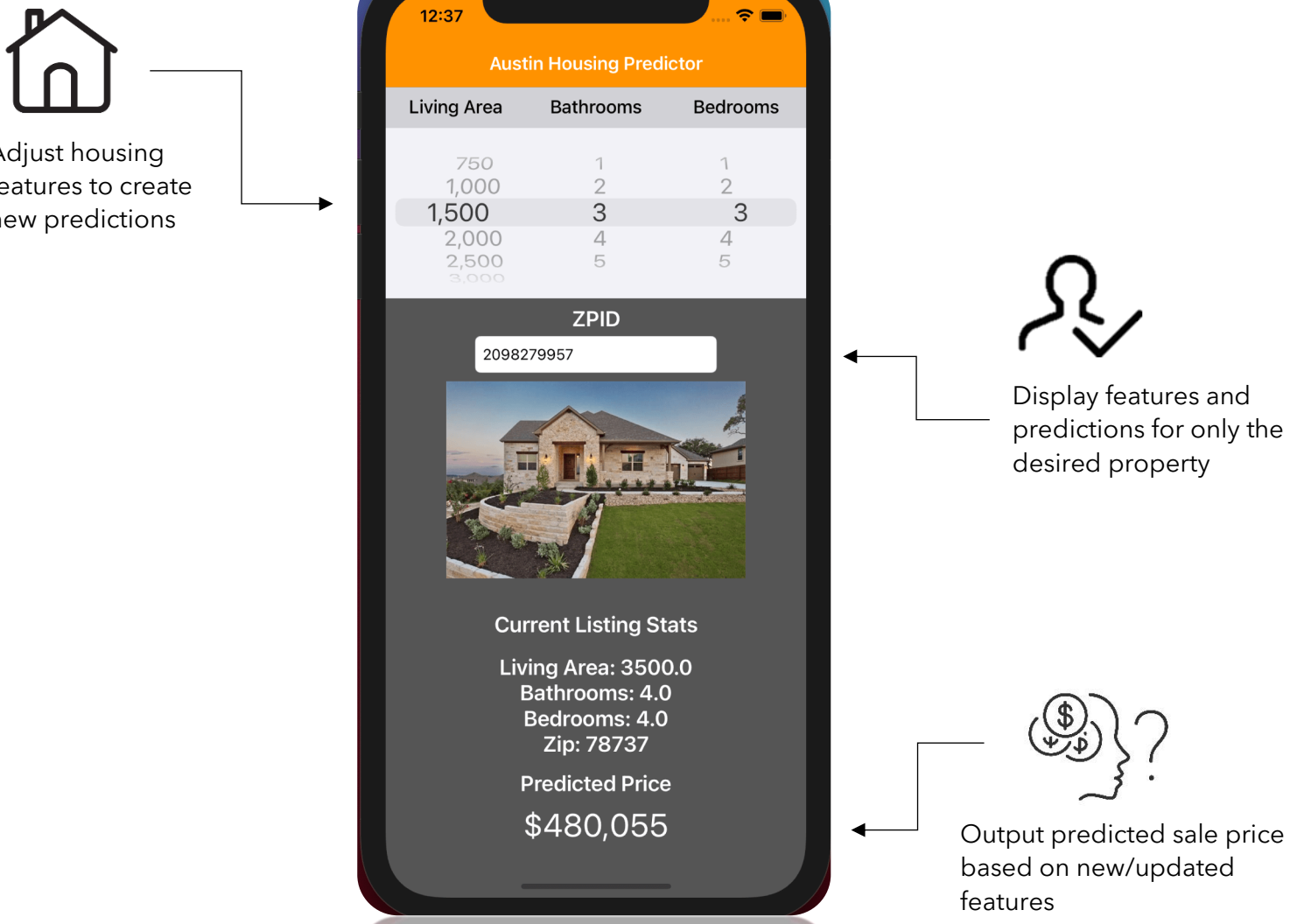


Mobile Application

The Austin Housing Field Agent is a simple mobile app to show home listers how their specific property listing would be impacted by making updates to their property.

The Area 51 team built a data pipeline which pulls listings from the Zillow site and distills a single listing into the key features which are used as inputs to our model. The user can then cycle through different configurations and modifications to their actual home to see what the impact to the price would be.

The user only needs to enter in the identifier for their Zillow listing, and the application takes care of the rest. The first image from the listing is also displayed in the application alongside some other



Insights Dashboard

An analytical dashboard has been created to assist agents in understanding the segmentation of existing home sales in Austin.

This dashboard has four components:

- Pricing Information
- School District Information
- Basic Home Information
- Location Demographics

Before diving into any of the components, the user can select the Filters button to fine-tune the selection, narrowing down to only suitable zip codes. One of the possible filters is the cluster outputs from earlier. This allows the user to look deeper into each cluster by location, price, etc.

Once all the desired fields are selected it is possible to toggle between the four pages.

Each tab provides a unique view into the top zip codes based on the selected filters. The goal of this visualization is to allow agents a quick and efficient way to begin a home search. Agents will have a starting point to begin searching for existing homes by knowing which zip codes most match the desirability of their client(s).

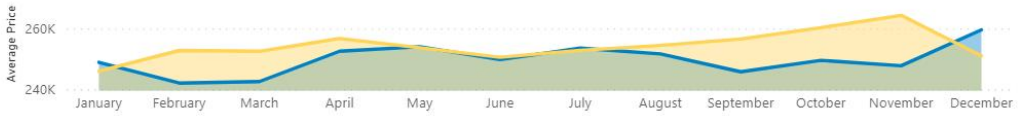
Sell in Austin!

Filters

Pricing Info | School District Info | Basic Home Info | Location Info

2 Year Average Sale Price by Month & Year

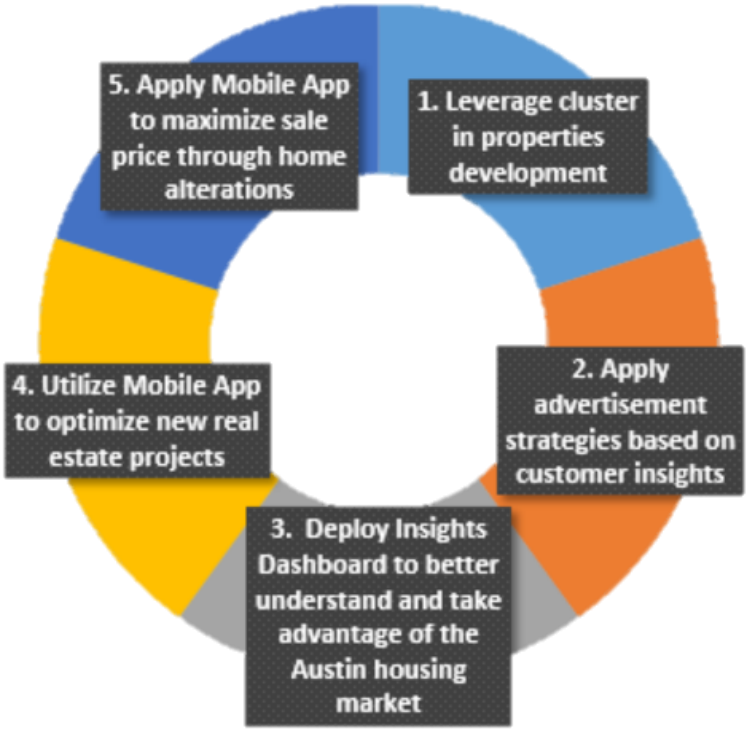
Sale Year 2019 2020



Neighboring School District Info

Zip Code	Students Per Teacher	Avg School Distance	Avg School Rating	Avg School Size	Elementary Schools	Middle Schools	High Schools
78746	14.33	1.53	9.00	1,623.00	3	3	0
78732	16.83	3.27	8.11	1,480.83	0	6	6
78733	14.00	4.08	8.00	1,373.00	8	24	0
78717	15.87	1.62	7.73	1,593.00	0	46	46
78759	15.92	1.86	7.43	1,554.20	0	25	25
78737	16.63	3.41	7.38	1,236.75	0	9	6
78739	17.00	1.03	7.33	1,633.00	0	1	1
78750	15.29	0.89	7.29	1,398.00	1	7	7
78726	16.00	1.93	7.00	1,277.00	0	1	1
78730	16.00	3.56	7.00	1,228.00	0	4	4
78731	16.00	0.74	7.00	1,483.00	0	11	11
78736	17.00	3.54	6.71	1,685.20	0	15	15
78722	15.00	1.40	6.67	1,071.14	0	7	7
78751	15.86	1.35	6.57	1,088.57	0	7	7
78705	15.69	1.42	6.41	1,179.38	0	13	13
78703	16.00	0.72	6.29	1,220.14	0	7	7
78701	16.00	1.47	6.00	1,189.00	0	1	1

Recommendations



1.Leverage clusters in new property developments

Brookfield can use the created clusters to expertly plan new development projects. Targeting properties to the right client demographic would enhance profits and improve ROI.

2. Apply advertisement strategies based on customer insights

Adjusting our strategies to target a desirable client base would reduce marketing costs and increase conversion rate.

3. Deploy Insights Dashboard to better understand and take advantage of the Austin housing market

The dashboard has a variety of use cases ranging from property identification to the planning of future projects. Increasing awareness of the surrounding area will help to keep Brookfield ahead of the market.

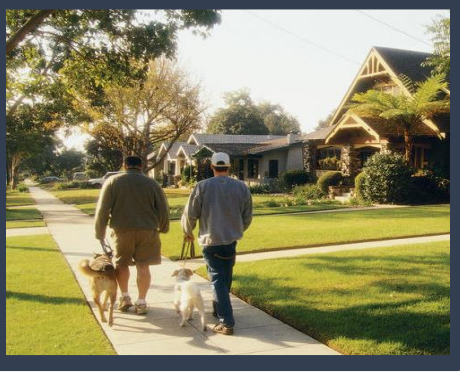
4. Utilize Mobile App to optimize new real estate projects

Our application enables real estate agents to leverage model-based pricing recommendations for properties based on key features as model inputs. This would prevent the sellers from mispricing their properties; it would also allow buyers to find good deals and to avoid overpaying.

5. Apply Mobile App to maximize sale price through home alterations

The app enables the agent and the customer to conduct what if scenarios to determine which investments are most impactful to their predicted selling price to maximize ROI.

Recommendation Details for 1 and 2

	Title	Property Development Plan	Advertisement Strategy
	**High Class Living**	Choose isolated areas that have large lot sizes. Use top-of-the-line materials. Do not cut corners to save construction expenses.	Focus on safety and comfortability. Portray ourselves as trustworthy and luxurious at the same time.
	**Family & Features**	Be in good school districts. Choose suburban locations with large lot sizes. Go for modern architecture style.	Paint a positive future outlook. Highlights the modern, comfortable, and convenient amenities. Showcase family values.
	**Metro Movers**	Ensure zero or very low association fee. Choose destinations that have multiple nearby public transports.	Showcase our convenient locations. Utilize viral marketing. Celebrity appearance could make a positive impact.
	**Blue Collar Families**	Ensure zero or very low association fee. Build large parking garages. Choose locations that are near free or cheap parking lots. Choose amenities that require low utility usage.	Highlight the cost-efficiency of the properties. Create marketing events that have price discounts.

Conclusions

The Austin housing market is hotter than ever. In this final report, we have provided an in-depth overview of our analytical process by following the best industry practices outlined by CRISP-DM.

Once data is cleaned, it is ready for modeling using clustering and regression analysis. We used K-Means clustering that find 4 clusters to describe customers in Austin: High Class Living, Families & Features, Metro Movers, and Blue Collar Families. These clusters can drive better leads for real estate agents.

We compared 16 models to predict home price and found our best model to be the AutoML regression. This model found that the most important features for determining home price are living area, average school rating, and number of bedrooms and bathrooms. Some zip codes also lead to higher home prices.

We have provided an application which can predict the price of a new listing for agents that input home features. For existing listings, we have provided a separate mobile application which agents and customers can go through to improve the value of a home. Last, we have provided a dashboard that provides insight into the Austin area.

We are confident that we have provided all the deliverables. We have provided our analysis and analytical models. Our insights dashboard provides unique visualizations and informative insights into the Austin-Texas area. We have leveraged the models as a mobile application and web application.

Final Report

Overview

Brookfield Residential Properties Inc. is a leading North American land developer and homebuilder with operations in 11 major markets. The company entitles and develops land to create master-planned communities and build and sell lots to third-party builders, as well as to their own homebuilding division.

In March 2015, Brookfield Residential Properties announced the acquisition of Austin-based Grand Haven Homes, which operates about 15 active communities with homes. This acquisition brings in an experienced team and enables Brookfield Residential to increase their presence and capabilities in the Austin market. With the use of data analytics, Brookfield Residential Properties has contracted Area 51 Consulting to explore the Austin market and make recommendations to improve business operations.

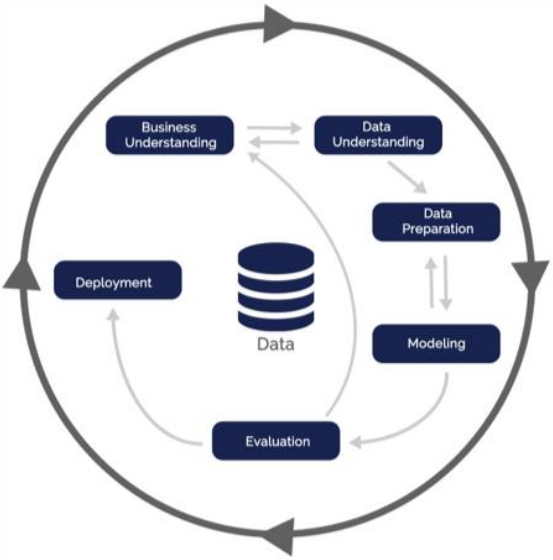
The goal of our engagement with the client is to create a comprehensive suite of tools that will allow the brokerage's real estate associates to maximize their potential lead conversions, increase understanding of which property features maximize return on investment, predicting an appropriate listing price for the property, and if time permits a recommendation engine that allows the associate to identify potential investments that the seller can make on the property to increase their overall profits.

Analytical Process

There are some dangers of neglecting market analysis:

- Never utilizing your competitive advantages
- Abandoning customer-centric marketing models
- Increasing strategic and operational risks
- Inability to make data-backed enterprise decisions
- Forgoing future business opportunities

To help Brookfield Residential Properties avoid these repercussions, the Area 51 Consulting team conducted a comprehensive analytics engagement through the CRISP-DM method. The first step is to access the situation and generate the business's primary objective. Afterwards, the team will check the data quality to know what can be expected and achieved from it. All members then collaborate to generate the finest models and to extract the best value of the pieces of information. Lastly, the team will produce actionable insights along with deploying an informative dashboard and a mobile application.



