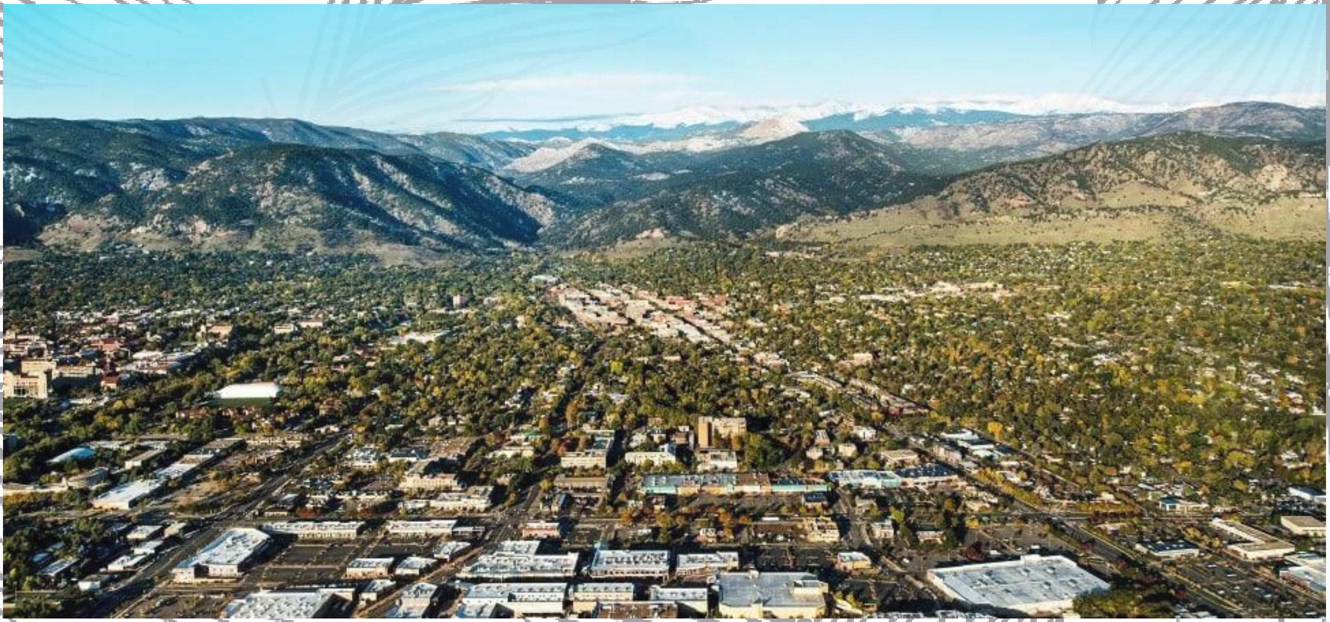


Analysis of Residential Real Estate Market and home price prediction in Boulder County, Colorado

Benazir Rowe

Table of Contents

Problem

Our goal is to identify trends and see what factors contribute to real estate prices in Boulder County, Colorado. We are focusing on home's intrinsic characteristics: physical characteristics and location. We are not using macroeconomic factors to limit the scope of our analysis and we do not incorporate school quality data, since there is an open enrollment policy, and a child is eligible to enroll in any school in the State of Colorado and a child often attends the school that fits their needs better and not the one they are zoned for.

About Boulder County, Colorado

Boulder County, Colorado is known for its vibrant economy and is considered one of the most prosperous regions in the state. Here are some economic trends in Boulder County:

1. **Strong job growth:** Boulder County has experienced steady job growth over the past decade, with an unemployment rate below the national average. The county is home to many innovative companies in industries such as technology, renewable energy, and healthcare, which have contributed to the region's strong job market.
2. **High median household income:** The median household income in Boulder County is above the national average, which is driven in part by the high-paying jobs available in the area. This has led to a high standard of living and a strong local economy.
3. **Growing population:** Boulder County has experienced steady population growth over the past few decades, which has helped to fuel the local economy. The county's population is well-educated and highly skilled, which has attracted many businesses to the area.
4. **Booming real estate market:** Boulder County's real estate market has been strong in recent years, with home values increasing faster than the national average. The region's desirability, coupled with limited housing inventory, has driven up home prices and made it a seller's market.
5. **Emphasis on sustainability and innovation:** Boulder County is known for its focus on sustainability and innovation, with many companies and organizations in the area focused on renewable energy, green technology, and sustainable practices. This emphasis on sustainability and innovation has helped to drive economic growth and attract new businesses to the area.

These factors make Boulder County Colorado and attractive destination, and residential real estate is an important driver of the Boulder County economy. We are focused on wealth creation: rising home values can contribute to wealth creation for homeowners. So, making an informed decision when buying or selling real estate in the market is very important. Our analysis aims to help with that.

Trends in Boulder County residential real estate sale prices

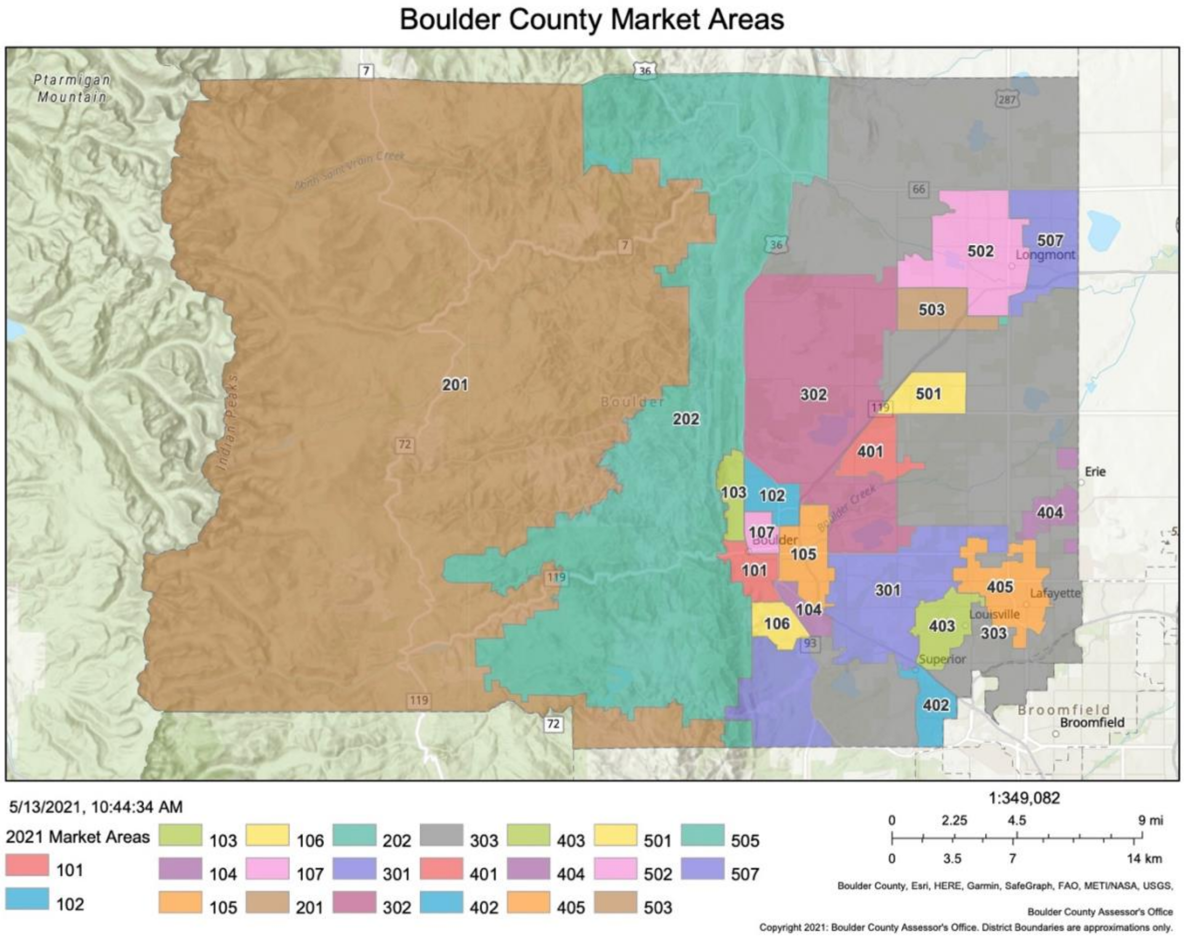
There are some notable trends in Boulder County's residential real estate sale prices. Here are a few:

1. **Overall price appreciation:** Over the past decade, residential real estate sale prices in Boulder County have appreciated significantly. According to data from the Boulder Area Realtor Association, the median sale price for a single-family home in Boulder County was $410,000 in 2010, and had risen to $730,000 by 2020, representing an increase of approximately 78%. This trend is in line with national trends of rising home prices.
2. **Limited inventory:** Boulder County's real estate market is characterized by limited inventory, which can contribute to rising prices. According to the Colorado Association of Realtors, there were only 317 active listings for single-family homes in Boulder County as of February 2021, compared to 1,219 in Denver County, which has a similar population. Limited inventory can create a seller's market, where buyers are willing to pay more for homes due to the scarcity of available properties.
3. **Location matters:** As with most real estate markets, location is a significant factor in determining sale prices in Boulder County. Properties in desirable neighborhoods, such as Boulder's downtown area, the foothills, or near the University of Colorado Boulder, tend to command higher prices than properties in less desirable areas. Properties with views of the mountains or open space are also highly valued.
4. **Luxury market growth:** Boulder County's luxury home market has seen significant growth in recent years, with sales of homes priced over $2 million increasing by 60% in 2020 compared to the previous year. The pandemic has contributed to this trend, as many high-income earners seek out larger homes with more amenities and outdoor space.

Overall, Boulder County's residential real estate market has seen significant price appreciation in recent years, driven in part by limited inventory and strong demand for homes in desirable locations.

Location matters: Market Area Maps for Single Family Properties

Comparable properties are often found in the same Market Area, which are groups of neighborhoods that have similar trends in the market. Boulder County uses these areas to establish market value for mass appraisal purposes. Since we established that location plays an important role in the price of the property, we use market area as a more refined proxy of location and expect it to be important in our analysis.



Below are the Market Areas

1. Single Family Residential Properties

Sales of single-family homes, duplexes, and triplexes.

Boulder

- Market Area 101 – Boulder
- Market Area 102 – Boulder
- Market Area 103 – Boulder
- Market Area 104 – Boulder
- Market Area 105 – Boulder
- Market Area 106 – Boulder
- Market Area 107 – Boulder

Countywide area

- Market Area 201
- Market Area 202
- Market Area 301
- Market Area 302
- Market Area 303
- Market Area 304

Erie Lafayette Louisville Superior Gunbarrel

- Market Area 401– Gunbarrel
- Market Area 402 – Superior
- Market Area 403 – Louisville
- Market Area 404 – Erie
- Market Area 405 – Lafayette

Longmont &Niwot

- Market Area 501 – Niwot
- Market Area 502 – Longmont
- Market Area 503 – Longmont
- Market Area 505 – Longmont
- Market Area 507 – Longmont

2. Residential townhomes

- Market Area 108 – Boulder, East of Broadway and All Other Townhomes
- Market Area 109 – Boulder, West of Broadway and Downtown Townhomes
- Market Area 407 – Erie, Gunbarrel, Lafayette, Louisville, Niwot, & Superior Townhomes
- Market Area 506 – Longmont Townhomes

3. Residential condominiums

- Market Area 630 – Boulder Condos
- Market Area 632 – Erie, Gunbarrel, Lafayette, Louisville, Niwot, & Superior Condos
- Market Area 633 – Longmont Condos
- Market Area 634 – Mountain Condos

Client

There are many stakeholders in the residential real estate market, including:

1. **Homeowners:** Homeowners are the most obvious stakeholders in the residential real estate market, as they own the properties that are being bought and sold. Homeowners have a financial interest in the market, as the value of their homes can impact their net worth and ability to access credit. Homeowners also have a personal interest in the market, as the quality and location of their homes can impact their quality of life.
2. Buyers: Buyers are another key stakeholder in the residential real estate market. They are looking to purchase properties that meet their needs and desires at a price they can afford. Buyers may also have a financial interest in the market, as they may be taking out loans or investing their own funds in a property.
3. **Sellers:** Sellers are another important stakeholder in the market, as they are looking to sell their properties for a price that meets their needs. Sellers may be motivated by a desire to downsize, move to a different location, or capitalize on market trends.
4. **Real estate agents:** Real estate agents are professionals who help buyers and sellers navigate the market. They have a financial interest in the market, as they earn commissions on the properties they help buy and sell.
5. **Lenders:** Lenders are financial institutions that provide loans to buyers. They have a financial interest in the market, as the value of the properties being bought and sold can impact their risk exposure and ability to generate profits.
6. **Developers:** Developers are stakeholders in the market who build new properties, often with the goal of selling them for a profit. They have a financial interest in the market, as the value of the properties they build can impact their profits and ability to secure financing for future projects.
7. **Local governments and policymakers.** Local governments have an interest in the market as it can impact the local economy, tax revenues, and quality of life for residents. Local governments may be involved in zoning and land use regulations that can impact the supply and demand for housing, as well as the affordability of housing. They may also provide incentives or programs to encourage affordable housing development, or work to ensure that new development is consistent with community plans and goals.