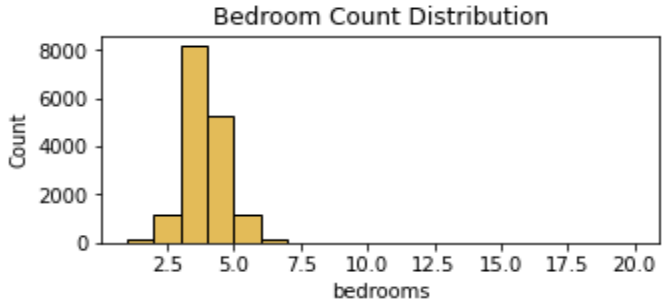
Data Overview

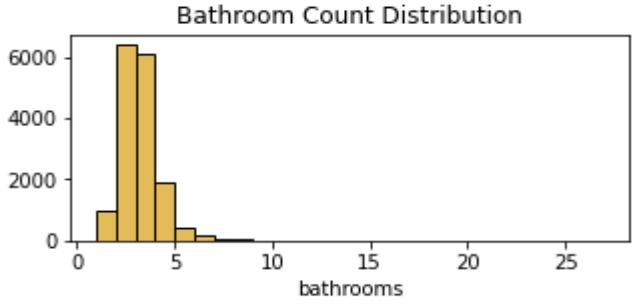
What do houses in Austin look like?

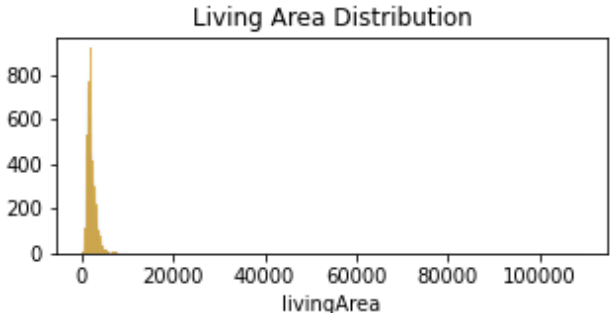
The majority of houses in Austin is single-family type. They have between 3 and 4 bedrooms, and they tend to have 2 or 3 bathrooms. The mean living area is just over 2,200 square feet. Considering the median size of an American single-family home is in the neighborhood of 1,600 or 1,650 square feet, Austin houses are significantly bigger than the rest of the nation.

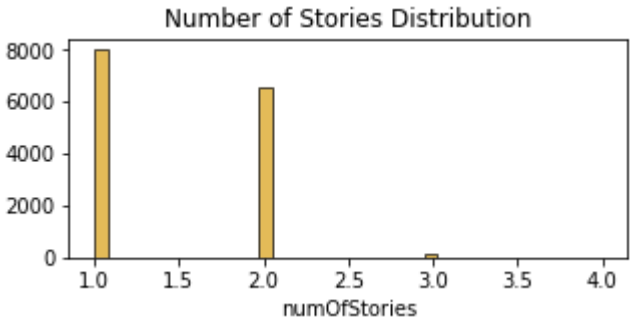
We could use this as a marketing point to lure potential investors, especially those coming from crowded places, into moving to Austin. However, this shows that Austin residents are accustomed to living with extra spaces, which could mean that future housing development projects need to be large.

Figure 1: Home Distributions









The word cloud further illustrates the importance of "large" and "spacious" in the Austin lifestyle.

We can also see that "walk" and "downtown" are often mentioned. This showcases the importance of location which is something we will dig deeper later.

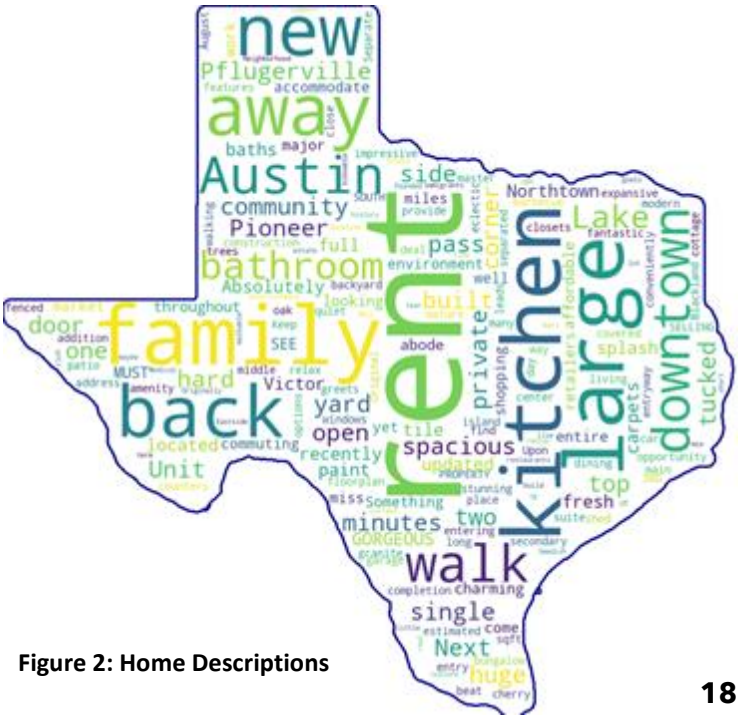


Figure 2: Home Descriptions

When to invest in the Austin market?

Regarding the mean price, the dataset shows that prices are the highest during the months of July, August, and December. It also indicates that the prices are relatively constant throughout the week with a small increase on Friday. We should adjust our transaction strategy to buy properties during the cheaper time, then sell them during the more profitable months.

Figure 3: Sale Distribution by Weekday

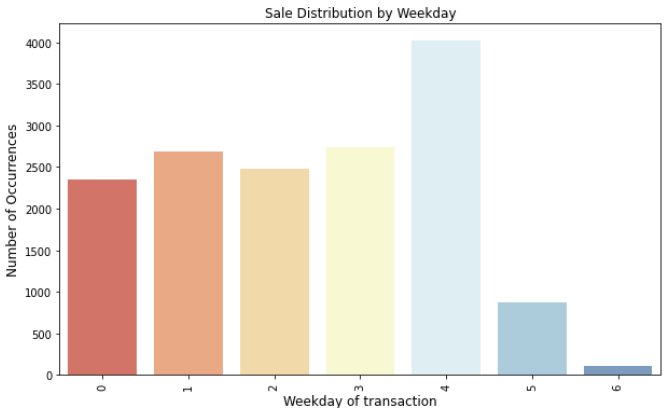
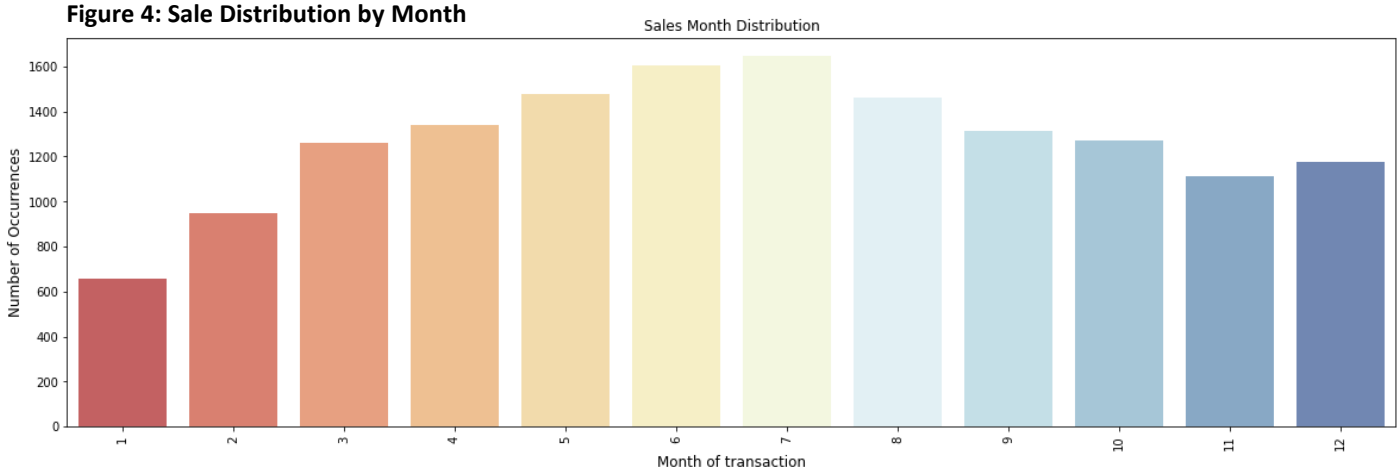
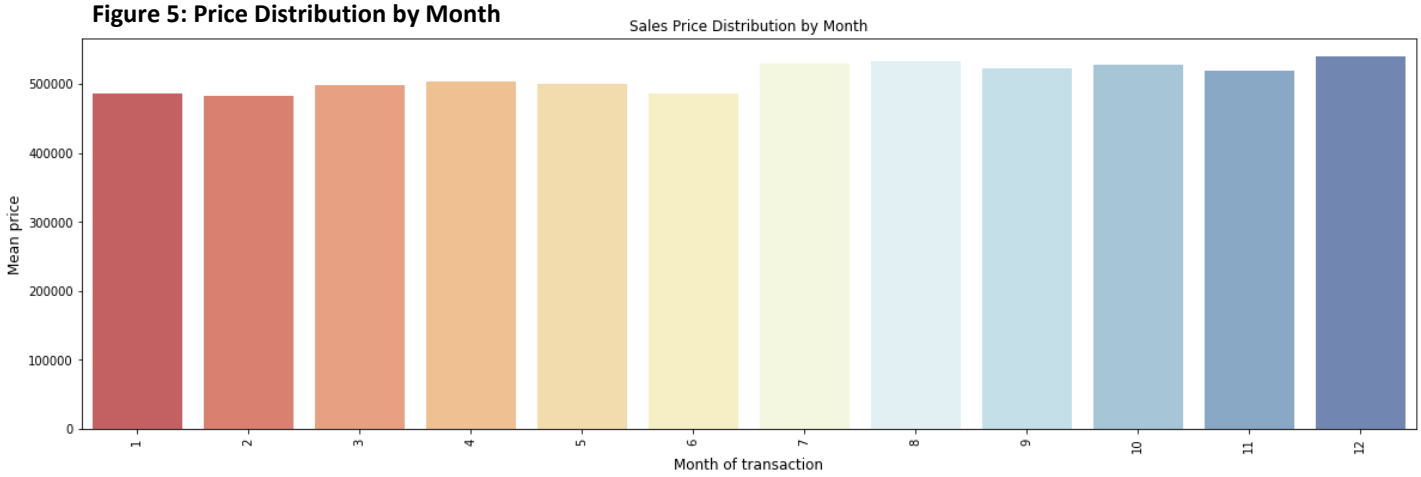


Figure 4: Sale Distribution by Month



Investigating the number of transactions, we can see that the housing market is livelier during the summer months, peaking in June and July. Checking into the transactions frequency per day, it can be noticed that sales happen most commonly on Wednesday and rarely at the end of the week. Hence, if a property needs to be sold quickly, a good sales strategy would aim for a Wednesday in July.

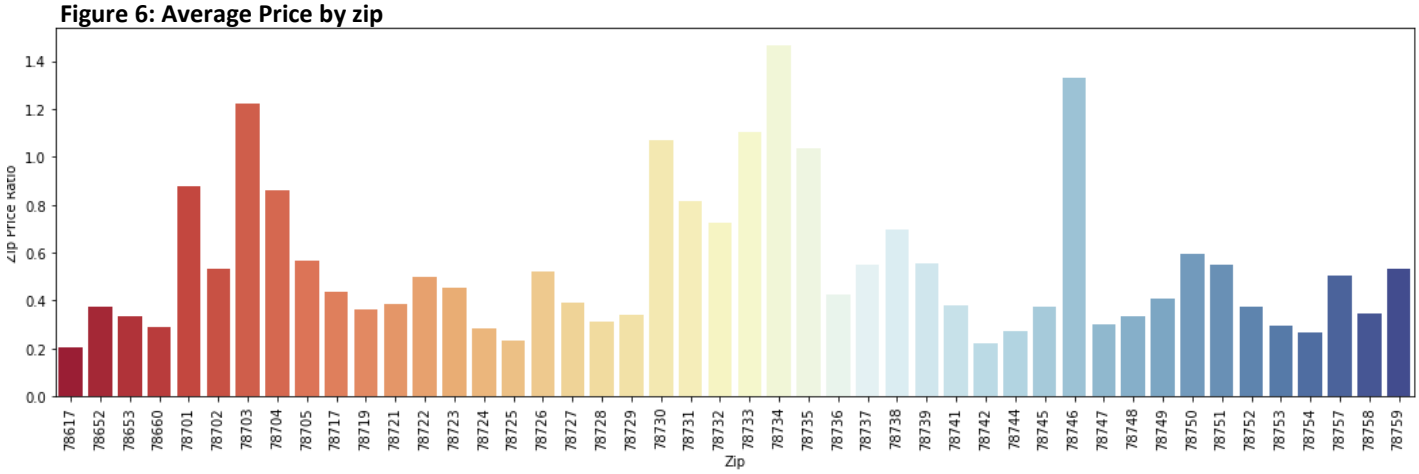
Figure 5: Price Distribution by Month



Where to invest in the Austin market?

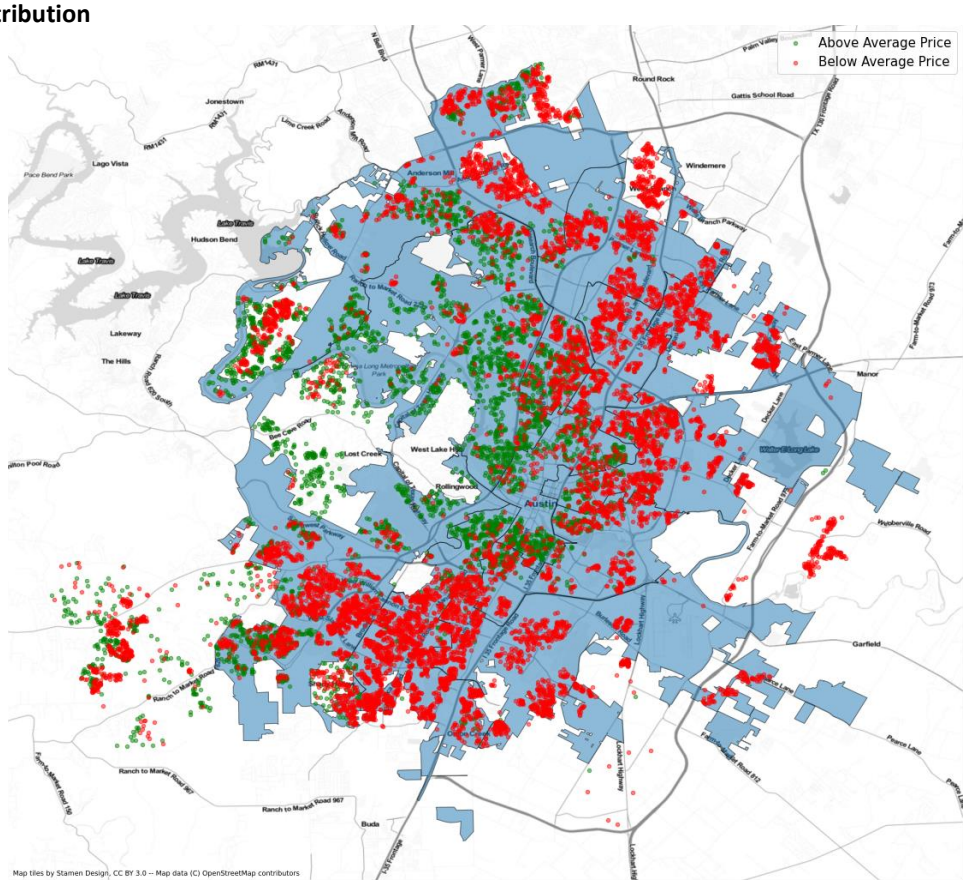
A deeper analysis was performed to understand the effect of geographical location on housing price. We could see that the price fluctuates significantly when moving from one zip code to the next. Some of the most expensive zip codes are: 78734, 78746, and 78703, while some of the cheapest properties can be found at these locations: 78617, 78742, and 78725.

Figure 6: Average Price by zip



We then categorized all real estate purchases based on whether the price was above or below the average. The map below clearly indicates a difference in prices based on locations. Properties in the north west are much more expensive than those in the east or south regions. We must plan our future projects accordingly, taking into account the impact of geographical location on prices.

Figure 7: Price Distribution

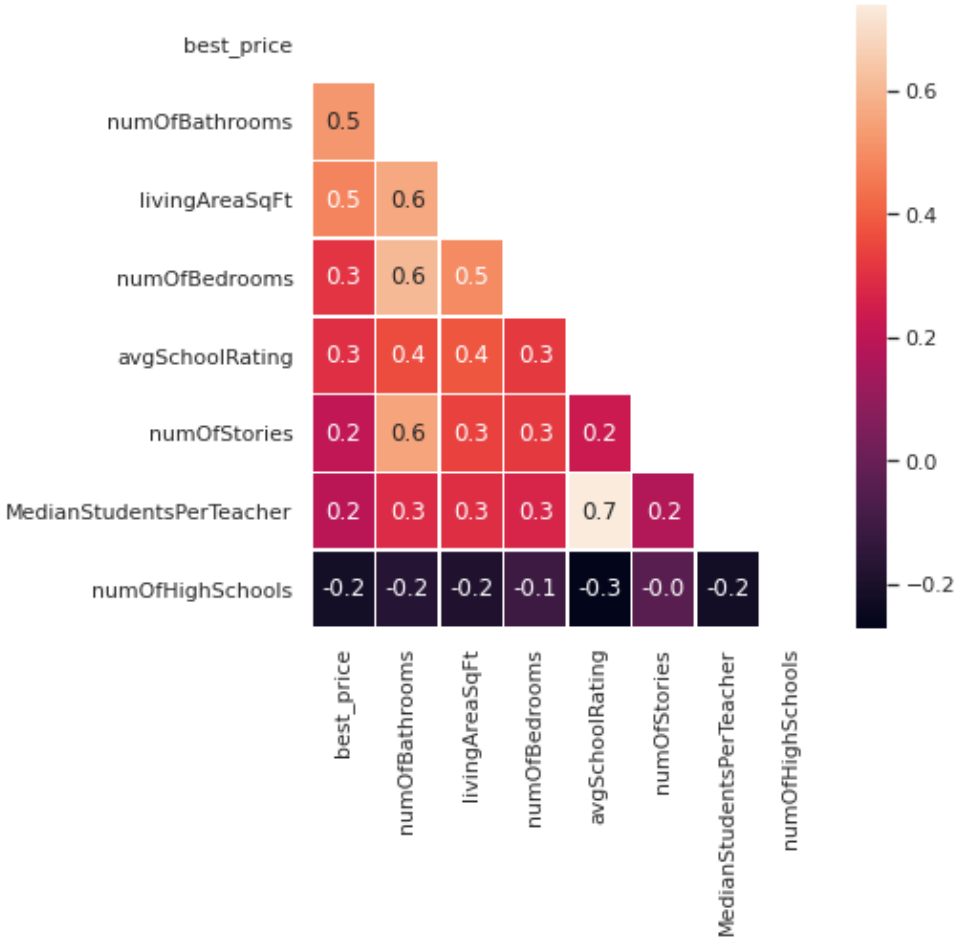


Which features are corelated with sale price?

At this point we are well educated on what makes up the characteristics of a typical Austin home. However, how does that correlate with our ultimate end goal, sales price?

A correlation plot and output of scores highlights the most and least correlated variables. There are strong correlations with sales price for bedroom/bathroom counts, living area size and average school ratings. Additionally, there is a lesser but noticeable negative correlation between sale price and the number of high schools.

Figure 8: Correlation Matrix



Feature Correlation to Sale Price

numOfHighSchools	-0.20
longitude	-0.18
numOfPrimarySchools	-0.17
address_zipcode	-0.14
propertyTaxRate	-0.06
hasAssociation	-0.00
numOfCommunityFeatures	-0.00
hasHeating	0.00
zpid	0.00
hasCooling	0.00
numOfAccessibilityFeatures	0.01
lotSizeSqFt	0.02
numOfAppliances	0.03
latest_salemonth	0.03
numPriceChanges	0.03
numOfWindowFeatures	0.05
numOfSecurityFeatures	0.05
yearBuilt	0.05
latest_saleyear	0.05
latitude	0.07
hasGarage	0.07
avgSchoolSize	0.08
numOfPatioAndPorchFeatures	0.08
avgSchoolDistance	0.09
numOfParkingFeatures	0.09
numOfWaterfrontFeatures	0.12
numOfMiddleSchools	0.12
hasView	0.13
numOfPhotos	0.15
numOfElementarySchools	0.15
parking	0.15
garageSpaces	0.15
hasSpa	0.17
MedianStudentsPerTeacher	0.19
numOfStories	0.19
avgSchoolRating	0.29
numOfBedrooms	0.29
livingAreaSqFt	0.46
numOfBathrooms	0.50
best_price	1.00

Description of Data

The data is sourced from Zillow real estate listings for the Austin-Texas area, including surrounding cities. The focus for our data is on housing prices and factors that may or may not affect this. Some of these factors can range from number of bedrooms, bathrooms, living area, types of schools, school sizes, school ratings, location-based information (latitude and longitude), and description of the dataset.

The initial pull of the data from Zillow yielded a dataset approximately 379 MB of data consisting of 59,534 rows and 1758 columns. After extensive data cleaning, the final dataset is approximately 10.9 MB in a single .CSV file yielding 15,269 unique properties and 47 housing features in the Austin-Texas area. A full list of the engineered features can be found in Table 1 below. In addition to this, Table 1 contains a description of the types of features created (count,

Table 1: Variable Descriptions

Feature Type	Variable Descriptions	Description
Housing	Count features • # of garage spaces • # of price changes • # of photos • # of accessibility features • # of appliances • # of community features • # of parking features • # of patio and porch features • # of security features • # of waterfront features • # of window features --- House-specific features • # of bathrooms • # of bedrooms • # of stories • Living area (sq. ft) • Lot Size (sq. ft) • Home type • Price --- Bundled features • Association • Cooling • Garage • Heating • Spa • View	Count features are unique entries across multiple columns for each property. House specific features are information we expect when looking for a home, such as number of rooms and stories or the square footage of a home. Bundled features are features that may or may not be present when purchasing a home, such as air conditioning.

Geographical	• City • Street address • Zip code • Latitude and Longitude	Related to the location of the property
Listing Information	• Description • Zillow Property ID • Sale datetime/date/month/year • Latest sale price	A text-based description from the property listing on Zillow, a digital ID used to uniquely identify the property, date information which agent listed on the property.
School Features	• # of Primary schools • # of Elementary schools • # of Middle schools • # of High schools • Average school distance • Average school rating • Average school size • Median students per teacher ratio	Information about schools that are listed for each property. Each property lists up to 4 schools in the area, including their distance in miles, rating, student body sizes and student to teacher ratio.

Transformation of Data

We gathered data by pulling property listings from Zillow. The unclean dataset consists of 59,534 rows and 1,758 columns. Many of the columns contain missing data, and many of the desired features require feature engineering across multiple columns. For data exploration and modeling, we require a clean dataset. Data is cleaned linearly as described in the following sections below.

Deduplication and Filtering

To begin, we drop any duplicate data based on the Zillow Property ID. The Zillow Property ID is a unique identifier used to identify a property listing on the Zillow website. We kept the first instance of each unique property. Next, we drop any columns that contain 100% null values. These steps whittle down the dataset to 22,083 rows and 1,337 columns.

The most important part is to make sure we are sticking to the Austin-Texas area. During data sourcing, some properties are in other states. The best way to do this without losing data points was by filtering by latitude and longitude between 30 to 31 and -99 to -97, respectively. This enables us to capture property listings in the Austin-Texas area, including some surrounding cities. The dataset identifies some properties that are rentals, so we try to filter them out. We reduce the number of properties to 19,297 while the columns remain the same.

Feature Creation and Cleaning

We need to extract features from this dataset by collating multiple columns together. We extract price by picking the most recent price for a property across 30 price columns. We use a similar approach for other features such as number of price changes, bathrooms, bedrooms, stories, and the year the property was built. This not only helps us extract those features but to fill in missing data.

We create datetime features such as the date, month, and year a property is listed for sale. The years range from 2018 to 2021. We rename a few features to make them easier to work with.

We count the number of unique entries across multiple columns in what we call count features. A list of count features and their descriptions can be found in Table 2. We extract other features such as living area and lot size and respectively convert them to square feet from acres when appropriate.

We created school-related features, such as counts of school types (i.e., middle school, high school). We averaged the school ratings, distance, and size across all schools per property. For student to teacher ratios, we opted for using the median because some schools have small student-to-teacher ratios, which can bring down the average.

In sum, we created 31 new features and cleaned some existing features.

Table 2: Count Features Created

Variable	Description
# of Accessibility Features	Number of unique accessibility features listed
# of Appliances	Number of unique appliances listed
# of Community Features	Number of unique community features listed (i.e., conference rooms, cluster mailboxes)
# of Elementary schools	Number of Elementary schools in area listed
# of High schools	Number of High schools in area listed
# of Middle schools	Number of Middle schools in area listed
# of Parking Features	Number of parking features (unique entries such as an attached garage or a parking spot)
# of Patio and porch features	Number of patio and porch features listed
# of Photos	Number of photos provided for a listing
# of Primary schools	Number of Primary schools in area listed
# of Security Features	Number of unique security features listed
# of Waterfront Features	Number of waterfront features listed (if a property is near a body of water.)
# of Window Features	Number of window features listed (if windows come with blinds, double pane, and vinyl windows, etc.)

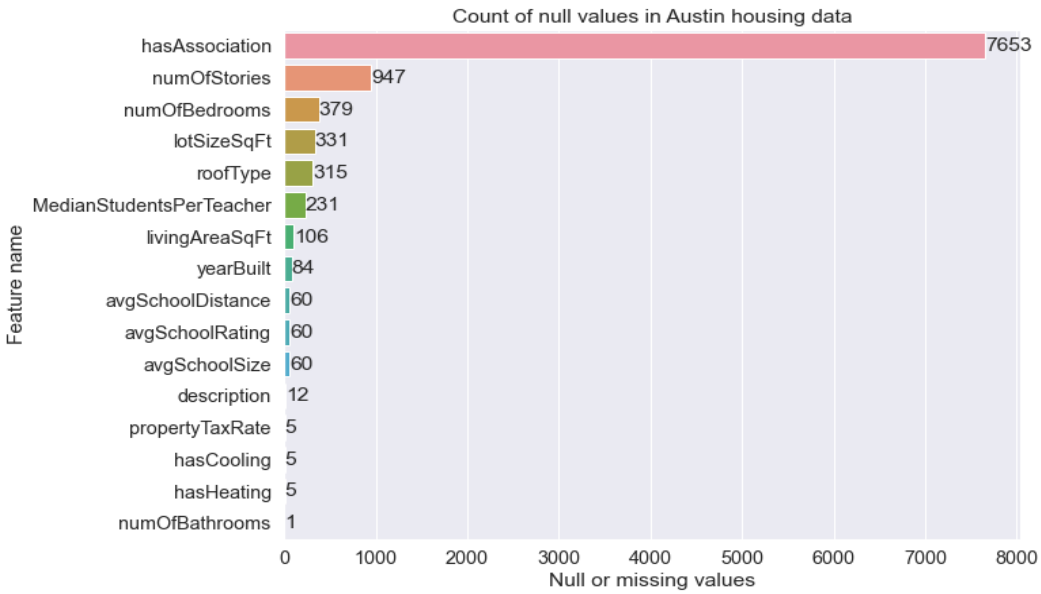
Additionally, binary 1 or 0 'has' features were created to indicate the presence of certain features such as desirable views and garages.

Missing Data

After feature engineering, we drop any data used for creating price, count, and school features. We removed columns that contained mostly null values and comprised entirely only 1 unique value, such as listings for neighborhood or subdivision. Our criteria for removing null values is when it is unfeasible for the column to be substituted in. Last, we exclude any properties below the $5,000 price point to hopefully capture any rental properties that may have slipped through during data filtering. The results from all these steps bring us down to 16,181 properties with 52 features across.

Many of the columns still have missing data, so they need to be addressed individually. To get an idea of the magnitude of these null values, please refer to Figure 9 below.

Figure 9: Null Counts in Austin data



We opted to remove any properties missing description, property tax rates, cooling, heating, number of bathrooms and number of bedrooms. We dropped the roof type column because it is unclean and contains information not directly related to a property, such as “contacting the agent”. We remove properties before the year 1900 and any outliers for living area and lot size that are less than 100 square feet reported. We fill in the rest of the missing values using a strategy of common sense, averages, and rounding.

For the number of stories a property has, it is safe to assume that a property is 1 story if that field missing. For school distance, rating, size, and median student to teacher ratio, we use the average for that column.

We use summary statistics (minimum, maximum, mean, median, standard deviation) to find any peculiarities in individual columns, such as decimal values for columns that should have a value of only whole numbers. For example, we found that school size and median student to teacher ratio should be a whole number, so it was rounded appropriately.

Table 3 summarizes the steps taken to deal with missing data by presenting the variable, action, and justification behind cleaning a specific column.

Key	
	Filling in missing data
	Dropping missing rows
	Dropping columns

Table 3: Strategy for columns with missing or bad data

Variable	Action	Justification for action
Average school distance	Fill in with average for column	To reflect the entire Austin area where missing.
Average school rating	Fill in with average for column	To reflect the entire Austin area where missing.
Average school size	Fill in with average for column and round	To reflect the entire Austin area where missing.
Cooling	Drop rows where missing	Low number of missing values.
Has association	Fill in missing values with False	If an association is not listed with a property, it is safe to assume that there is no association.
Heating	Drop rows where missing	Low number of missing values.
Living area	Drop rows where missing, filter out rows >= 100 sq. ft	Outliers of 1 square foot exist for living area, which is unrealistic.
Lot size	Drop column	Column is filled with dirty data and includes living area, front yards, and backyards.
Median students per teacher	Fill in with average for column and round	To reflect the entire Austin area where missing.
Number of bathrooms	Drop rows where missing	Low number of missing values.
Number of bedrooms	Drop rows where missing	A safe approach over averaging because it can affect modeling and correlation results
Number of stories	Fill in missing values with 1	Most properties are found to have 1 story.
Property tax rate	Drop rows where missing	Low number of missing values.
Roof type	Drop column	Too many unique values found which were not related to roof type.
Year built	Drop rows where missing, filter out rows greater than the year 1900	Homes before the year 1900 do not reflect most homes in Austin.

Removing Outliers and Irrelevant Data

Some columns do not properly reflect most properties found in the dataset or serve no purpose for our analysis. These columns are found in Table 4 below, along with justification for removal.

At the end of this step, our final dataset has 15,269 unique rows and 47 columns.

Table 4: Summary of columns to drop

Column name	Justification to remove
Home Status	Used for filtering out rental properties but only consists of 2 values: 'sold' and 'recently sold'. A recently sold property is any property sold in the last 365 days, although we already have datetime data.
Furnished	Only 39 homes out of 15,269 properties are found to be furnished.
Has Attached Garage	There already exists a feature called 'hasGarage'.
Has Carport	Only 7 homes out of the 15,269 properties have a carport.