Data

We are looking at the current 2021/2022 tax years: residential sales of Boulder County, Colorado. Base Period Sales from July 1, 2018–June 30, 2020. In areas with few sales, the assessor is allowed to use comparable sales going back up to five years.[1]

There is a separate dataset for each market area, totaling 31 datasets. Each dataset contained the following fields:

Account Number
Property Type
Property Address
Street Number
Property Address
Street Dir
Property Address Street Name
Property Address Street Suffix
Property Address Unit Number
Location
Design
Quality
Eff Yr Built
Above Grd SF
Basemt Tot SF
Basemt Fin SF
Basemt Unf SF
Garage Type
Garage SF
Est Land SF
Reception No
Sale Date (Mon-Yr)
Sale Price
Time Adjust Sales Price
Market Area

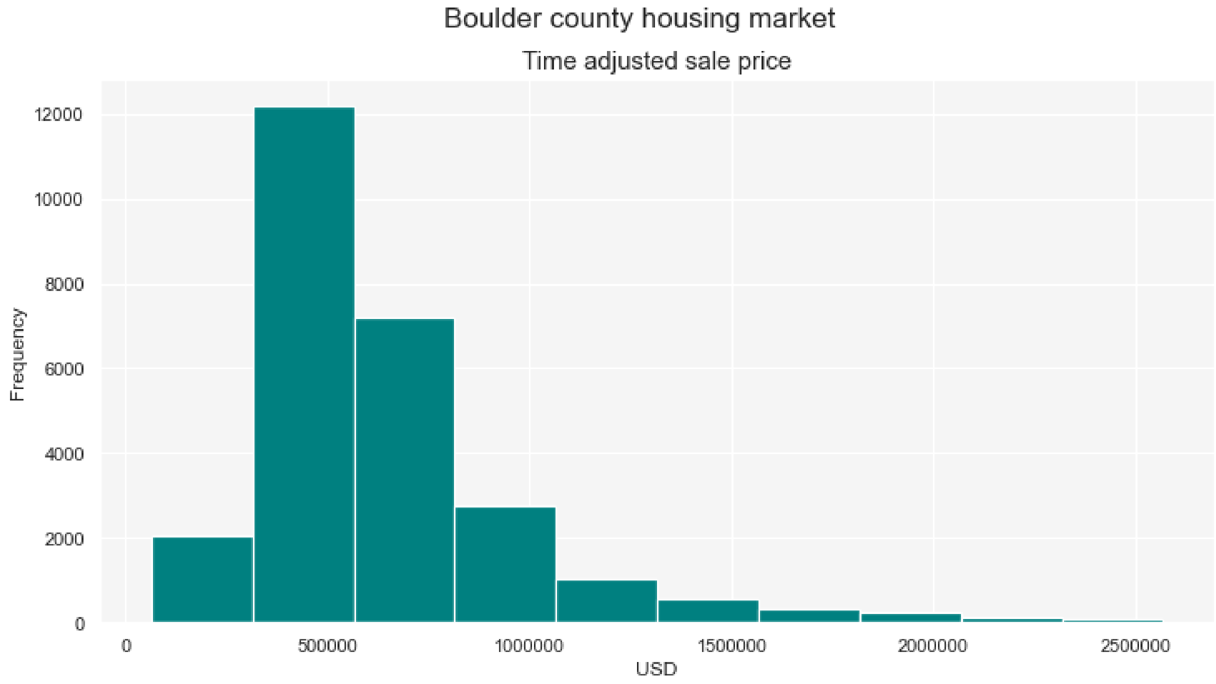
[1] https://bouldercounty.gov/property-and-land/assessor/sales/comps-2021/residential/

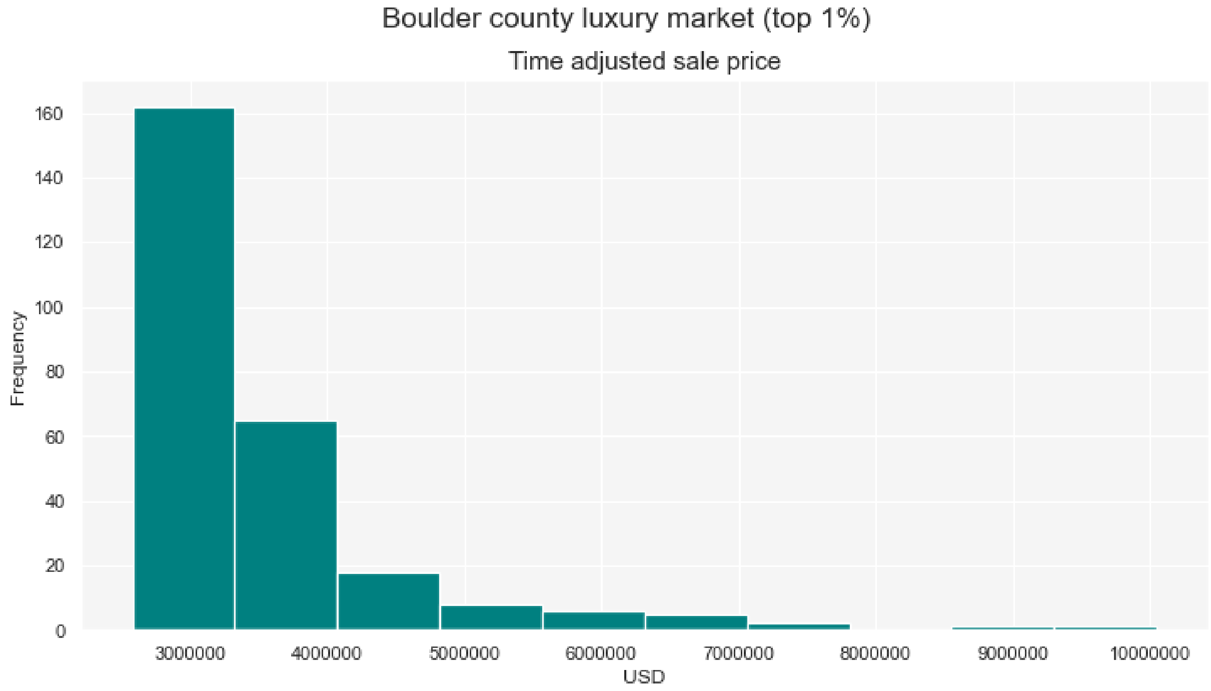
Exploratory Data Analysis

Univariate analysis

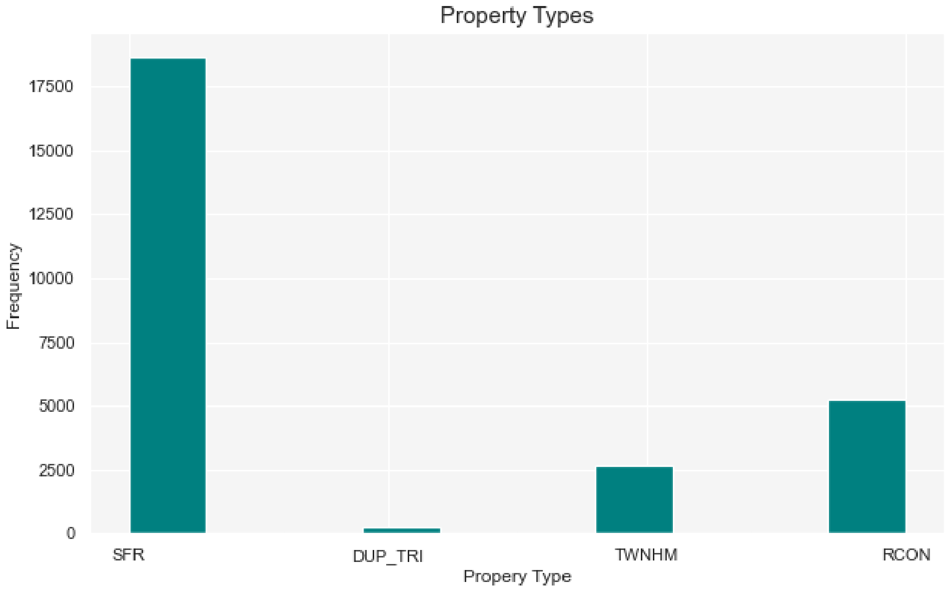
Sale price.

We separated 99% of the houses from top 1% to see the differences better.



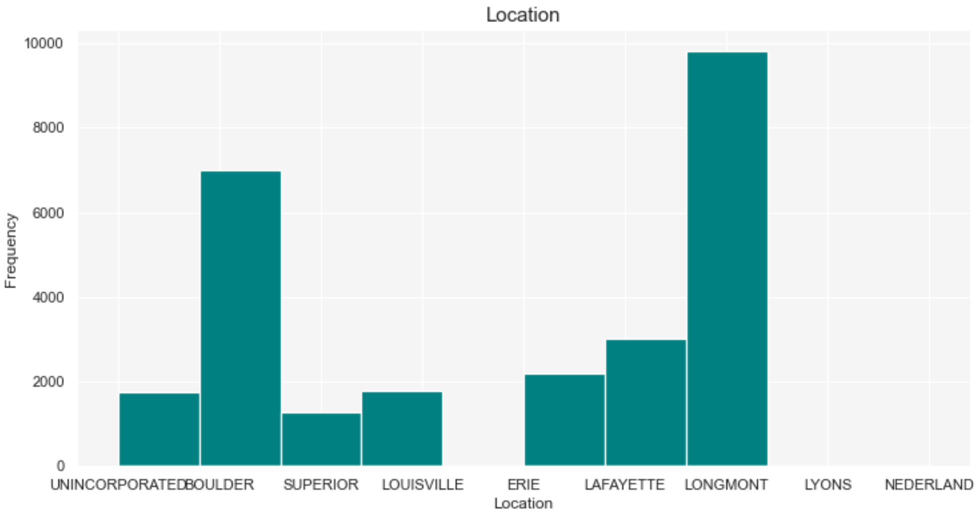


Property type.



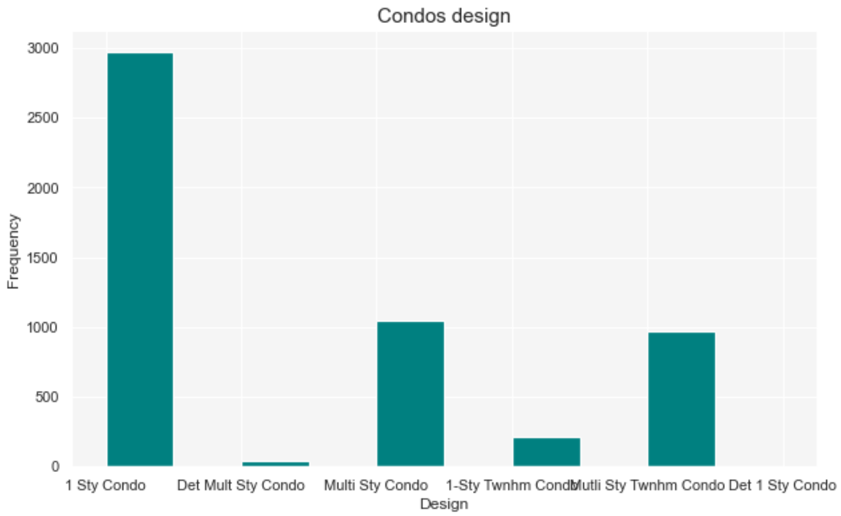
The majority of houses sold on the market are single family houses, followed by condos and townhomes and duplexes/triplexes.