Exporting Data

We need to export data for exploration and modeling amongst the team. Before exporting the cleaned dataset to a CSV file (comma-separated values), we need to convert data types where appropriate. We trim off the decimal point for any values that should be integers. We can find a summary of the dataset before and after data cleaning in Table 5 below.

Table 5: Summary of before and after data cleaning

File	Uncleaned data	Cleaned data
Columns	1,758	47
Rows	59,534	15,269
File size	379 MB	10.9 MB
Memory usage	793 MB	4.7 MB

Modeling Approaches

The primary objectives of our engagement are to create three distinct models for both price prediction and customer segmentation for advertisement targeting. Our modeling approaches vary depending on the objective; however, for all modeling approaches we follow the CRISP-DM methodology in which we first aim at understanding the business problem, followed by understanding the data, preparing it for modeling, developing our models, evaluating the results, and then presenting it to the client as a deliverable for future deployment.

For the advertisement targeting, we are using unsupervised learning algorithms to create groups of sold homes which share similar characteristics, both derived from the housing listings themselves, as well as location based demographic metrics. We create the development of our price prediction models in contrast using a series of regression techniques and a model tournament format to arrive at the most optimal and predictive model.

Additionally, we standardized and centered our data prior to modeling. This helps us control the variability of our data and ensures that we are not putting more weight on variables that have smaller variances. As we continue to evaluate and stabilize our dataset, we will refine our two different clustering approaches: Hierarchical clustering and k-means clustering. Along with the cluster definitions, we append the zip code demographics data to our clusters to define the advertising targets for the model's output.

The second model that we developed was the Price prediction and Home value ROI model. The primary goal of this model is to predict what a home will sell for, and what are the best investments the seller can make in their home that would result in a higher selling price while also giving the most bang for your buck. We used several regression methods and established a model tournament as a means of comparison and chose a champion model that will be used in our final production deliverable.

For our regression modeling tournament, we selected models that range in complexity from simple ordinary least squares regression, all the way to a multi-layer deep neural net regression and Ensemble regression models. We evaluate model performance using several metrics such as the Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-Squared. The goal of the models is to minimize the error while retaining the highest levels of explained variance or prediction power. Unquestionably, we understand that the more complex the models the more difficult it will be to explain to our audience so the preference would be that the winning models are simpler to explain than the more complex "black box" models. Our final model selection will consider both ease of mathematical explanation as well as accuracy and variance explanation.

Wherever applicable, we will partition our data using 70% of the data for training and 30% for testing and validation. Additionally, we will conduct several cross-validation tests to ensure that we are using the most valid scoring practices while eliminating the possibilities of data leakage within our models.

Advertising Target Clustering Model

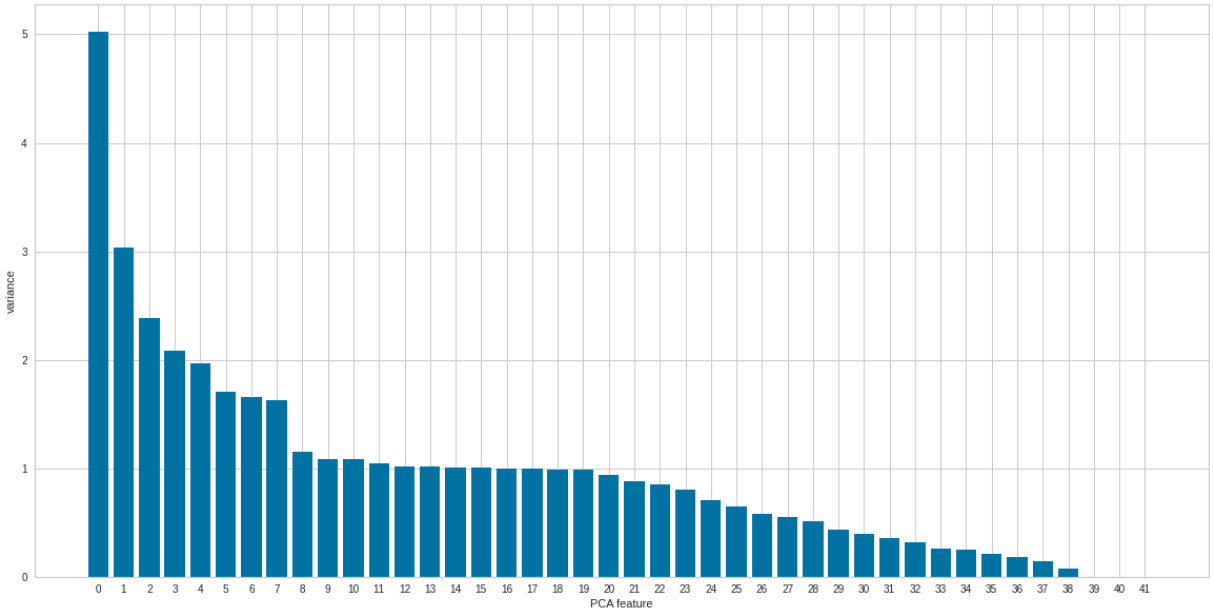
The goal of this model is to create listing segments that our client can use to align advertisements to customers, based on specific listing, geographic, or demographic interest. The value in this approach is that with targeted ads, we should be able to reach an audience that is more inclined and interested in properties that the agents are selling and align listings to customers which may not be considering them. Our aim is to drive better leads, lower time to conversion, and more conversions.

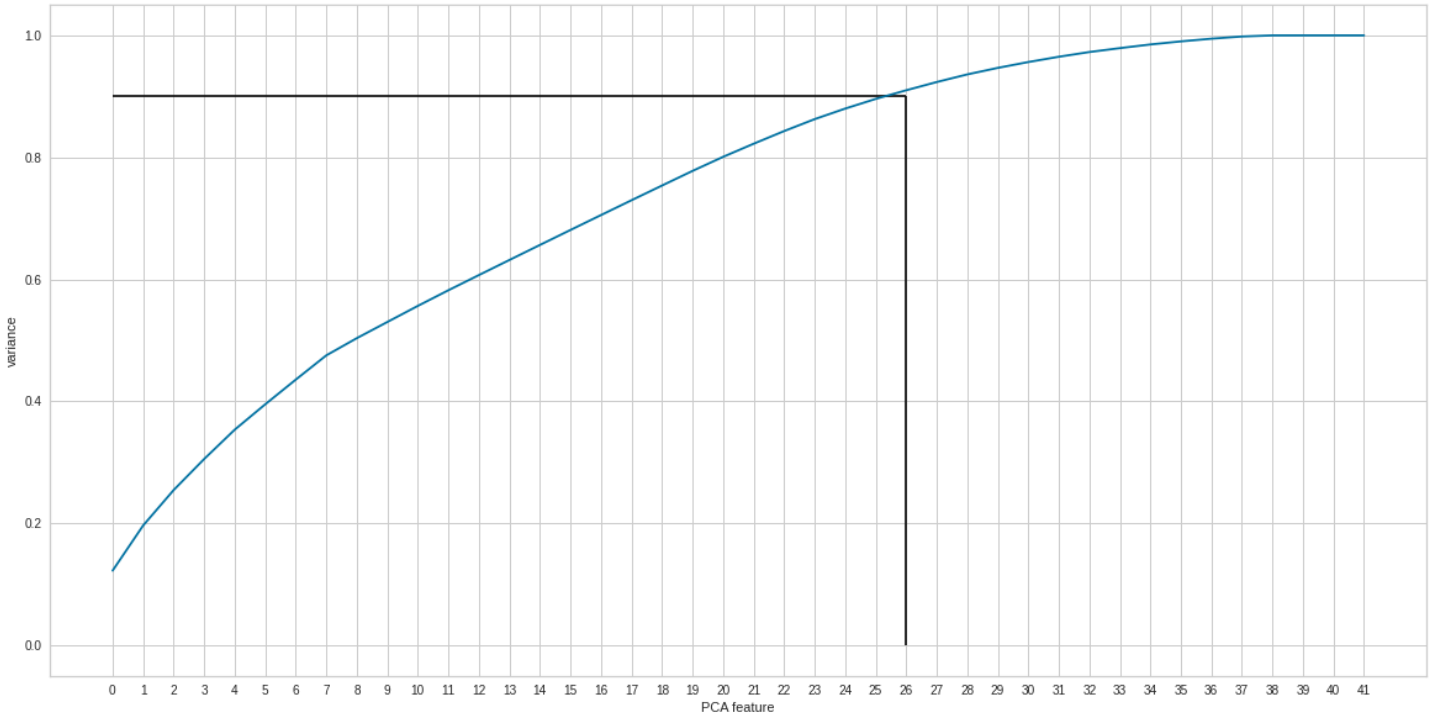
Our data preparation for the clustering model includes the following modifications to our data:

- Log transform features with high variability to remediate unbalanced importance in the clustering model.
- Simplify features such as "number of patio features" to boolean variables like "has patio features" to reduce model complexity.
- Encode categorical variables using the one-hot method to enable our models to consider category membership.
- Center and scale continuous variables to have a mean value of 0 and a standard deviation of 1 to balance importance by normalizing scale and range.

While we conducted principal component analysis (PCA) across our data, we ultimately determined that limiting the features listing agents provide would generate questions around feature importance and would limit future use. Our initial PCA work did produce interesting results and found that roughly 90% of the variability in our data can be explained with only half (26) of our features.

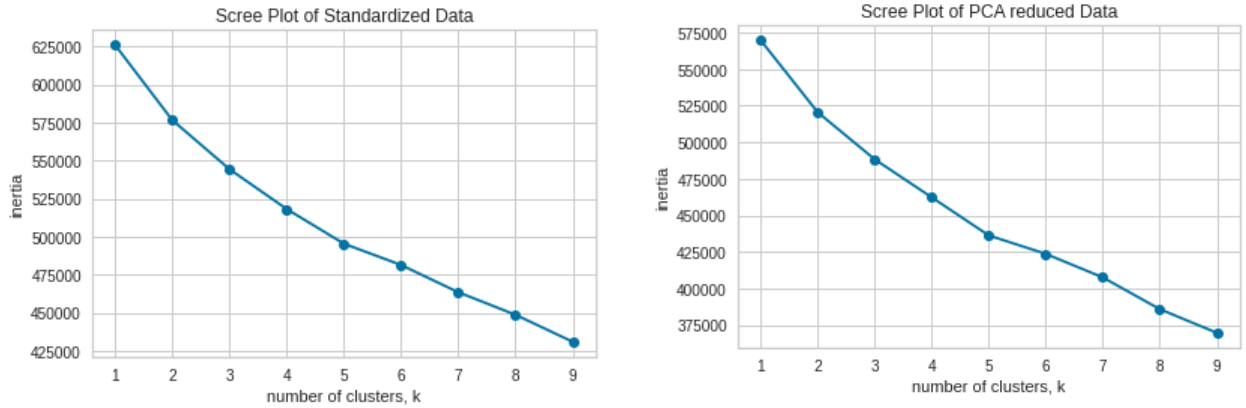
Figure 10: Principal Component Analysis





We employed both the Scree Plot (Elbow Method) and Silhouette Plot methods to determine the optimal number of segments against our original and PCA transformed datasets. As can be seen in Figure 11, the Scree Plots highlighted diminishing inertia, but were linear and thus did not highlight an optimal cluster number.

Figure 11: Scree Plots for Standardized and PCA reduced datasets



We used K-means clustering to segment homes based on nearest mean value, and hierarchical clustering to segment homes by features ranked over euclidean distance.

Evaluating the results of both clustering models, we can see that K-Means yields clearer clusters (Figure 12) as compared to the hierarchical method (Figure 13). We proceeded with K-Means as our methodology of choice.

Figure 12: K-Means Clustering

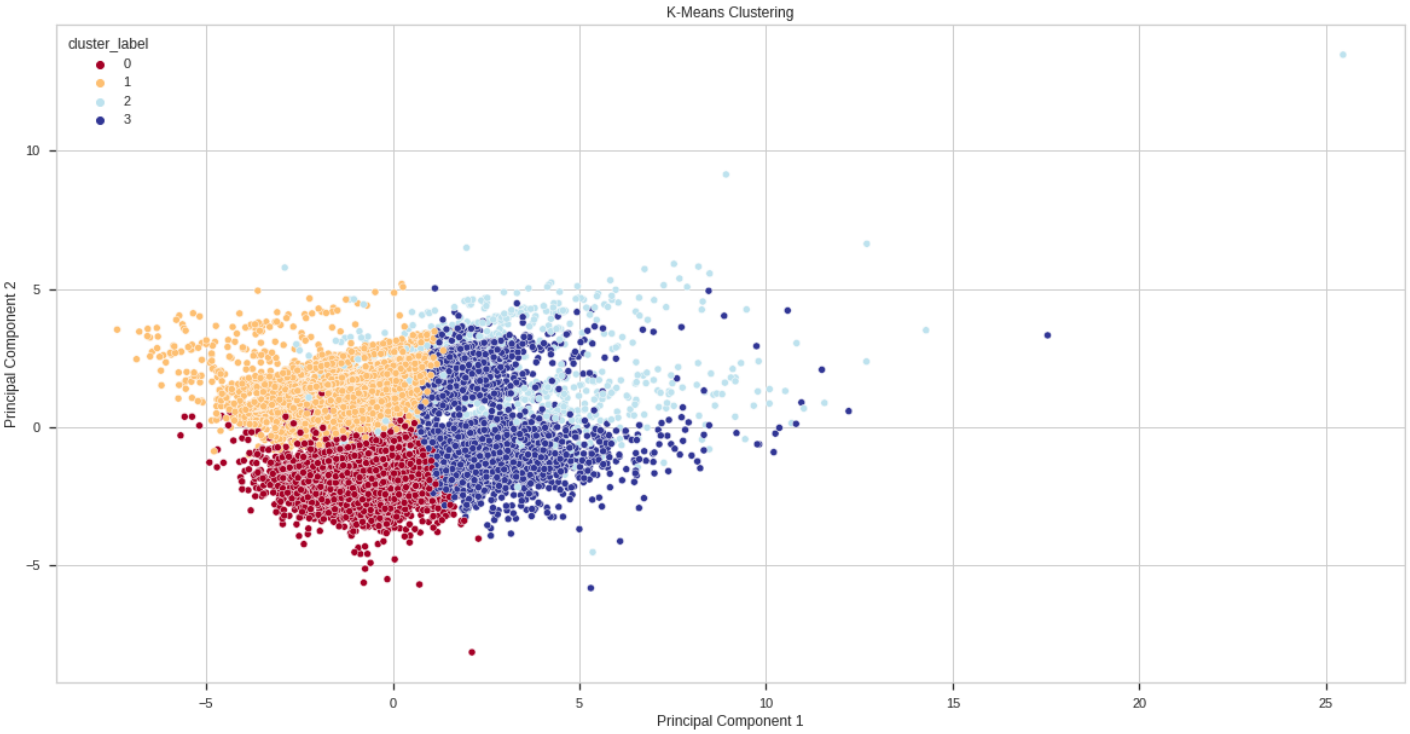
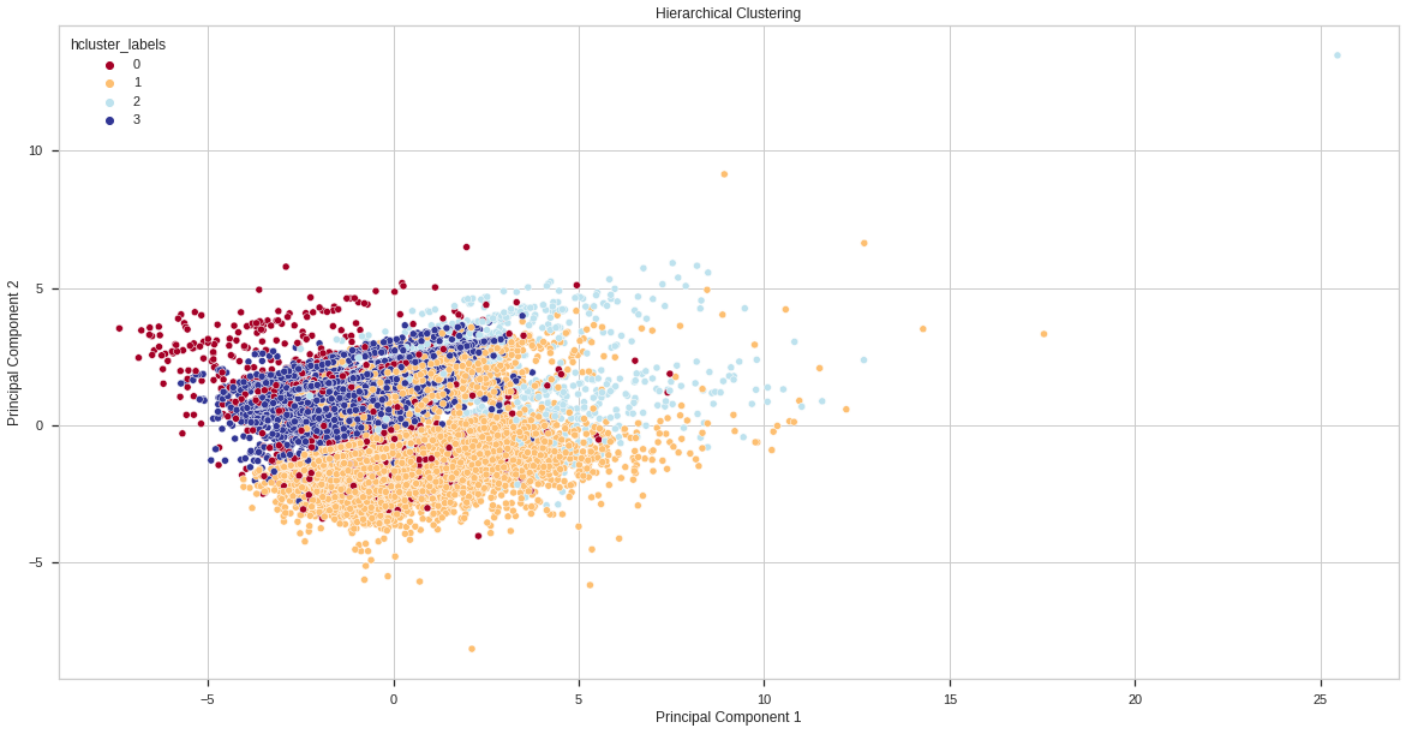


Figure 13: Hierarchical Clustering



Based on the defining characteristics of each cluster, we developed the following descriptions for each grouping.

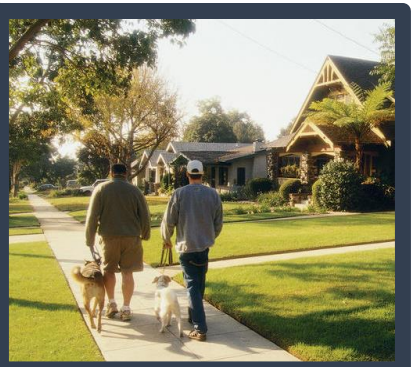
				
	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
Title	High Class Living	Family & Features	Metro Movers	Blue Collar Families
Characteristics	• Large lot sizes • Less Neighbors • $600k+ homes • $121k+ household income • Mostly college grads • 25% post grad • Families	• Large living area • Big lot sizes • $575k+ homes • Newer construction • Most have amenities • Large bed/bath count • Must be near all schools • Suburban • $94k+ household income • Mostly college grads • 25% post grad • Families	• High population • No associations • Public transportation • Value location • Mid-range price • Entry-mid salary • College students • New professionals	• High population • No associations • Value parking spaces • Mid-range price • Entry – mid salary • Some college
Cluster Size	5,000	>5,000	<1,000	~4,000

Table 6: Cluster descriptions

Clusters 0 (High Class Living) and 1 (Family & Features) had the greatest number of observations while cluster 2 (Metro Movers) represented the least.

As previously shown, center and west Austin tend to have higher priced homes than average, therefore it is not surprising that our High Class Living and Family clusters are located central and west while the Blue Collar cluster is situated farther east.

Figure 14: K-Means Observations

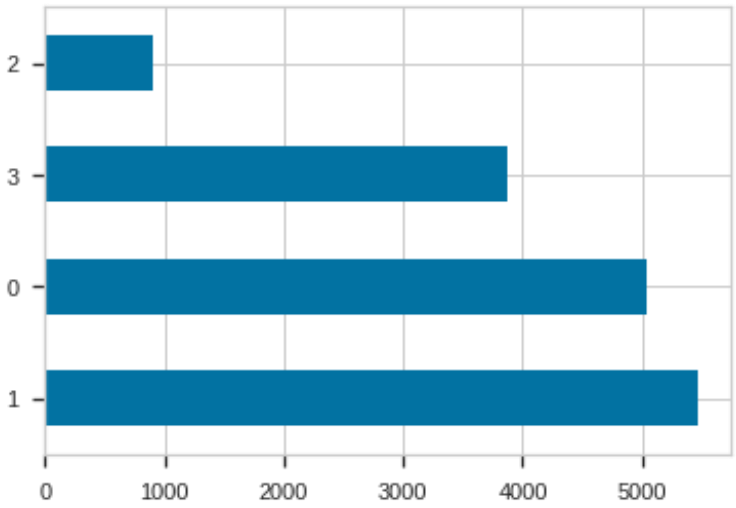
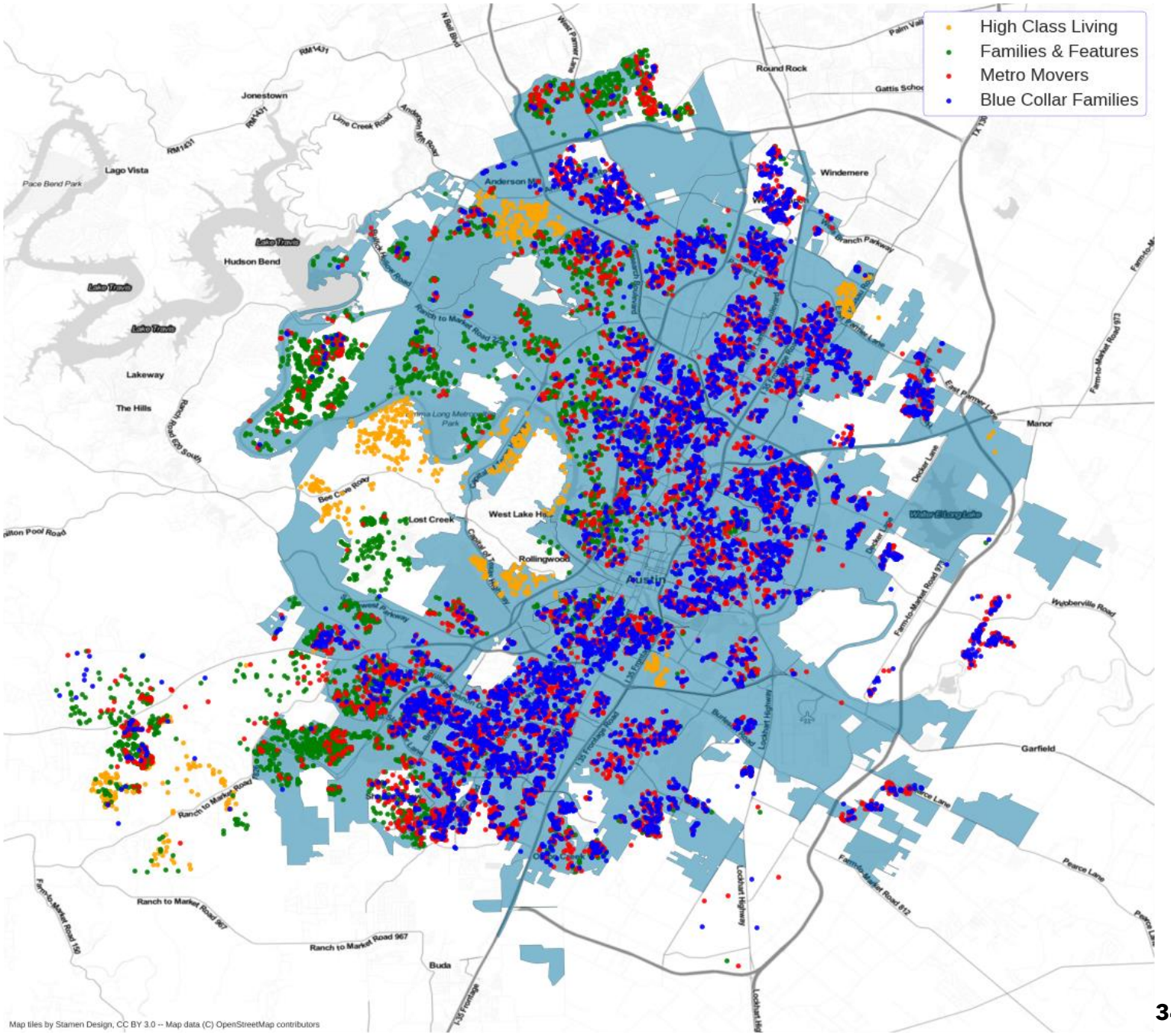


Figure 15: Locations of individuals by cluster



Price Prediction and Investment ROI Model

Our housing price prediction model will serve two purposes. First to assist agents to list the best possible selling price for a property based on its features, location, and amenities. This will ensure that the agent can give their customers a confident price prediction for their home that has a high likelihood of resulting in a conversion for the asking price. Second, we will develop an application to enable the agent and the customer to conduct what if scenarios to determine which investments are most impactful to their predicted selling price to maximize ROI.

As we are using regression techniques for this model, we included a correlation analysis to determine whether there are strong linear relationships between any two variables. This serves to identify variables with multicollinearity (where two or more variables are highly correlated with each other as opposed to the target variable), and isolate variables with independent explanation of variability. We created a correlation matrix and eliminated relationships that have over 80% correlation.

To combat and quantify overfitting (when models too closely align to data they were trained on but perform poorly on new data) we employed a 70/30 percent split for our training and testing data.

Our regression modeling approach leverages a model tournament system that executes several models and then ranks their performance to determine a champion model. The champion model will be used in the production deliverable for the client. We evaluated our models on both the full dataset, and the reduced dataset derived from Principal Component Analysis as noted above.

We evaluated the following modeling approaches in our model tournament:

- Linear Regression
- Linear Regression with PCA Reduction
- Lasso Regression with 5-fold Cross-Validation
- Lasso Regression with 5-fold Cross-Validation and PCA Reduction
- ElasticNet Regression with 5-fold Cross-Validation
- ElasticNet Regression with 5-fold Cross-Validation and PCA Reduction
- Random Forest Regression
- Gradient Boosting Regression
- Bagging Regression
- AdaBoost Regression
- Decision Tree Regression
- Ensemble Regressor (Random Forest, Gradient Boosting)
- XGBoost Regressor
- Tuned Random Forest Regressor
- Automated Machine Learning Regressor (AML)
- Deep Neural Net Regression

The Automated Machine Learning (AutoML) model is the clear victor in the tournament. This is primarily because the AutoML model conducts a secondary model tournament (Table 8) incorporating Gradient Boosting, XGBoost, and Stacked Ensemble models as well as additional protections against sample inaccuracies and leakage. AutoML also includes automated hyperparameter tuning, giving it a clear advantage over the manually tuned models.

Despite this, the tuned random forest and manual ensemble models came close in second and third place within less than a 1% difference amongst the three.

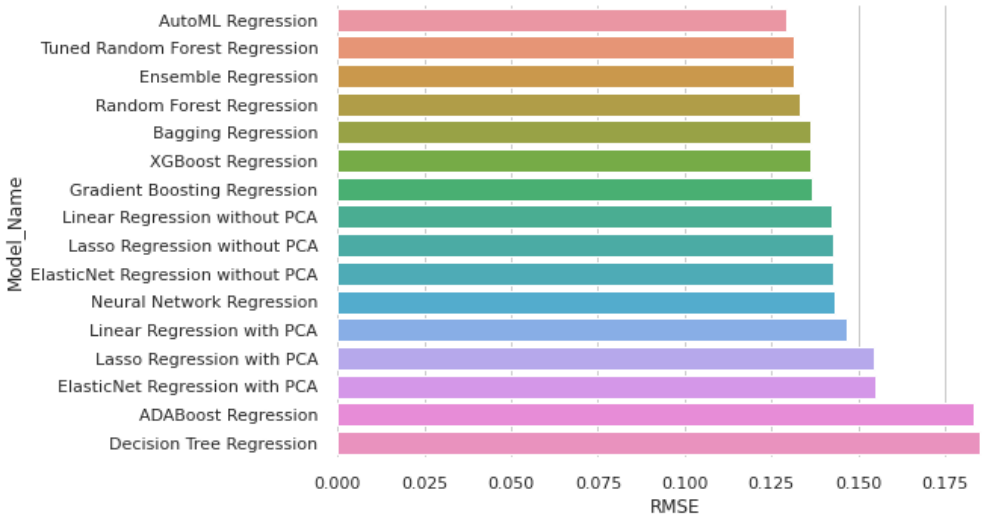
Table 7: Regression Models: Champion Tournament

Model_Name	MSE	RMSE	R2	MAE
AutoML Regression	0.02	0.13	0.71	0.07
Tuned Random Forest Regression	0.02	0.13	0.70	0.08
Ensemble Regression	0.02	0.13	0.70	0.08
Random Forest Regression	0.02	0.13	0.70	0.08
Bagging Regression	0.02	0.14	0.68	0.08
XGBoost Regression	0.02	0.14	0.68	0.08
Gradient Boosting Regression	0.02	0.14	0.68	0.08
Linear Regression without PCA	0.02	0.14	0.65	0.09
Lasso Regression without PCA	0.02	0.14	0.65	0.09
ElasticNet Regression without PCA	0.02	0.14	0.65	0.09
Neural Network Regression	0.02	0.14	0.65	0.09
Linear Regression with PCA	0.02	0.15	0.63	0.09
Lasso Regression with PCA	0.02	0.15	0.59	0.10
ElasticNet Regression with PCA	0.02	0.15	0.59	0.10
ADABoost Regression	0.03	0.18	0.42	0.13
Decision Tree Regression	0.03	0.19	0.41	0.11