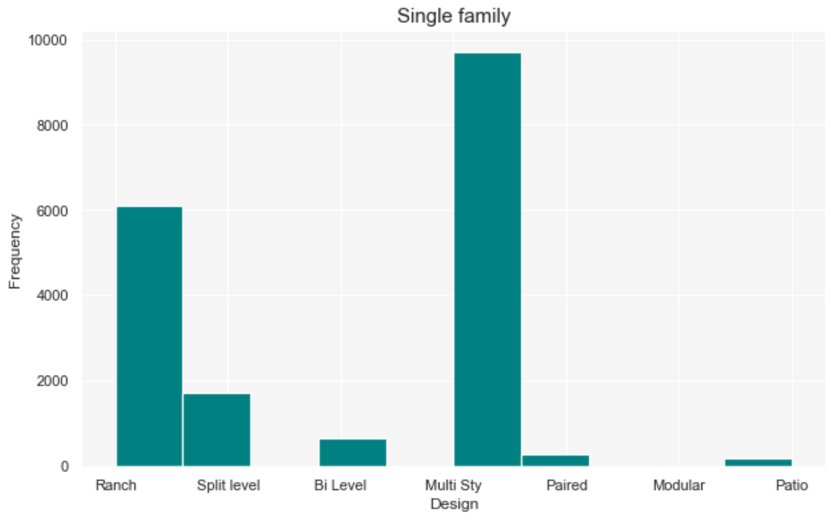
Location.

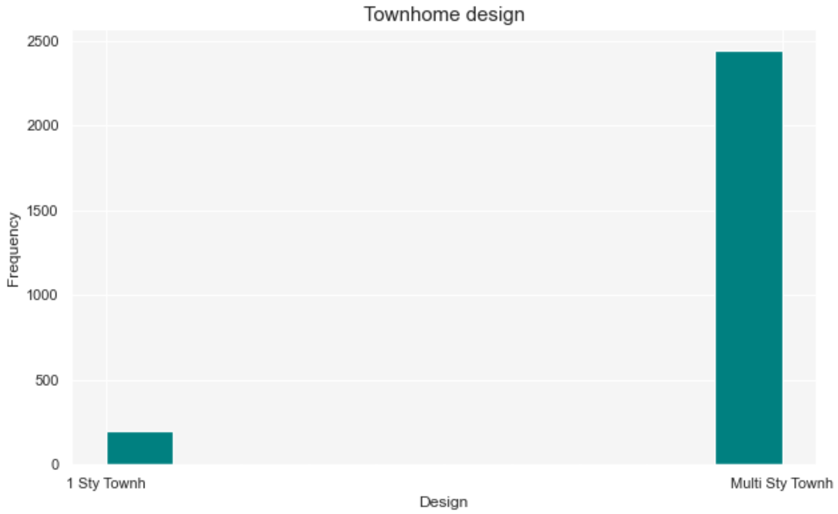


The biggest municipalities are Longmont, followed by Boulder, Lafayette, Erie, Louisville, Unincorporated, Superior, with very few in Lyons and Nederland.

Design.



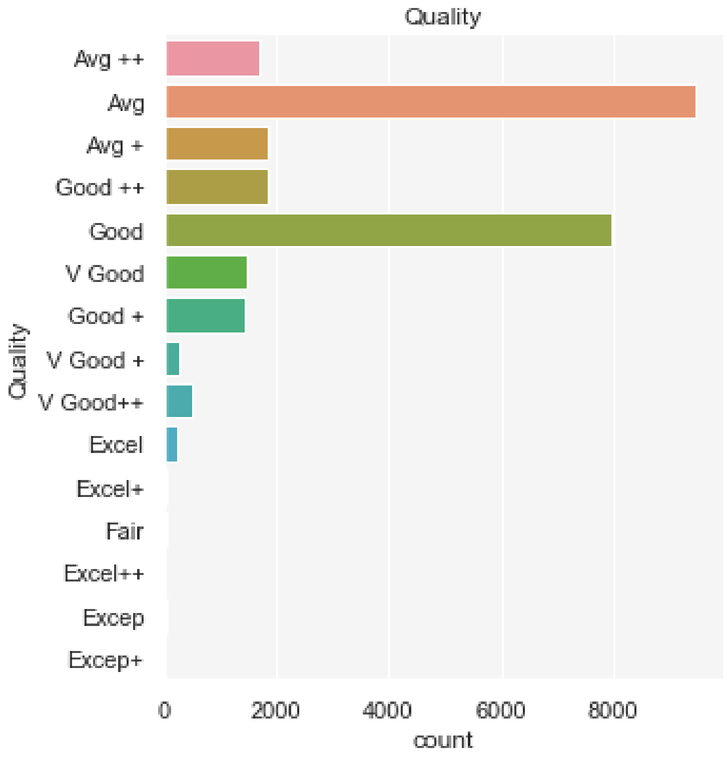




Most common design for condos is a one-story condo, for single family and townhomes multi-story. We combined condo designs into a single design ‘condo’ to limit the impact of additional variables.

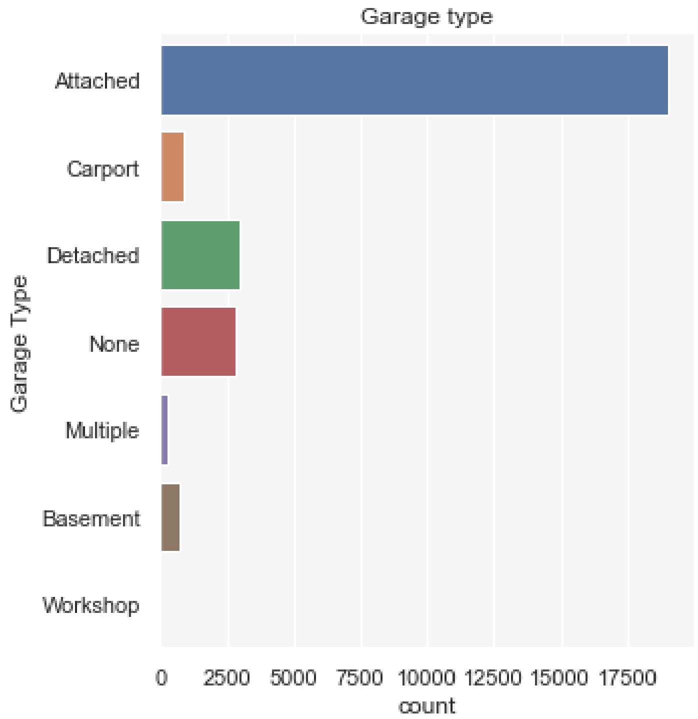
There are various condo designs, however since the proportion of condos is small in the total market volume, we will merge all condo designs as ‘condo’.

Quality.



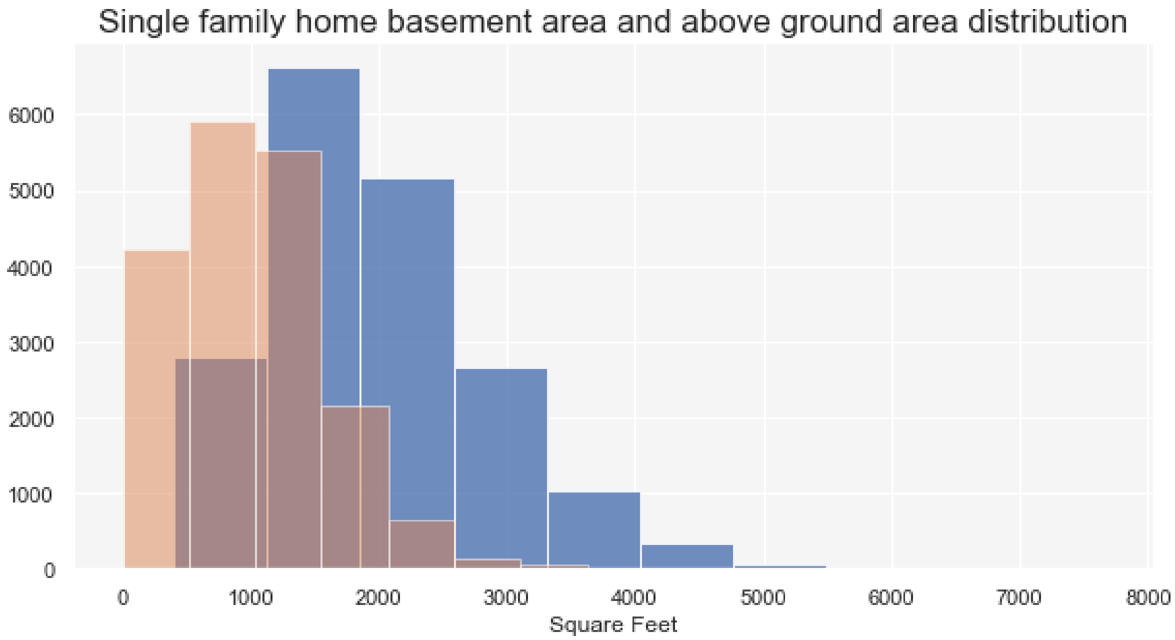
Most houses are rated Average and Good, with a few given + or ++ rating with almost no houses rated at V good or Excellent or Exceptional. Given this distribution, the value of such detailed scale is dubious, especially Exceptional and Very Good categories.

Garage Type.



Most houses have attached garage. We decided to binarize the ‘Garage type’ variable into attached vs everything else in order to avoid creating too many indicator variables.

Above ground and basement square footage (SF).

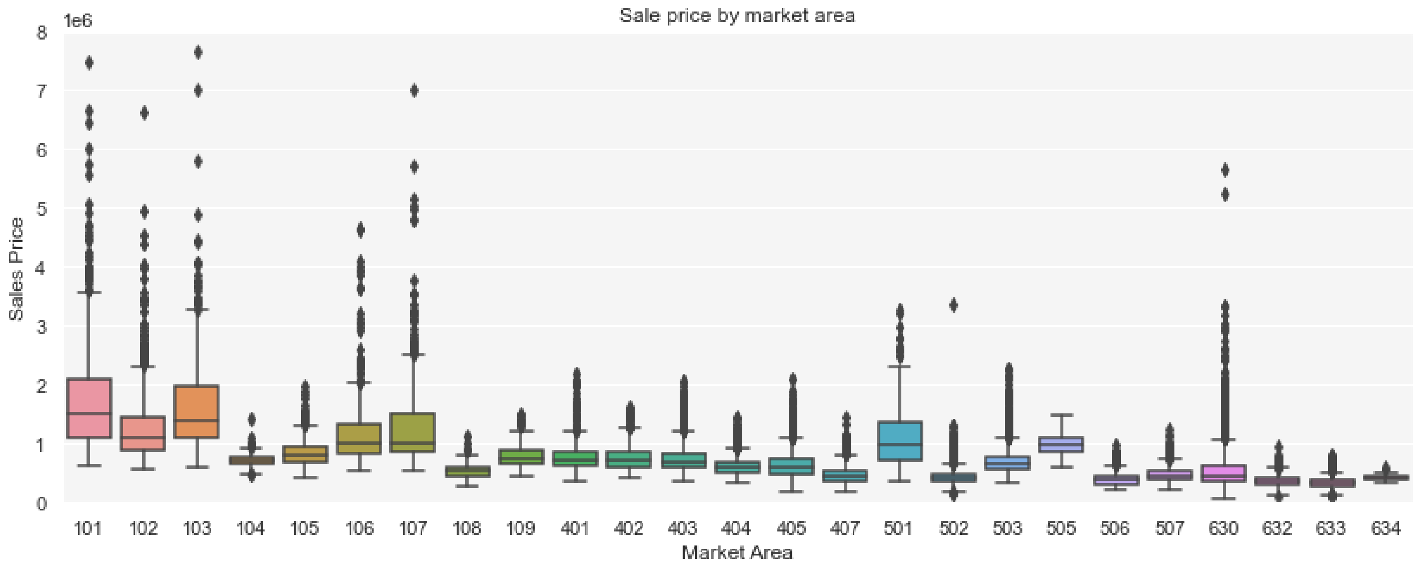


Most homes are around 2000 SF and basements around 1000 SF. This gives an idea what typical single-family house looks like.