Table 8: AutoML Internal Tournament Results

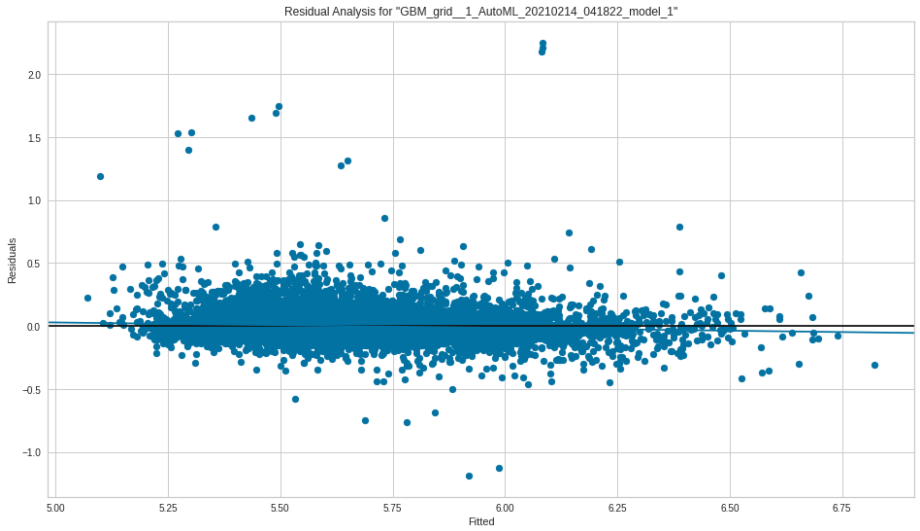
model_id	mean_residual_deviance	rmse	mse	mae	rmsle
GBM_grid__1_AutoML_20210214_041822_model_1	0.017	0.129	0.017	0.075	0.020
XGBoost_grid__1_AutoML_20210214_041822_model_1	0.017	0.130	0.017	0.076	0.020
XGBoost_grid__1_AutoML_20210214_041822_model_4	0.017	0.131	0.017	0.078	0.020
XGBoost_3_AutoML_20210214_041822	0.018	0.133	0.018	0.080	0.021
XGBoost_grid__1_AutoML_20210214_041822_model_2	0.018	0.134	0.018	0.080	0.021
XGBoost_1_AutoML_20210214_041822	0.018	0.135	0.018	0.085	0.021
XGBoost_2_AutoML_20210214_041822	0.018	0.135	0.018	0.082	0.021
GBM_grid__1_AutoML_20210214_041822_model_2	0.019	0.137	0.019	0.084	0.021
XGBoost_grid__1_AutoML_20210214_041822_model_3	0.019	0.137	0.019	0.083	0.021
GLM_1_AutoML_20210214_041822	0.021	0.145	0.021	0.093	0.022
DeepLearning_grid__1_AutoML_20210214_041822_model_1	0.021	0.146	0.021	0.095	0.023
GBM_4_AutoML_20210214_041822	0.023	0.152	0.023	0.100	0.023
GBM_3_AutoML_20210214_041822	0.023	0.152	0.023	0.100	0.023
GBM_2_AutoML_20210214_041822	0.024	0.155	0.024	0.103	0.024
GBM_1_AutoML_20210214_041822	0.025	0.157	0.025	0.106	0.024
DeepLearning_1_AutoML_20210214_041822	0.025	0.158	0.025	0.102	0.024
DRF_1_AutoML_20210214_041822	0.025	0.159	0.025	0.096	0.025
GBM_5_AutoML_20210214_041822	0.026	0.161	0.026	0.107	0.024
StackedEnsemble_AllModels_AutoML_20210214_041822	0.027	0.165	0.027	0.113	0.025
StackedEnsemble_BestOfFamily_AutoML_20210214_041822	0.027	0.166	0.027	0.113	0.025

Figure 16: Root Mean Squared Error by Model



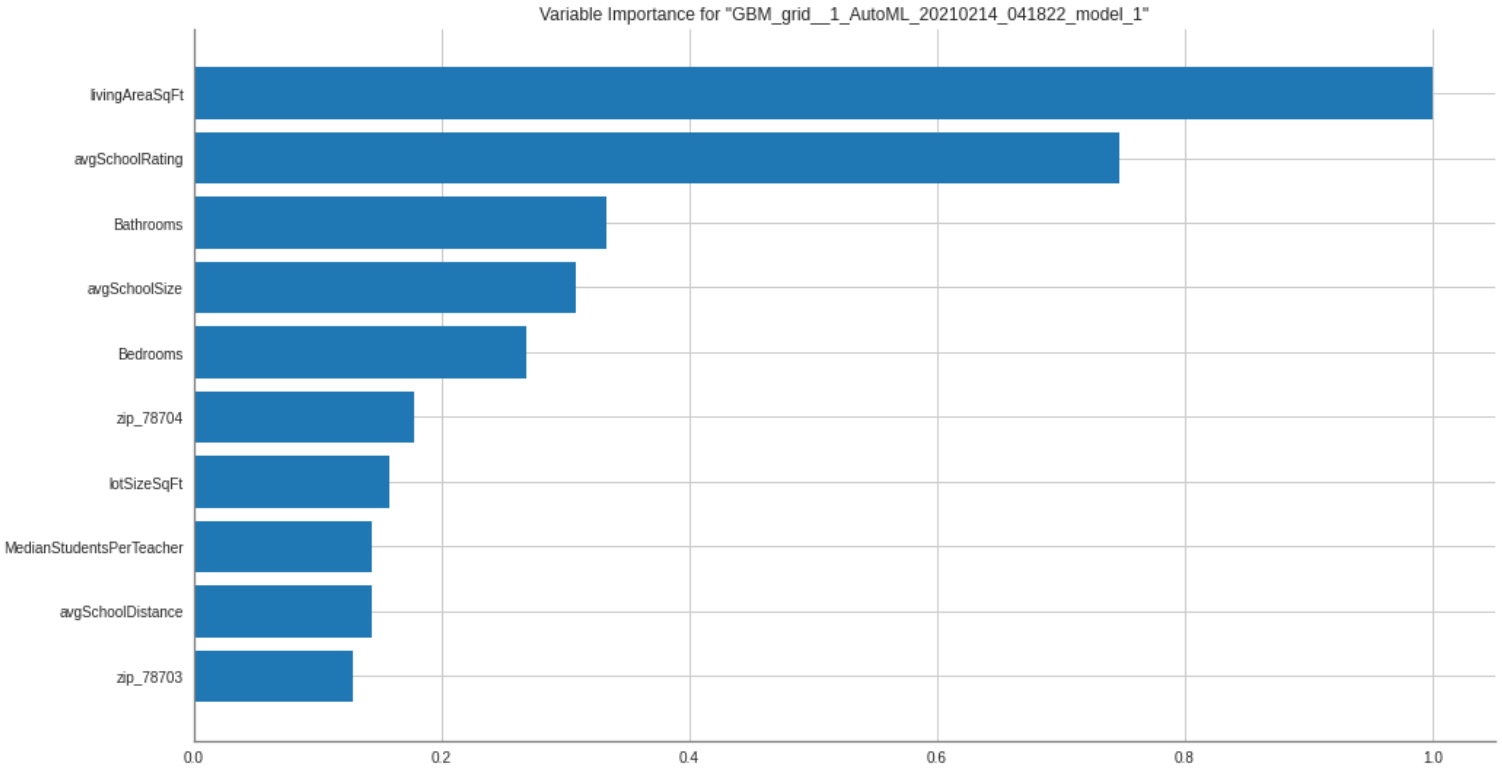
The residual plot for our champion model (Figure 17) shows a pattern of randomly distributed points, which is what we are looking for in terms of goodness of fit in a model.

Figure 17: Residual Plot for AutoML



The champion model determined that the most important variables in home price prediction are related to the size of the living area, the quality of the school system, school system size, the relative size of the home in terms of number of bedrooms and bathrooms, and a home's lot size. We also saw zip code and location data playing a significant role in ultimate house price.

Figure 18: Variable Importance



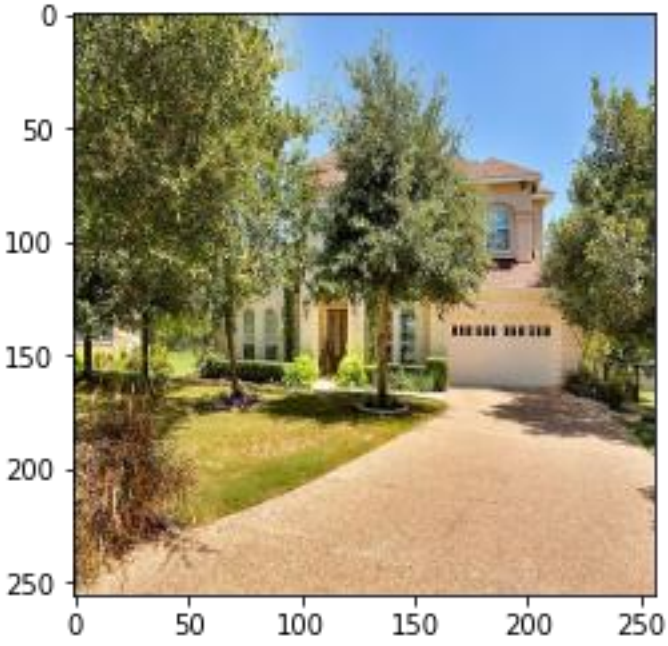
Computer Vision

The data we pulled from Zillow included links to the images uploaded to the posting. As image data can be highly valuable and potentially predictive, we leveraged computer vision modeling methodologies to determine how useful the image is in terms of price prediction.

As listings can have more than 150 images associated with them, we only examined the first image for a given listing. In order to prepare the images for use in a computer vision model we scaled them to 128x128 pixel squares, and retained the three color channels associated with them.

Before Scaling

After Scaling



For model architecture we built a Convolutional Neural Network, as can be seen to the right. This model uses a fully connected linear activation function, enabling us to use image data as input for a regression based deep neural network.

We use mean absolute percentage error as the loss evaluator, and trained the model over 25 epochs with a batch size of 8.

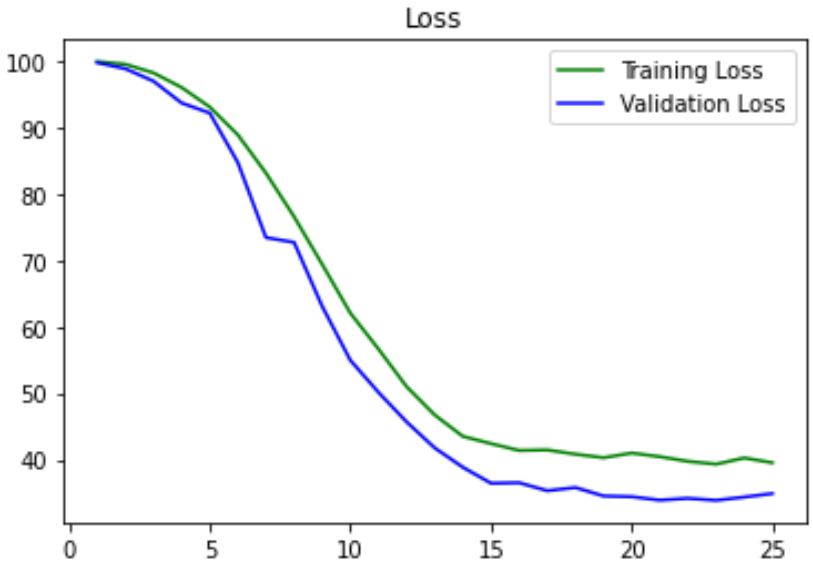


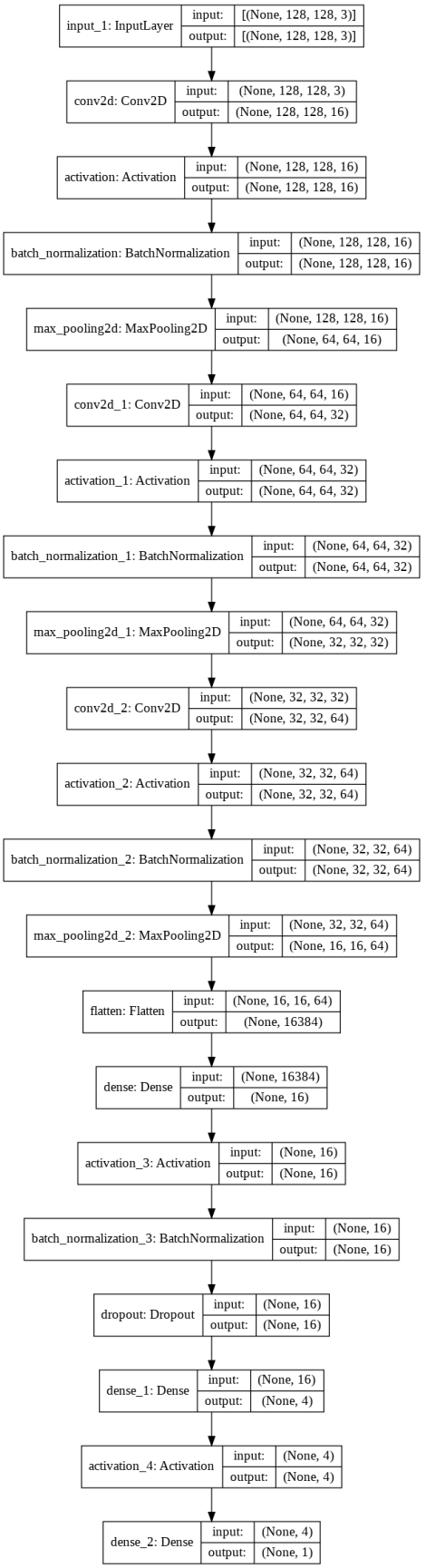
Figure 19: Loss Evaluation

Ultimately, our model had a mean absolute percentage error of 34.87%, indicating that by using the image data alone, on average our predicted price is around 35% off of the actual predicted price. While this surprisingly accurate for just image data alone, we felt it was not accurate enough for the standards and expectation of the Area51 team and our clients.

We believe that there are several reasons for our model results, including the limited amount of information that can be captured in a single image about a home listing, such as location, features, bedroom count, bathroom count, and so on. The only thing that the images will be able to base prices off of is "curb appeal", which is a rather subjective and non-specific thing to measure.

In the future the Area51 team may evaluate using image based pricing predictions as an input feature to our production models.

We also made the image data available on Kaggle here: **https://www.kaggle.com/ericpierce/austinhousingprices**



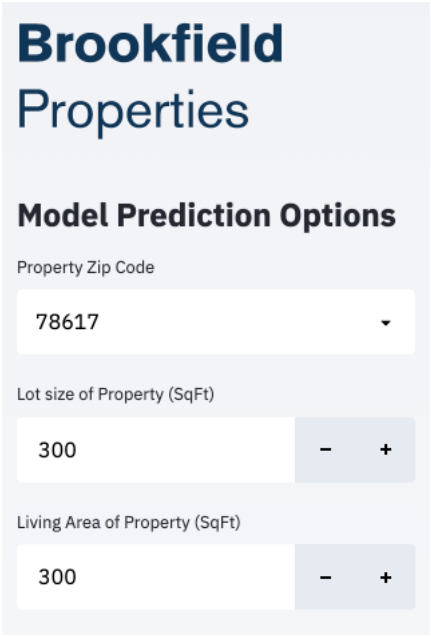
Web Application

Our Home Pricing Prediction web application is built on a responsive web framework that scales not only to all PC browsers, but also all mobile and tablet browsers. The application is custom hosted at the following url: **http://app.area51austin.com/** and can be accessed by our client and their customers alike at anytime from anywhere.

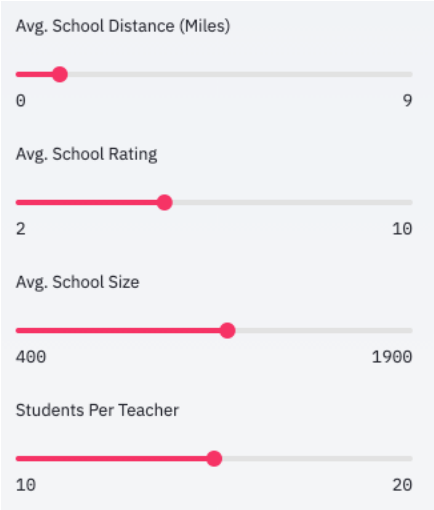
Navigation Panel

The left navigation panel of the application is where all the user inputs are entered.

The first input will be the zip code. We have made it easier and intuitive for the user to select from a drop down which only includes the zip codes for the Austin area. This allows us to add constraints and ensure that no errors are inputted into our models and that the user does not have a bad experience while using the tool. The next two inputs are the property lot size and the living area of the property. Again, we have made it simple by adding the minimum size for any of the properties in the Austin area so that the user knows not only the scale, but also that any number they select should likely be greater than the default selections.

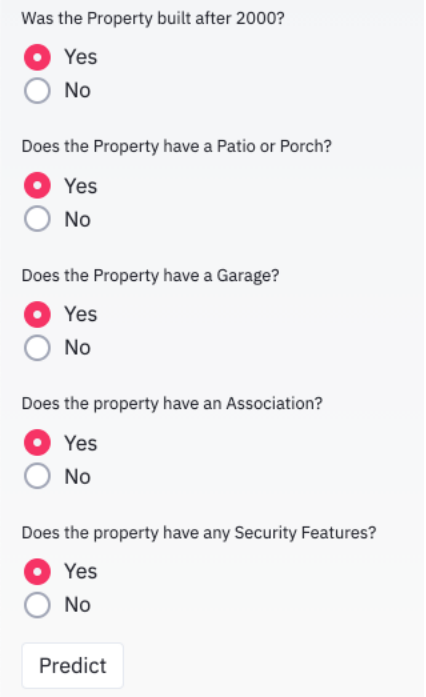


The next section asks for the Total number of bedrooms, bathrooms, and stories of the property. We made it very intuitive and easy to use by incorporating sliders for user selection. In addition, we set the max numbers based on the max number for the properties in Austin which will once again ensure data quality standards on user input and result in less confusion when utilizing the tool.



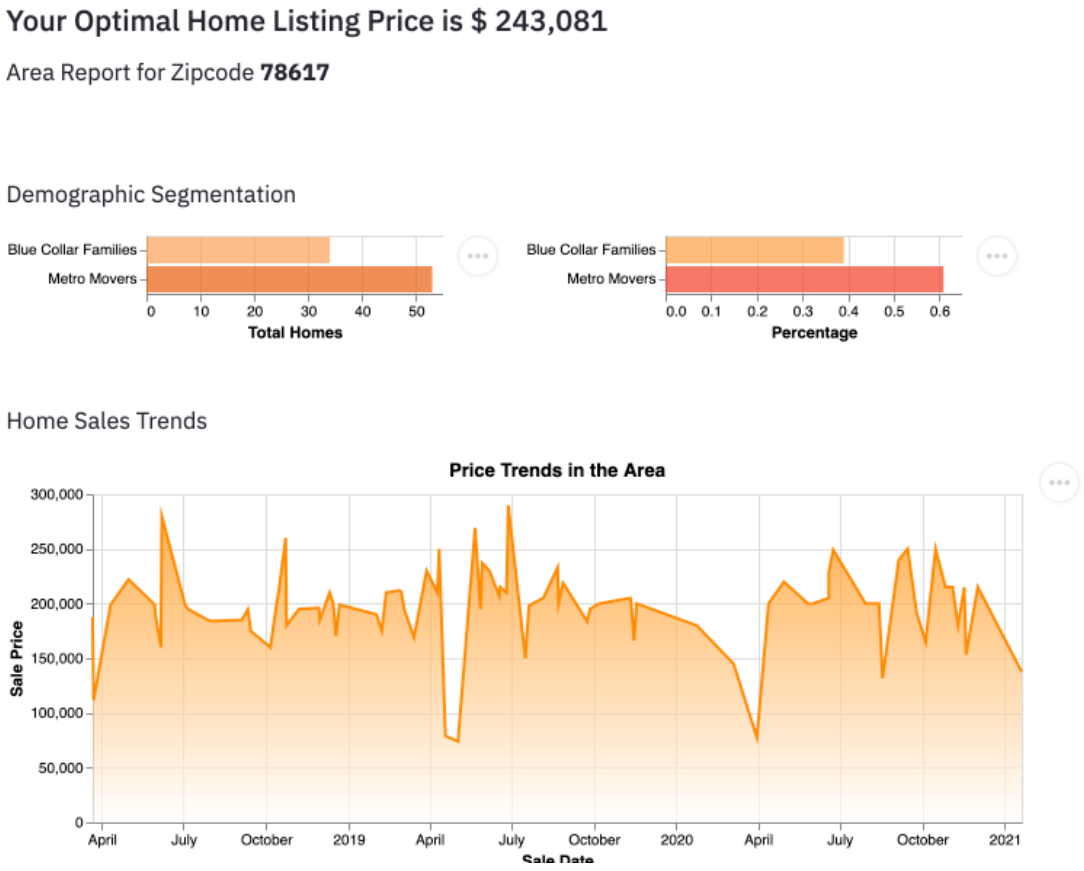
The school preference selection section allows the user to choose from a variety of options related to the Avg. School Distance in the area relative to the property, the Avg. School rating, the Avg. School Size, and the median number of students per teacher. We understand this is not something that most people will know off hand, so we made it easier for the users by setting the default values for all these fields to the mean values for the selected zip code in the Austin area.

The final section is all about features of the property. We ask the user some simple Yes and No questions about features of the property to help them assess the value of their property and the value of the property if they were to invest in these changes or have these features implemented on their property. From a design perspective, we have made it very simple for the users to simply click on a radio button and make their selection and we take the input on the back end and apply it to the pricing model. The final step in the process once all selections have been completed is to simply click on the "Predict" button and it will display the final prediction as well as a basic report of the selected Zip code.



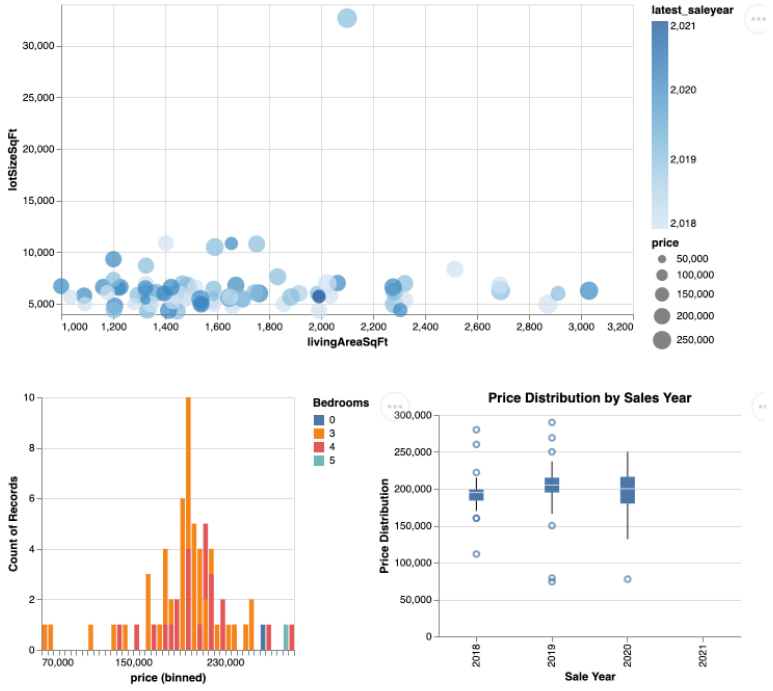
Prediction Report

The initial output for the prediction application will be an Optimal Home Listing Price. This is the optimal price at which the seller should consider listing the home. This is a result based on a model that considers not only location and size of the property, but also its features and the importance of the many factors that are part of the inputs. This result is based on thousands of listings spanning several years.



The next output is a segmentation of our population clusters. This will give the user a good indication of the type of customers or buyers that reside within the area and the density and distribution of each segment. This will allow the agent to better target their offers on the home to the audience that will have the greater propensity to respond to the advertisement and hopefully result in quicker conversions.

In addition, we also display a trend of sales prices over time for the area so that the agent and seller can have an indication of where their predicted price falls among the trending sales and give them an index of what the landscape looks like today and how it compares over time for the area.



We also provide the users with more comprehensive pricing comparison graphs to help them understand what the pricing landscape looks like for the area.

The scatter plot is based on the size of the living area, lot size, sales year, and sales price. It helps the user understand how the pricing is distributed based on the size of the property over time.

Further, we provide pricing distribution based on the number of bedrooms and finally boxplots that show the distribution of prices over the past few years so we can see how the housing market has changed in the area recently.

The last graphic in the prediction report is a map display of the density of properties within the selected zip code. This allows the seller to see how many other homes are within the area and the proximity.

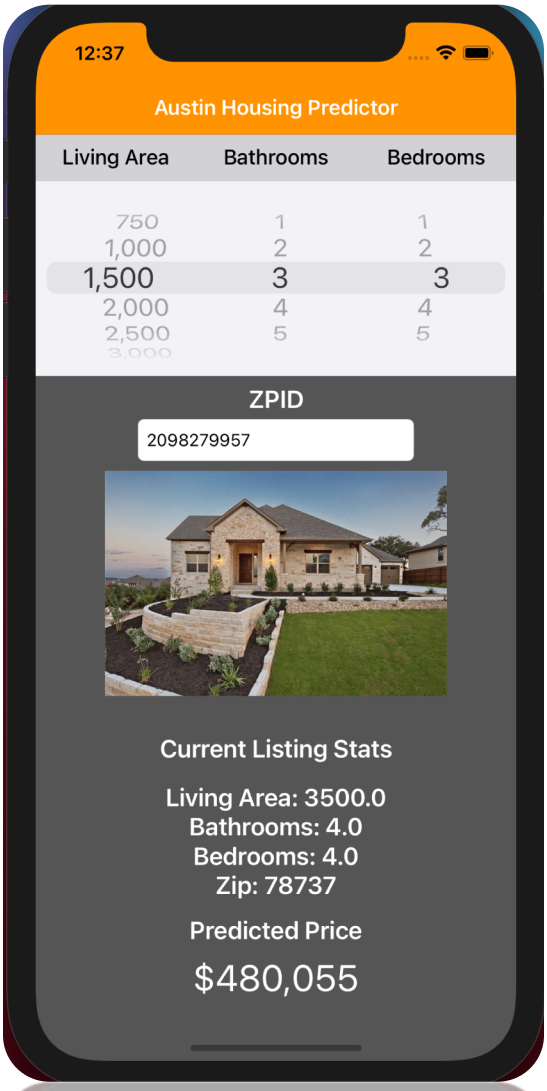
Mobile Application

The Austin Housing Field Agent is a simple mobile app to show home listers how their specific housing listing would be impacted by making updates to their property.

The Field Agent is a native iOS application, developed using Apple's XCode development environment. This app is differentiated from the web-based pricing prediction application as it enables users to identify what the impact of making changes to their existing house would be, using over 80 features from the actual housing listing on Zillow. A video demo of the app can be found here: **https://vimeo.com/518376293** and the source code for the app can be found here **https://github.com/eric-pierce/Austin-Housing-App**

Remodeling and Housing Update Price Projection

Unlike our web-app which recommends a price for a home given some input factors, the *Austin Housing Field Agent* app uses the actual listing for the customer's house and uses 86 different features as input to make pricing recommendations. The *Field Agent* allows customers to toggle through several different updates to make to their house, and see how they will impact their potential asking price.



Adjust housing features to create new predictions

Display features and predictions for only the desired property

Output predicted sale price based on new/updated features

Dashboard

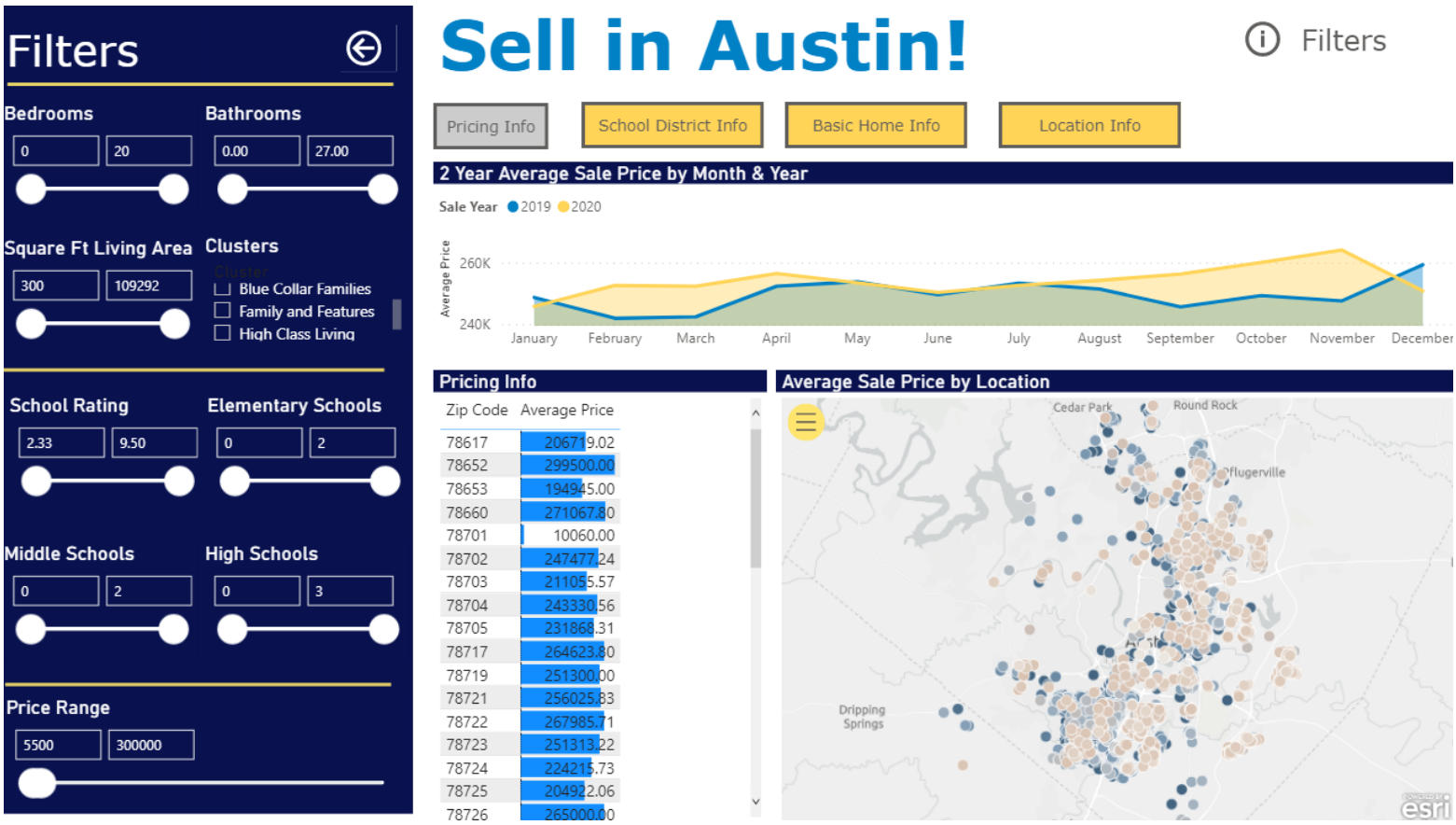
An analytical dashboard has been created to assist agents in understanding the segmentation of existing home sales in Austin.