Bivariate analysis

Sale price vs Market Area.

For someone who is considering entering the residential real estate market, the graph of sale prices by market areas can be an incredibly useful tool. It provides a comprehensive view of the various price ranges and can help buyers understand what options are available in different areas. By examining this information, buyers can gain a better understanding of what type of property might be a good fit for their preferences and budget. Ultimately, this graph can help buyers make more informed decisions about their real estate investment and find a property that meets their needs and expectations.

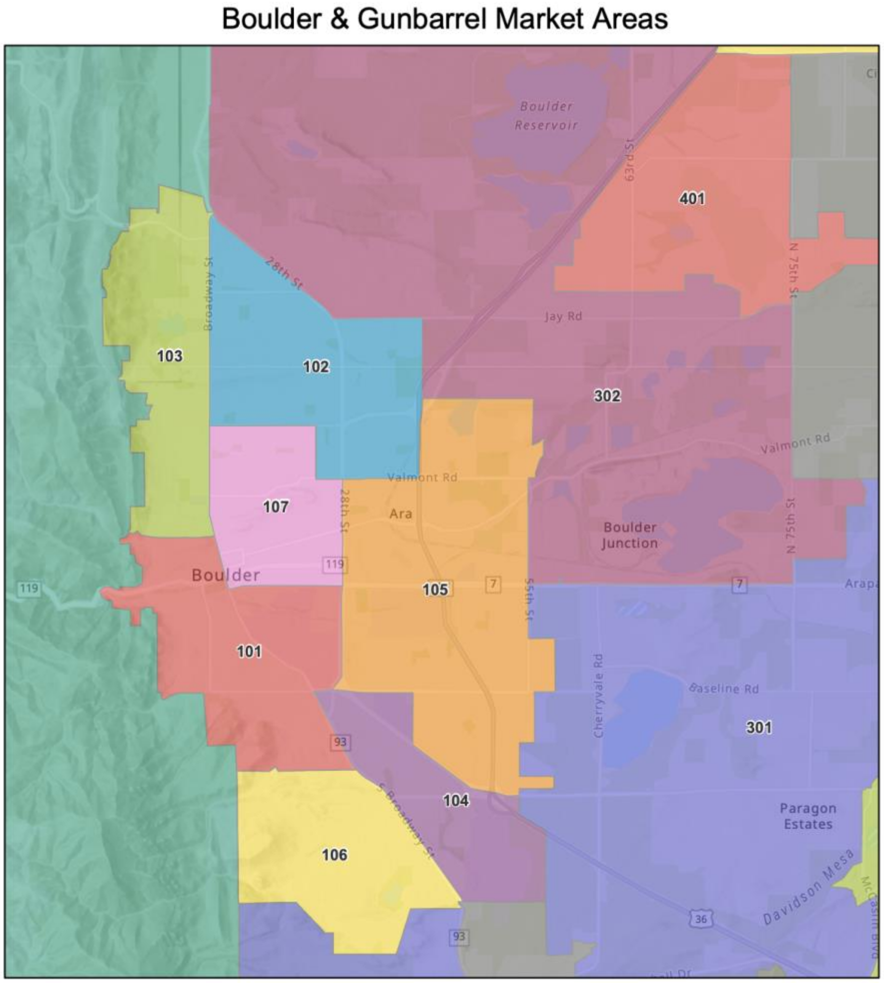


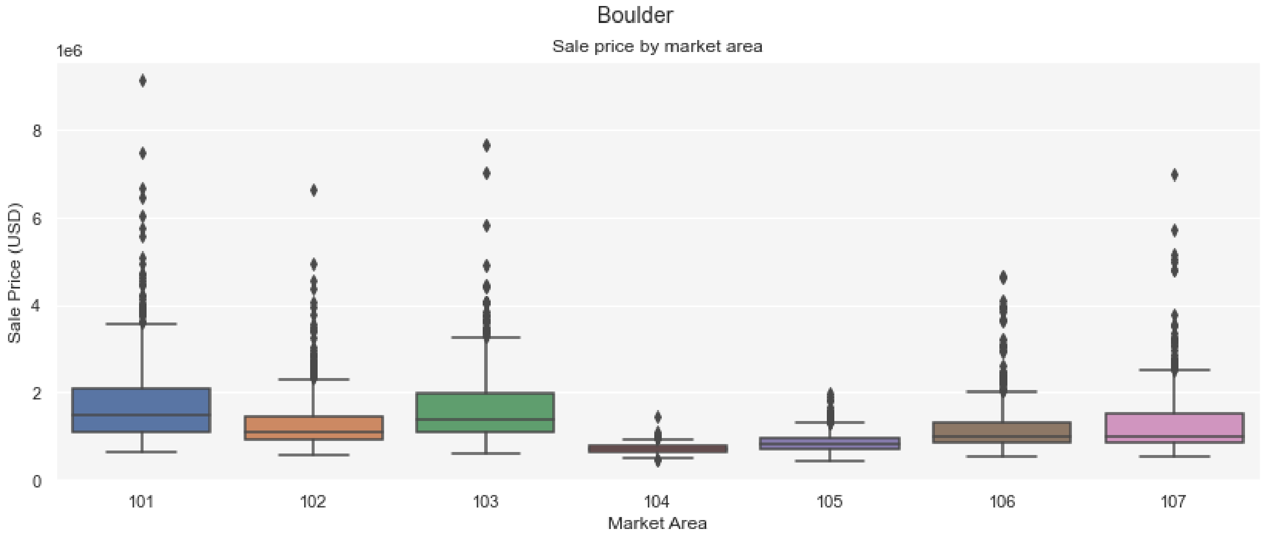
The graph provides a useful overview of the price ranges, as well as the floors and ceilings, for properties available in different areas. For example, the single-family homes in Boulder (101-107) generally have high ceilings, with the exception of 104. The next highest price pocket is Niwot (501). Boulder condos are also priced at the higher end of the range, although Niwot condos are priced more similarly to those in Longmont and Lafayette.

Suburban areas (400s) generally provide a medium price range. Townhomes (108, 109, 407, 506) show some interesting variations, with 108 being much lower priced than 109, and 407 and 506 being priced about the same.

Among condos (630, 632, 633, 634), Boulder condos have both a higher average price and higher ceilings than those in other areas, with mountain condos coming close but being very rare.

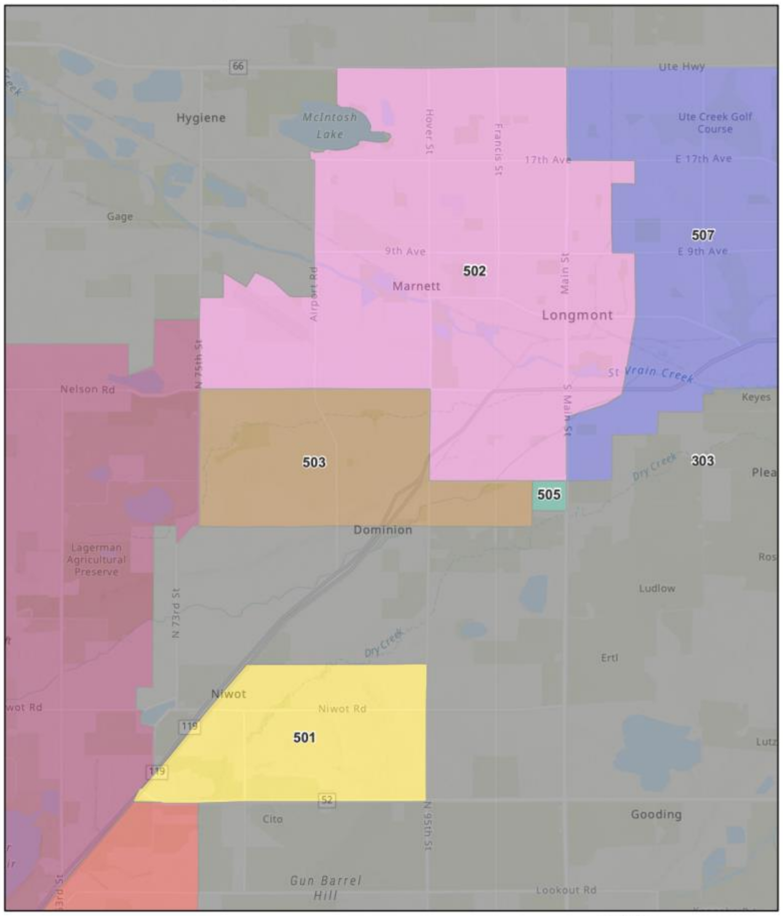
Below we will look closer at specific areas.

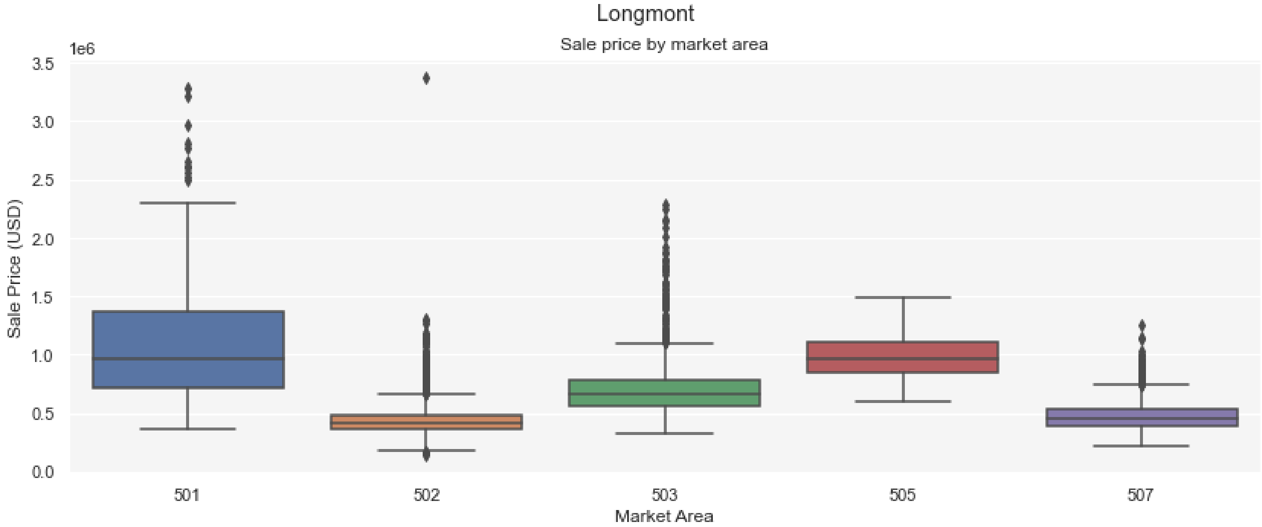




The graph shows that the market area with the lowest average sale price is 104, which encompasses South Boulder between Broadway and 36. South Boulder is located just south of the University of Colorado Boulder and is known for its family-friendly atmosphere and access to many parks and outdoor recreation opportunities. This area appears to have fewer sales compared to other market areas. The next lowest average sale price is in Market Area 105, which is located between 28th and 55th and also has low ceilings, typically under $2 million USD. Market Area 106, which covers Table Mesa South, follows closely behind, with Market Areas 102 and 107 not far behind. The highest average sale prices are found in Market Areas 103 and 101. By analyzing this data, potential buyers can determine the level of exclusivity they are seeking within their budget, especially in the luxury market.