This dashboard has four components:

- Pricing Information
- School District Information
- Basic Home Information
- Location Demographics

Before diving into any of the components, the user can select the Filters button to fine-tune the selection, narrowing down to only suitable zip codes. One of the possible filters is the cluster outputs from earlier. This allows the user to look deeper into each cluster by location, price, etc.

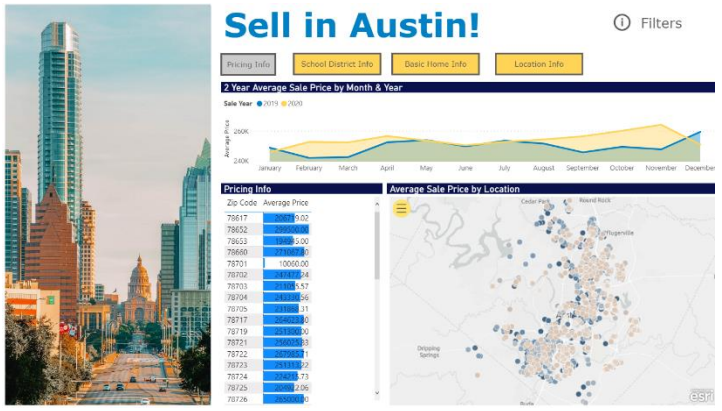
Once all the desired fields are selected it is possible to toggle between the four pages.



Pricing Information

The landing page of the dashboard that shows, at a high level, each zip code's average price, the price over time and the locations of each sale.

The map is fully interactive, allowing the user to select single sales or a collection in a desired area.

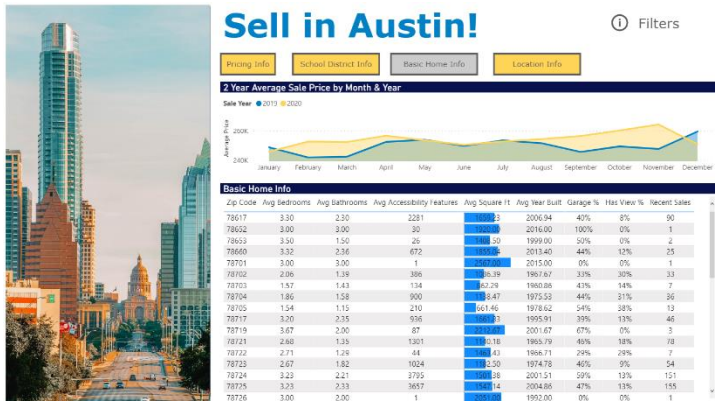


School District Information

Valuable information on the surrounding school district such as student teacher ratio, school sizes and distances are displayed on this tab in descending order. This allows the user to quickly identify which zip codes in their selection has the most desirable schools.

Basic Home Information

The average features of the homes in each zip code are displayed on this tab. Each zip code shows the average number of beds/baths, square footage, etc. for each house. Additionally, one can gather information on the area by utilizing the percentage of homes in the zip code that have a desirable view or have a garage.



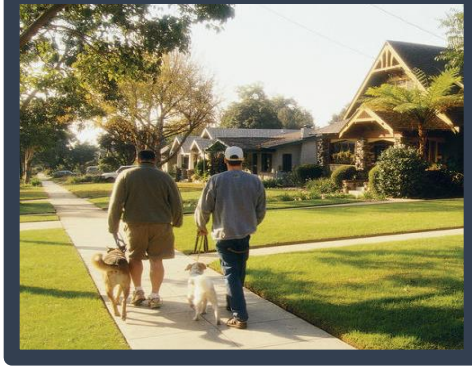
Demographics Information

Each zip code in Austin is unique. This tab highlights the uniqueness of each zip code by displaying demographic information. Information including degree holding percentage, poverty percentage, race distribution, age distribution and more.

Recommendations

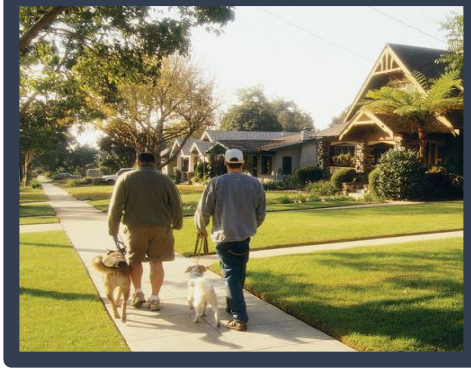
1.Leverage clusters in new properties development

Brookfield can use clustering information to expertly plan new development projects. Targeting properties to the right client demographic would enhance profits and improve ROI.

	Title	Property Development Plan	Rationale
	**High Class Living**	Choose isolated areas that have large lot sizes. Use top-of-the-line materials. Do not cut corners to save construction expenses.	These clients have high income and live with their families. They are willing to pay more to get the best possible property.
	**Family & Features**	Be in good school districts. Choose suburban locations with large lot sizes. Go for modern architecture style.	These are middle/upper-middle class families. They are modern and want their kids to be well-educated like them.
	**Metro Movers**	Ensure zero or very low association fee. Choose destinations that have multiple nearby public transports.	These people are educated but just started out their career. They have entry-level income and are careful with spending. They prefer to save money by avoiding associations and using public transports.
	**Blue Collar Families**	Ensure zero or very low association fee. Build large parking garages. Choose locations that are near free or cheap parking lots. Choose amenities that require low utility usage.	This customer group has lower income. They are extremely conscious spenders. They drive often but do not want to spend too much on parking spaces.

2. Apply advertisement strategies based on customer insights

Advertisement strategies come in all shapes and forms. We cannot rely on a magical one-size-fits-all campaign. Adjusting our strategies to target a desirable client base would reduce marketing costs and increase conversion rate.

	Title	Advertisement Strategies	Rationale
	**High Class Living**	Focus on safety and comfortability. Portray ourselves as trustworthy and luxurious at the same time.	These are high-earners who live with their families. They want their houses to be a safe haven that will provide an absolute peace of mind.
	**Family & Features**	Paint a positive future outlook. Highlights the modern, comfortable, and convenient amenities. Showcases family values through well-behaved children.	These clients have good income and do not mind spending on some shiny new things. They want a good future for their kids by providing them a quality education.
	**Metro Movers**	Showcase our convenient locations. Utilize viral marketing. Celebrity appearance could make a positive impact.	These people are young, educated, and trendy. They are conscious of spending.
	**Blue Collar Families**	Highlight the cost-efficiency of the properties. Create marketing events that have price discounts.	This customer group is very price sensitive. They are actively looking for new ways to save money.

3. Deploy Insights Dashboard to better understand and take advantage of the Austin housing market

The decision makers will be able to visualize data in a holistic manner. This is essential to understand hidden and influential trends in the market. From the information on the Insight Dashboard, here are what Brookfield can do:

- **Amend buying and selling schedule to get the best deal:** If a property needs to be sold quickly, a good sales strategy would aim for a Wednesday in July.

- **Plan future projects according to the effect of geographical location on housing price:** Price fluctuates significantly when moving from one zip code to the next. Some of the most expensive properties are located at: 78734, 78746, and 78703, while the cheapest zip codes are: 78617, 78742, and 78725.

- **Adjust investment strategies based on the geographical price distribution**: Properties in the north west are much more expensive than those in the east or south regions.

4. Utilize Web-Based Research Application to optimize new real estate projects

Our application enables real estate agents to leverage model-based pricing recommendations for properties based on key features as model inputs. This would prevent the sellers from mispricing their properties; it would also allow buyers to find good deals and to avoid overpaying. The two most important price predictors are: Living Area and School Rating. Neither of them are amendable once the foundation has been laid. Hence, it is critical for Brookfield to have a detailed plan before actioning any new projects.

5. Utilize Mobile Field Agent Application to maximize sale price through home alterations

The app enables the agent and the customer to conduct what if scenarios to determine which investments are most impactful to their predicted selling price to maximize ROI. According to the app, increasing the number of Bathrooms is a doable altercation that can make a significant positive impact on a property price.

Conclusions

The Austin housing market is hotter than ever. In this final report, we have provided an in-depth overview of our analytical process by following the best industry practices outlined by CRISP-DM.

Our data overview of the Austin Texas area found many homes to be single-family homes. Having a spacious home is a requirement for many single families in Austin. The average home in Austin is 2,200 square feet, outweighing the median single-family home size of 1,650 square feet. Homes in Austin sell well during the summer months of June and July and best on a Wednesday. The northwest area of Austin consists of expensive homes, whereas blue-collar families mainly live in the East side of Austin.

Data is prepared using a thorough data cleaning process.

Once data is cleaned, it is ready for modeling using clustering and regression analysis. We used K-Means clustering that find 4 clusters to describe customers in Austin: High Class Living, Families & Features, Metro Movers, and Blue-Collar Families. These clusters can drive better leads for real estate agents.

We compared 16 models to predict home price and found our best model to be the AutoML regression. This model found that the most important features for determining home price are living area, average school rating, and number of bedrooms and bathrooms. Some zip codes also lead to higher home prices.

We have provided an application which can predict the price of a new listing for agents that input home features. For existing listings, we have provided a separate mobile application which agents and customers can go through to improve the value of a home. Last, we have provided a dashboard that provides insight into the Austin area.

We are confident that we have provided all the deliverables. We have provided our analysis and analytical models. Our insights dashboard provides unique visualizations and informative insights into the Austin-Texas area. We have leveraged the models as a mobile application and web application.

Appendix

Data Dictionary

Variable	Variable Name	Data Type	Description
City	'address_city',	String	A lowercase name of a city or town in or surrounding Austin, Texas.
Street Address	'address_streetAddress',	String	A street address.
Zipcode	'address_zipcode',	Numeric	A 5-digit ZIP code.
Listing Description	'description',	String	A Zillow listing description of the property.
Latitude	'latitude',	Numeric	The latitude of the home.
Longitude	'longitude',	Numeric	The longitude of the home.
Property Tax Rate	'propertyTaxRate',	Numeric	The property tax rate for the home.
Garage spaces	'garageSpaces',	Numeric	The number of garage spaces a property comes with.
Association Membership	'hasAssociation',	Boolean	If the home is part of an association.
Has Air Conditioning	'hasCooling',	Boolean	If the home comes with air conditioning.
Has Garage	'hasGarage',	Boolean	If the home comes with a garage.
Has Heating	'hasHeating',	Boolean	If the home comes with heating.
Has Spa	'hasSpa',	Boolean	If the home comes with a spa.
Has View	'hasView',	Boolean	If the home comes with a view, determined by the property lister.
Home Type	'homeType',	String	The home type (i.e., Single Family, Townhouse, Apartment).
# of Parking Spots	'parking',	Numeric	The number of parking spots that come with a home.
Year built	'yearBuilt',	Numeric	The year the property was built.
Zillow Property ID	'zpid',	Numeric	A unique identifier to identify a property listing on Zillow.
Price	'best_price',	Numeric	The latest price the home is sold for.
# of Price Changes	'numPriceChanges',	Numeric	The number of price changes a home has underwent since being listed.
latest sale date time	'latest_saledatetime',	Datetime	The latest sale date time(YYYY-MM-DD 00:00:00).
latest sale date	'latest_saledate',	Datetime	The latest sale date (YYYY-MM-DD).
latest sale month	'latest_salemonth',	Numeric	The month the home sold (1-12).
latest sale year	'latest_saleyear',	Numeric	The year the property sold (2018-2021).
latest price source	'latest_price_source',	String	The party that provided the price estimate.
# of photos	'numOfPhotos',	Numeric	The number of photos in the Zillow listing.
# of accessibility features	'numOfAccessibilityFeatures',	Numeric	The number of unique accessibility features in the Zillow listing.
# of appliances	'numOfAppliances',	Numeric	The number of unique appliances in the Zillow listing.

# of parking feautres	'numOfParkingFeatures',	Numeric	The number of unique parking features in the Zillow listing.
# of patio and porch features	'numOfPatioAndPorchFeatures',	Numeric	The number of unique patio and/or porch features in the Zillow listing.
# of security features	'numOfSecurityFeatures',	Numeric	The number of unique security features in the Zillow listing.
# of waterfront features	'numOfWaterfrontFeatures',	Numeric	The number of unique waterfront features in the Zillow listing.
# of window features	'numOfWindowFeatures',	Numeric	The number of unique window aesthetics in the Zillow listing.
# of community features	'numOfCommunityFeatures',	Numeric	The number of unique community features (community meeting room, mailbox) in the Zillow listing.
Lot Size	'lotSizeSqFt',	Numeric	The lot size of the property reported in Square Feet.
Living Area	'livingAreaSqFt',	Numeric	The living area of the property reported in Square Feet.
# of Primary schools	'numOfPrimarySchools',	Numeric	The number of Primary schools listed in the area on the Zillow listing.
# of Elementary schools	'numOfElementarySchools',	Numeric	The number of Elementary schools listed in the area on the Zillow listing.
# of Middle schools	'numOfMiddleSchools',	Numeric	The number of Middle schools listed in the area on the Zillow listing.
# of High Schools	'numOfHighSchools',	Numeric	The number of High schools listed in the area on the Zillow listing.
Average school distance	'avgSchoolDistance',	Numeric	The average distance of all school types (i.e., Middle, High) in the Zillow listing.
Average school rating	'avgSchoolRating',	Numeric	The average school rating of all school types (i.e., Middle, High) in the Zillow listing.
Average school size	'avgSchoolSize',	Numeric	The average school size of all school types (i.e., Middle, High) in the Zillow listing.
Median students per teacher ratio	'MedianStudentsPerTeacher',	Numeric	The median students per teacher for all schools in the Zillow listing.
# of bathrooms	'numOfBathrooms',	Numeric	The number of bathrooms in a property.
# of bedrooms	'numOfBedrooms',	Numeric	The number of bedrooms in a property.
# of stories	‘numOfStories’	Numeric	The number of stories a property has

References

Works Cited

n.d. *Best High Schools in the Austin, TX Area.* https://www.usnews.com/education/best-high-schools/texas/rankings/austin-tx-12420.

n.d. *Council Districts Fill: Open DATA: City of Austin Texas.* https://data.austintexas.gov/Locations-and-Maps/Council-Districts-Fill/hdpc-ysmz.

n.d. *Deployment to Core ML.* https://apple.github.io/turicreate/docs/userguide/supervised-learning/export-coreml.html.

Pandey, Harshita. 2020. *House Price Prediction (Regression) with Tensorflow - Keras.* May. https://medium.com/analytics-vidhya/house-price-prediction-regression-with-tensorflow-keras-4f.

Rosebrock, Adrian. 2020. *Keras, Regression, and CNNs.* June. https://www.pyimagesearch.com/2019/01/28/keras-regression-and-cnns/.