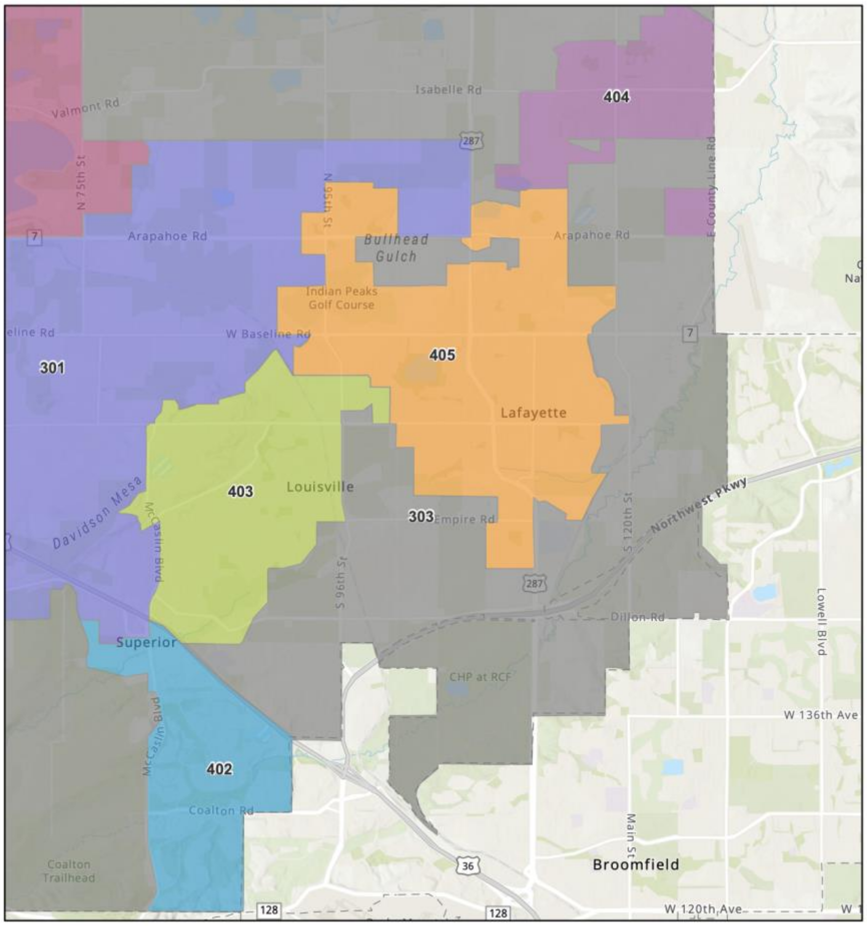
Longmont & Niwot Market Areas 2021

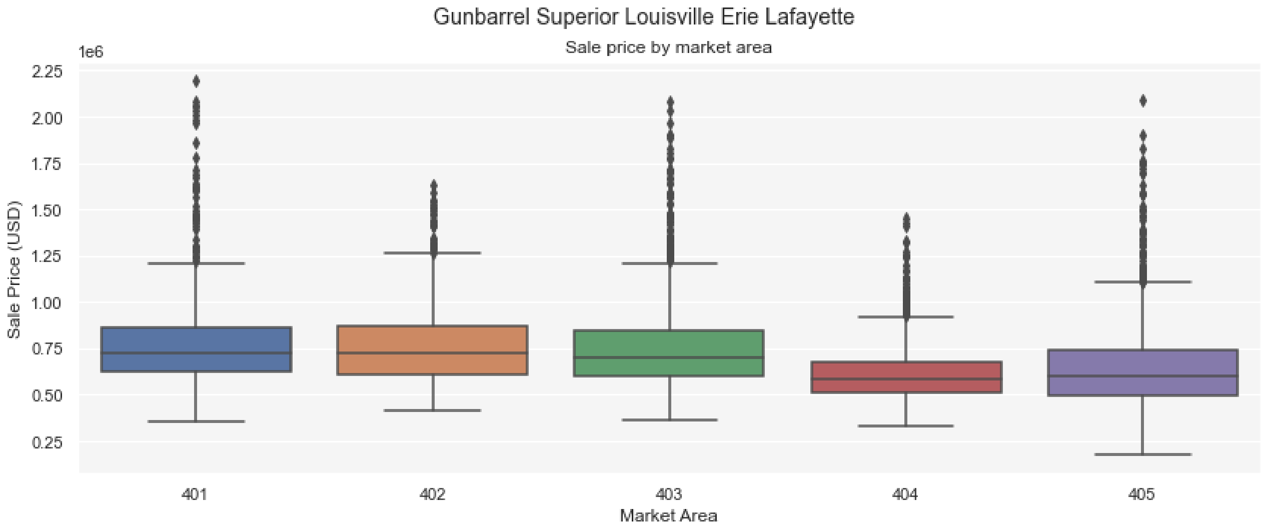




Notable areas in the region include Niwot (501) and South Longmont (503), along with a small pocket along HWY 287 (505). Most of Longmont's residential real estate can be found in areas 502 and 507, which appear to have comparable floors, ceilings, and averages. In general, Longmont offers a wide variety of housing options and exhibits a relatively consistent landscape, with a handful of luxury options located primarily in Niwot and the southern part of the city.

Erie, Lafayette, Superior, & Louisville Market Areas



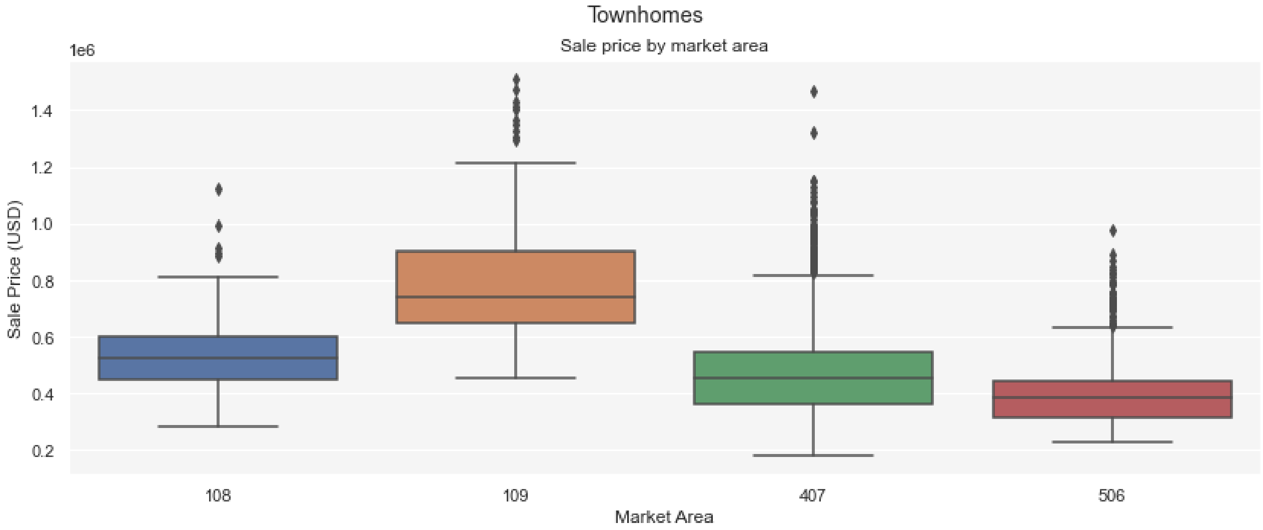


Gunbarrel, Superior and Louisville are slightly higher averages compared to Erie and Lafayette. Overall, the areas look very uniform.

For market participants interested in condos/townhomes here are the following graphs:

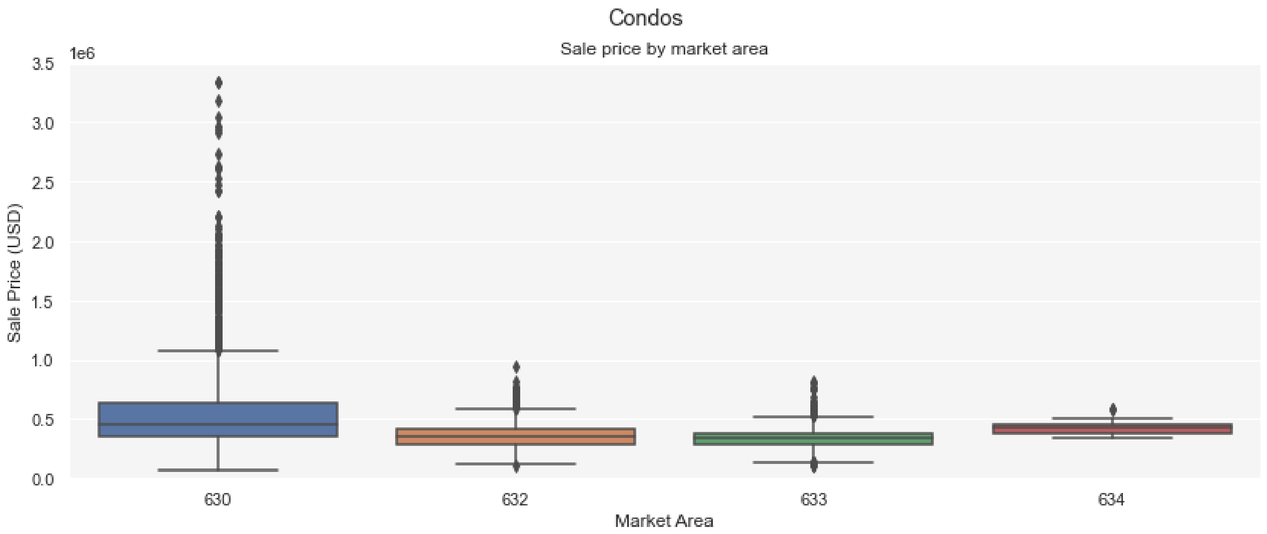
Price vs Market Area for residential townhomes

- Market Area 108 – Boulder, East of Broadway and All Other Townhomes
- Market Area 109 – Boulder, West of Broadway and Downtown Townhomes
- Market Area 407 – Erie, Gunbarrel, Lafayette, Louisville, Niwot, & Superior Townhomes
- Market Area 506 – Longmont Townhomes

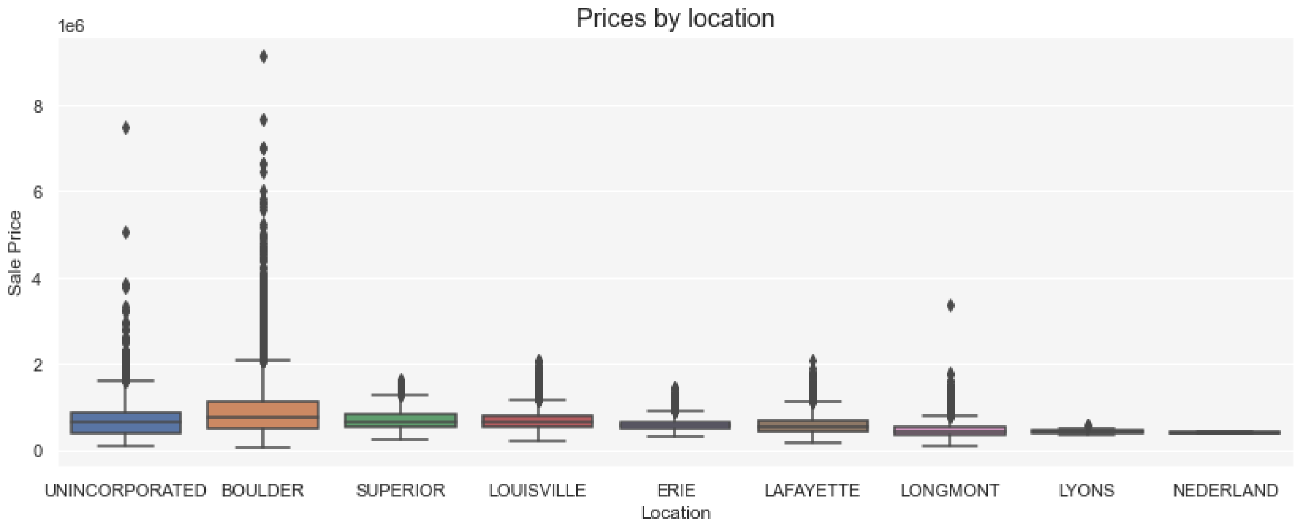


Price vs Market Area for residential condominiums

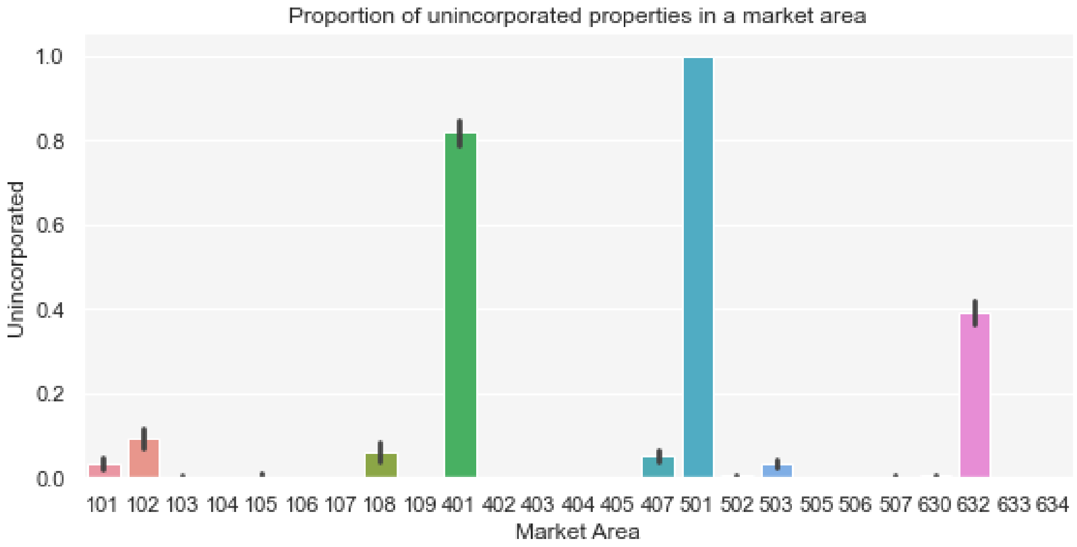
- Market Area 630 – Boulder Condos
- Market Area 632 – Erie, Gunbarrel, Lafayette, Louisville, Niwot, & Superior Condos
- Market Area 633 – Longmont Condos
- Market Area 634 – Mountain Condos



Price vs location.

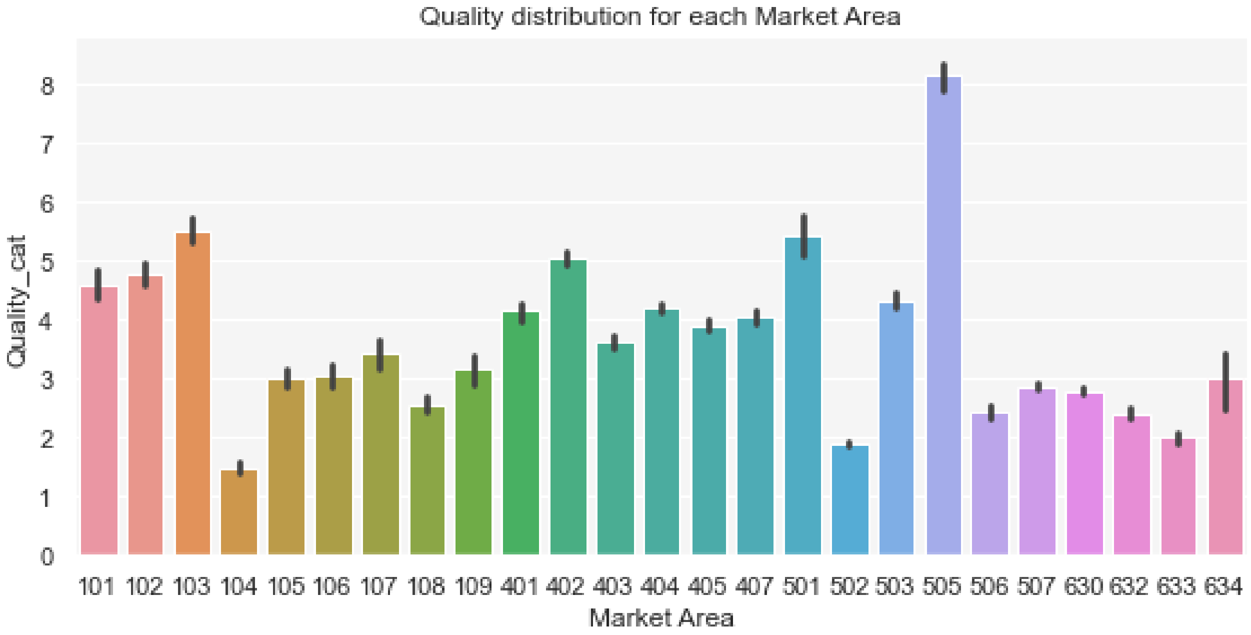


While not very noticeable, the highest floor seems to be in Erie, which is characterized by low diversity of housing options leading to a more mono feel.

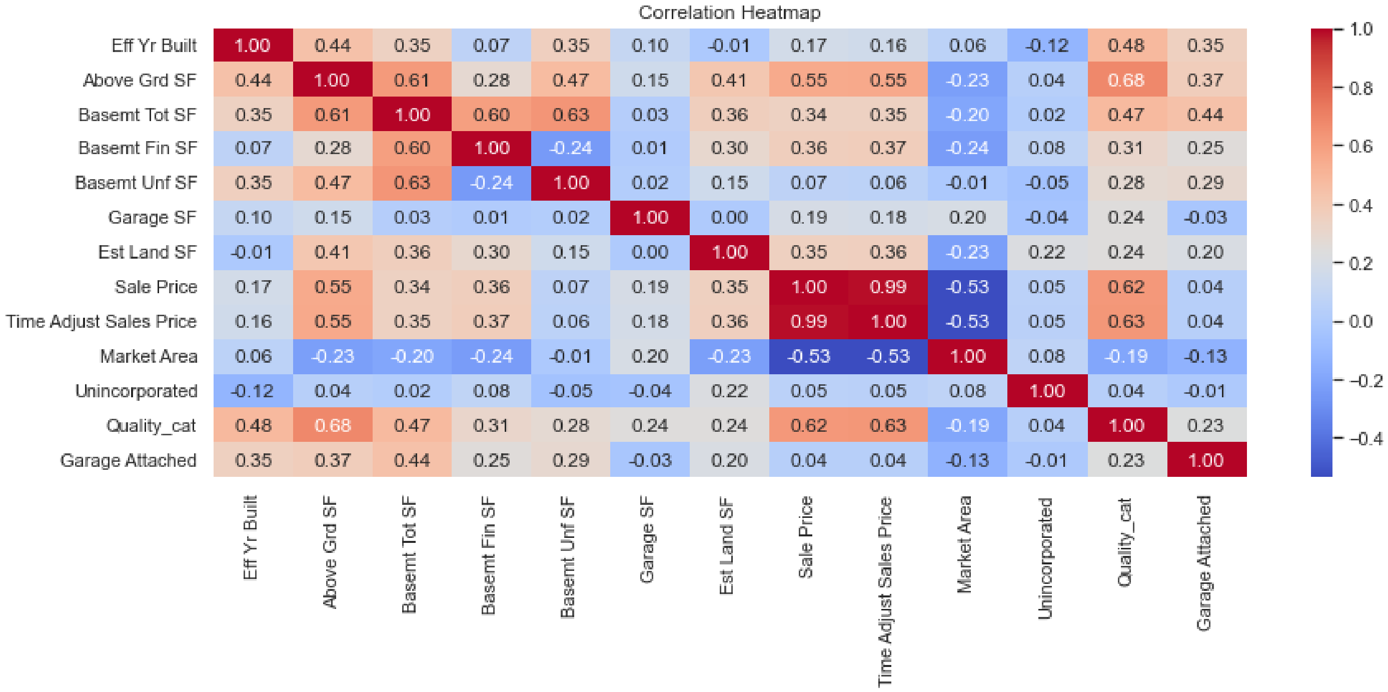


Most unincorporated properties are in Gunbarrel and Niwot with very few in Boulder and Longmont.

Quality rating of the house seems to be quite dependent on the market area as well, with higher quality ratings being given to the properties in the more expensive market areas.

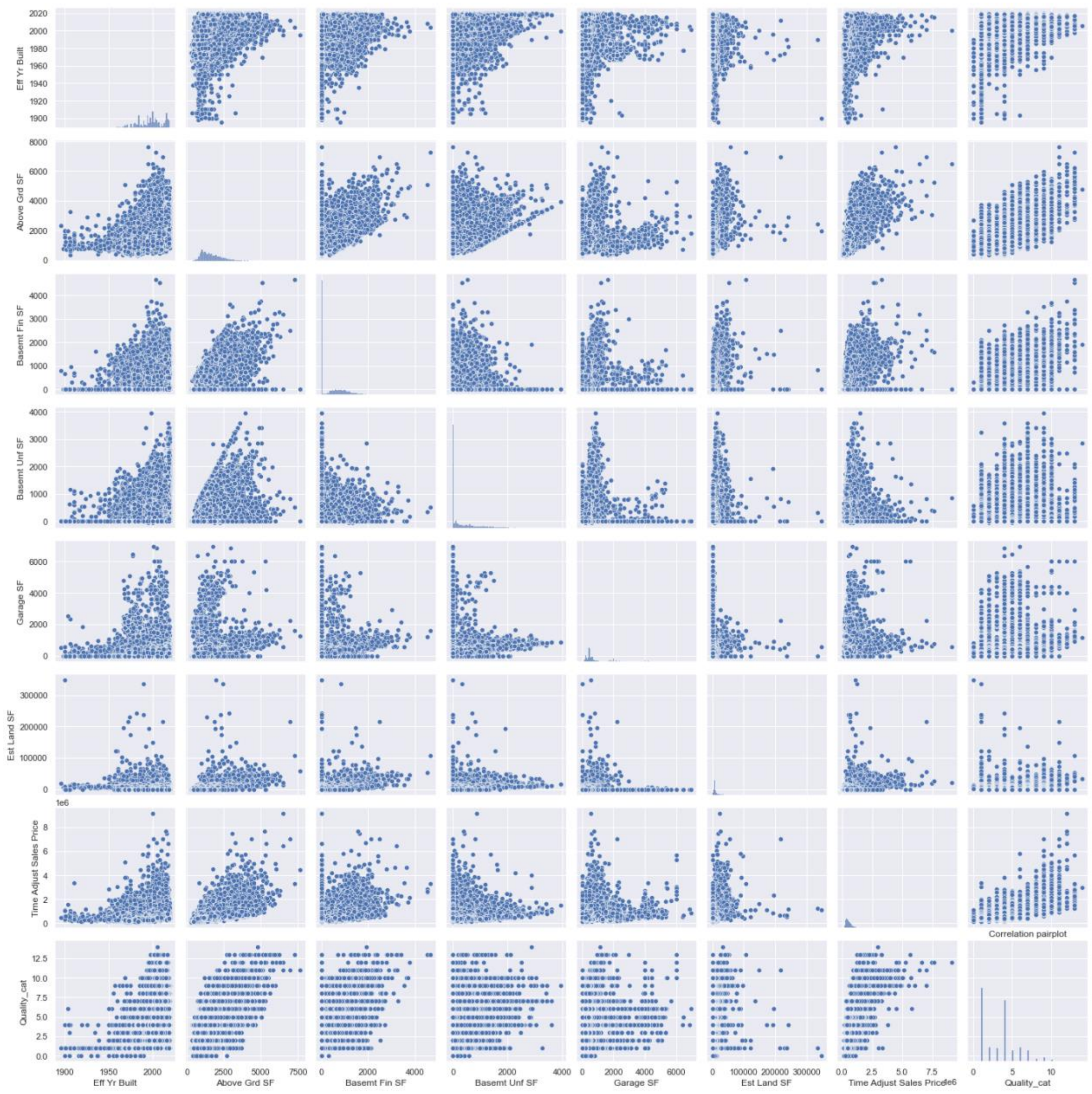


This looks like an interesting relationship requiring further investigation. We perform a chi squared test of independence and get p-value of 0.0, meaning that the two variables are not independent.



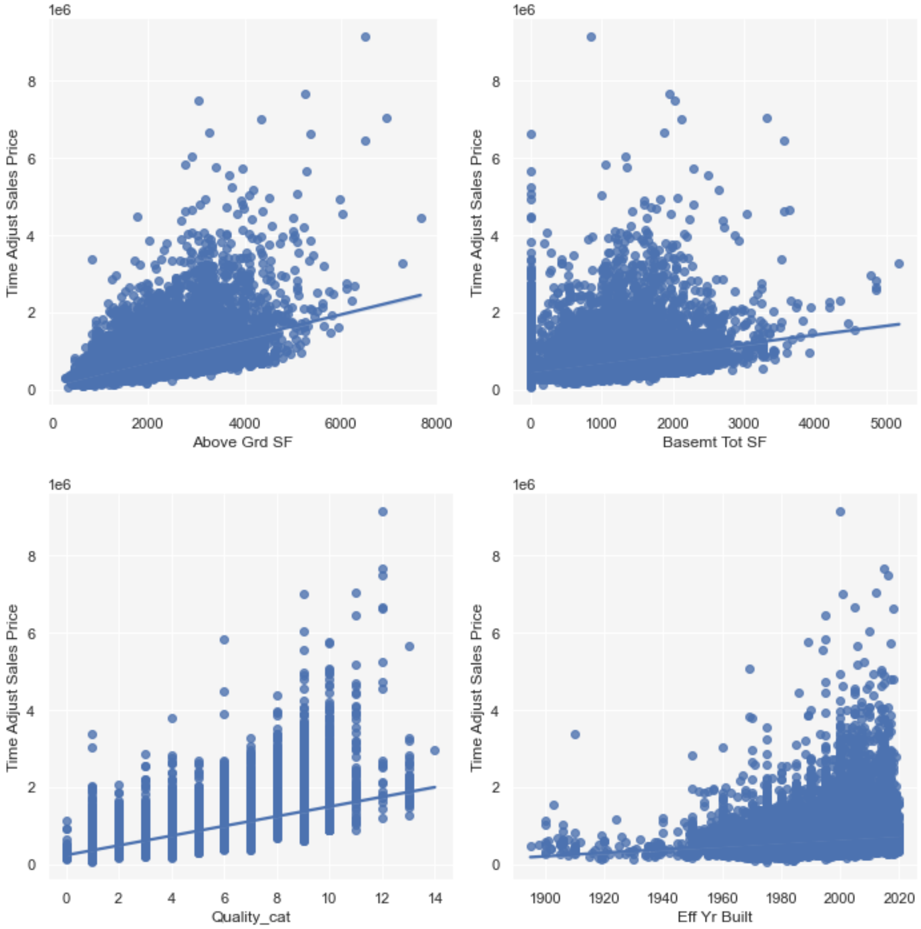
	Eff Yr Built	Above Grd SF	Basemt Tot SF	Basemt Fin SF	Basemt Unf SF	Garage SF	Est Land SF	Sale Price	Time Adjust Sales Price	Market Area	Unincorporated	Quality_cat	Garage Attached
Eff Yr Built	1.00	0.44	0.35	0.07	0.35	0.10	-0.01	0.17	0.16	0.06	-0.12	0.48	0.35
Above Grd SF	0.44	1.00	0.61	0.28	0.47	0.15	0.41	0.55	0.55	-0.23	0.04	0.68	0.37
Basemt Tot SF	0.35	0.61	1.00	0.60	0.63	0.03	0.36	0.34	0.35	-0.20	0.02	0.47	0.44
Basemt Fin SF	0.07	0.28	0.60	1.00	-0.24	0.01	0.30	0.36	0.37	-0.24	0.08	0.31	0.25
Basemt Unf SF	0.35	0.47	0.63	-0.24	1.00	0.02	0.15	0.07	0.06	-0.01	-0.05	0.28	0.29
Garage SF	0.10	0.15	0.03	0.01	0.02	1.00	0.00	0.19	0.18	0.20	-0.04	0.24	-0.03
Est Land SF	-0.01	0.41	0.36	0.30	0.15	0.00	1.00	0.35	0.36	-0.23	0.22	0.24	0.20
Sale Price	0.17	0.55	0.34	0.36	0.07	0.19	0.35	1.00	0.99	-0.53	0.05	0.62	0.04
Time Adjust Sales Price	0.16	0.55	0.35	0.37	0.06	0.18	0.36	0.99	1.00	-0.53	0.05	0.63	0.04
Market Area	0.06	-0.23	-0.20	-0.24	-0.01	0.20	-0.23	-0.53	-0.53	1.00	0.08	-0.19	-0.13
Unincorporated	-0.12	0.04	0.02	0.08	-0.05	-0.04	0.22	0.05	0.05	0.08	1.00	0.04	-0.01
Quality_cat	0.48	0.68	0.47	0.31	0.28	0.24	0.24	0.62	0.63	-0.19	0.04	1.00	0.23
Garage Attached	0.35	0.37	0.44	0.25	0.29	-0.03	0.20	0.04	0.04	-0.13	-0.01	0.23	1.00

Let's examine the pairwise correlations in greater detail. The most significant correlation of 0.68 is observed between the size of the house and its quality, followed by the basement's total area and the above-ground square footage at 0.61. The sale price and time-adjusted sale price have a high correlation of 0.99, and we will only consider the time-adjusted sale price for analysis purposes. Furthermore, the time-adjusted sale price demonstrates a moderate correlation with the quality of the house at 0.63 and a correlation of 0.55 with the above-ground square footage. Year built shows moderate correlation values of 0.48 with Quality and 0.44 with Above Ground SF. Additionally, various square footage measurements display moderate correlations with each other and sale price.



There are several linear associations worth noting, such as the correlation between quality and above ground SF (0.68), above ground SF and time adjusted sale price (0.55), finished basement and time adjusted sale price (0.37), and quality (0.31). The pair plot provides a more detailed visualization of the correlation heatmap.

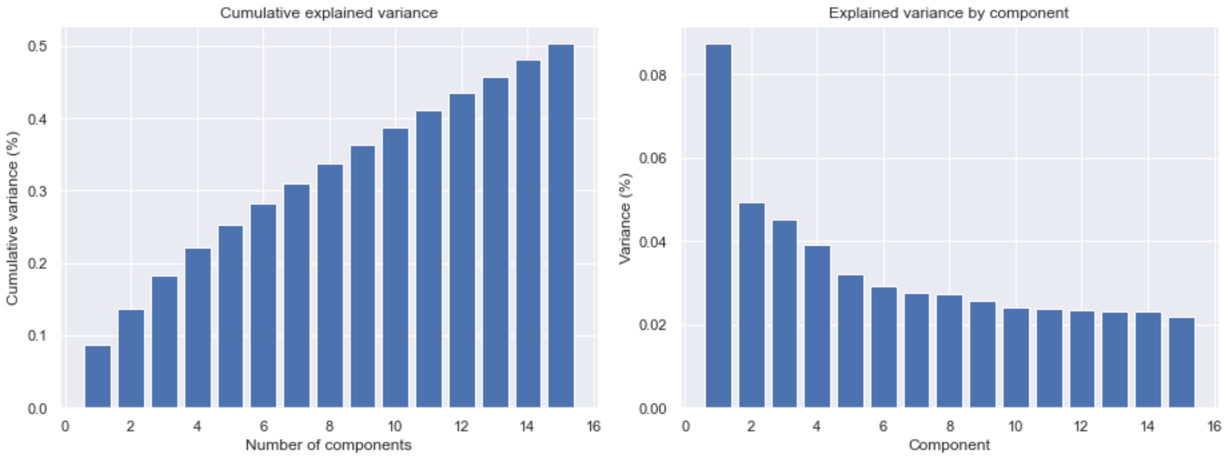
Below we will use Linear Regression plots to discuss the associations in more detail.



Principal component analysis

PCA is a technique that transforms the original set of correlated variables into a new set of uncorrelated variables called principal components. These principal components are ordered in terms of their ability to explain the variance in the data. By reducing the dimensionality of the data, PCA can help to simplify complex datasets and improve the performance of machine learning models. Furthermore, it can help to identify which features are contributing the most to the variation in the data, providing valuable insights for further analysis.

We calculate principal components and the amount of variance they explain below.



We can see that even the largest component only explains about 9% of the variation, with the next 5% of the variation and then slow decline with plateau at about 3% for every additional component. Given this result, it doesn't make sense to use the PCA for our problem since the variance that's captured doesn't tend to be in a few major directions, but instead distributed across features.

Modeling

Simple linear regression

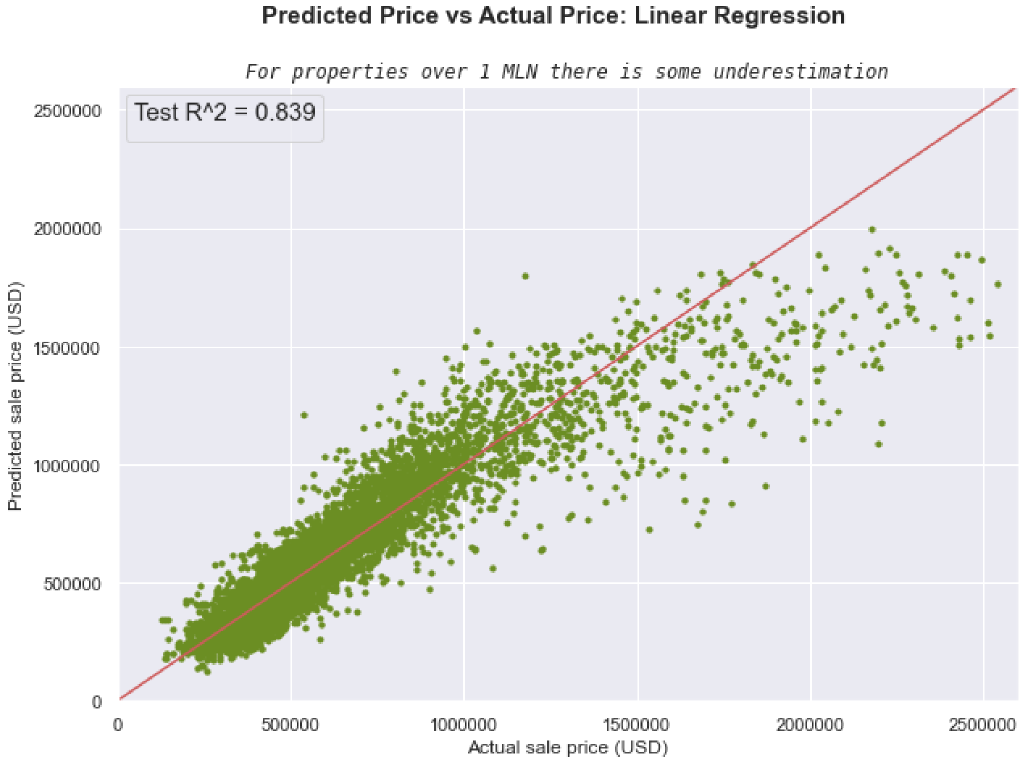
We start with simple linear regression model to use it as a benchmark to compare other more complicated and less interpretable models against it. First, we split the data into train and test set where training set contains 75% of the observations, which is 20127 and test set contains 6710 data points, which is 25% of the observations. Next, we fit simple linear regression model with an intercept term via statsmodels OLS.

Table 1. OLS Regression Results

Dep. Variable:	Time Adjust Sales Price	**R-squared:**	0.846
Model:	OLS	**Adj. R-squared:**	0.845
Method:	Least Squares	**F-statistic:**	2595
Durbin-Watson:	2.016	**Prob (F-statistic):**	0.00

The model's R-squared is 0.780. The test set R-squared is 0.846, which we set as our benchmark R-squared. The model doesn't overfit. t-statistics look good for all coefficients except "**Market**

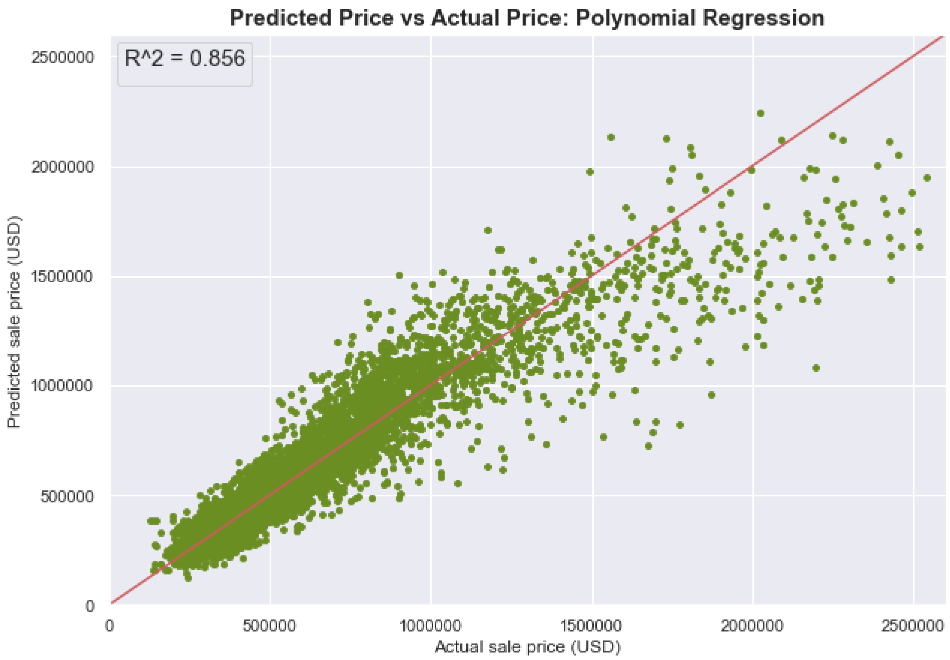
Area_105’ and “Market Area_104”. The Durbin-Watson 2.016 and between 1.5 and 2.5, so autocorrelation is likely not a cause for concern.



Clearly, there is a problem. The more expensive properties are not properly priced under linear regression model, however for properties under 1.3 million USD the fit is great.

Polynomial regression

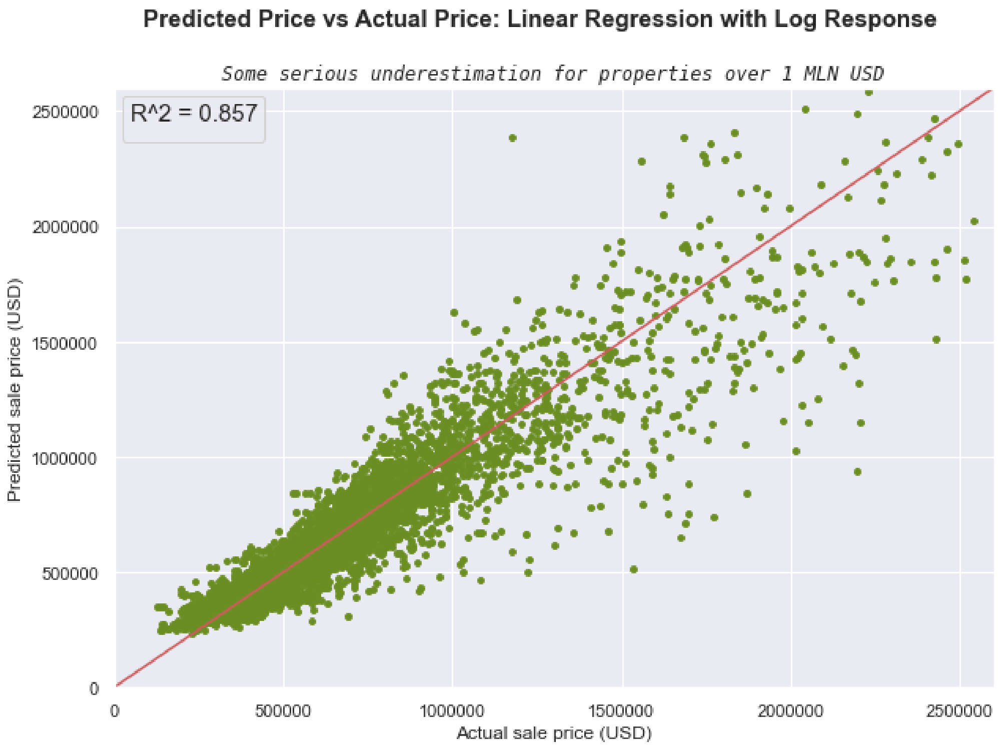
Based on the given information, we can see that polynomial regression with second-degree polynomial features and interactions performs better than simple linear regression. The model achieves a test set R2 of 0.856 and an RMSE of 129436. The train set R2 is 0.865, which indicates almost no overfitting.



There seems to be some issues with the relationship for high sale price properties, but the number of observations in that range is not very large. Further investigation is needed to identify what could be driving this discrepancy. Overall, the model appears to be performing well and does not seem to be systematically over or underestimating the prices.

Linear regression with log transformed response.

We took a logarithm of base 10 of the response variable and fit a simple linear regression model. The results we got were: Test set R^2: 0.857, RMSE: 129093. Train set R^2 was 0.884 indicating no overfitting.



XGBoost

XGBoost is one of the most popular machine learning frameworks among data scientists. According to the Kaggle State of Data Science Survey 2021, almost 50% of respondents said they used XGBoost, ranking below only TensorFlow and Sklearn[2].

XGBoost is an optimized distributed gradient boosting library. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples[3]. Here we want to solve our regression problem. As with many other ML algorithms, there is an issue of parameter tuning[4]. We perform a grid search using 3 parameters.
Here we focus on 3 main parameters:

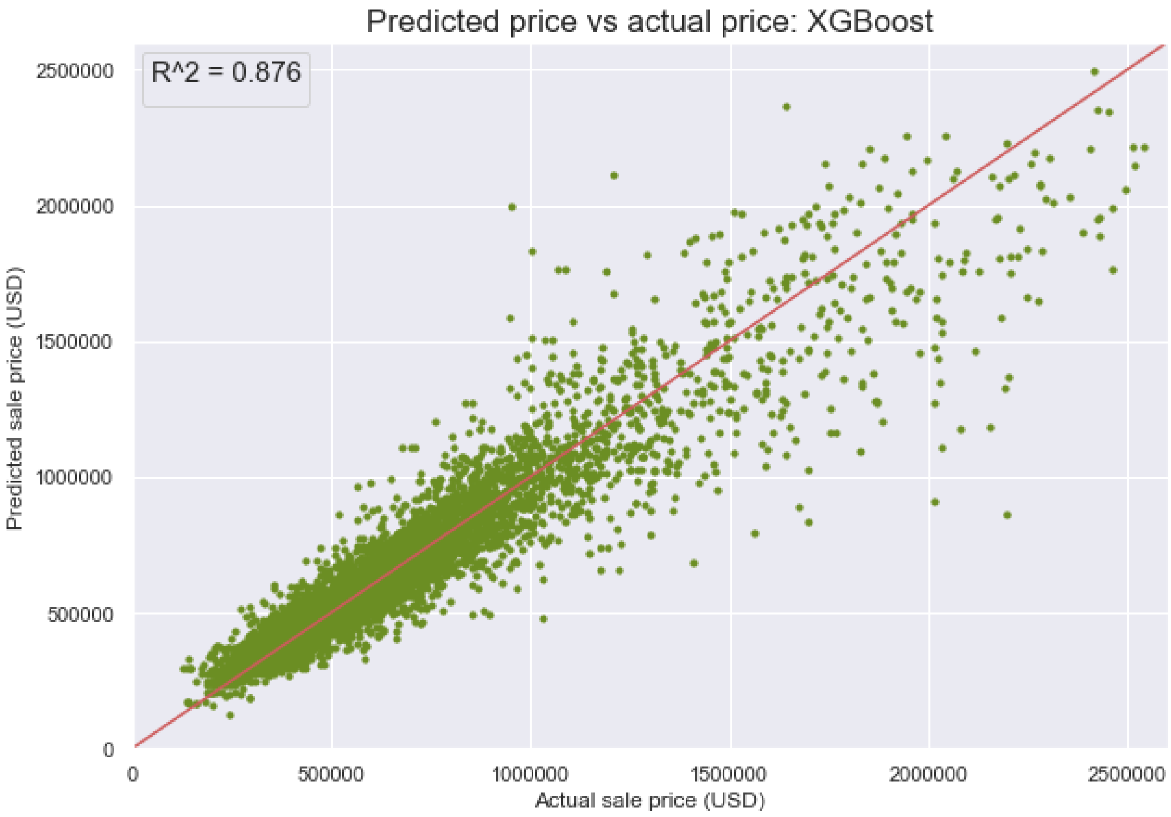
- learning_rate (eta) shrinks the feature weights to make the boosting process more conservative. Default 0.3 [0,0.2,0.5]

[2] https://www.datacamp.com/tutorial/xgboost-in-python
[3] https://xgboost.readthedocs.io/en/stable/
[4] https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook

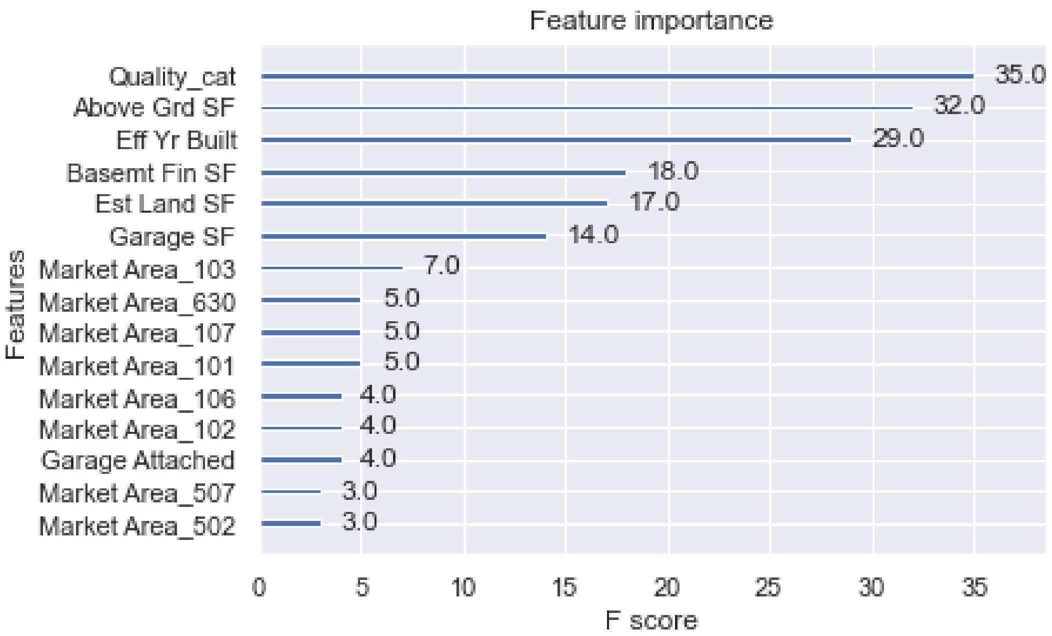
- min_split_loss (gamma) specifies the minimum loss reduction required to make a split. Default 0 [0,10,100]
- reg_lambda, default 1, L2 regularization term on weights (analogous to Ridge regression). Increasing this value will make model more conservative. [1,10,100]

The analysis reveals that the optimal eta value is 0.5, while the default values for the remaining parameters work well.

With R2 of 0.876 and RMSE of 120181, it is a strong candidate for a top method.



Let’s explore XGBoost’s feature importance table.



The feature that has the most significant impact on the predicted house price is the above ground square footage. Quality, year built, land square footage, basement finished square footage, garage square footage, and basement Next go market areas, with the top being the most expensive areas are in Boulder.

Random forest regression

We use Random Forest Regressor from sklearn.ensemble. A random forest is a meta estimator that fits several decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting[5].

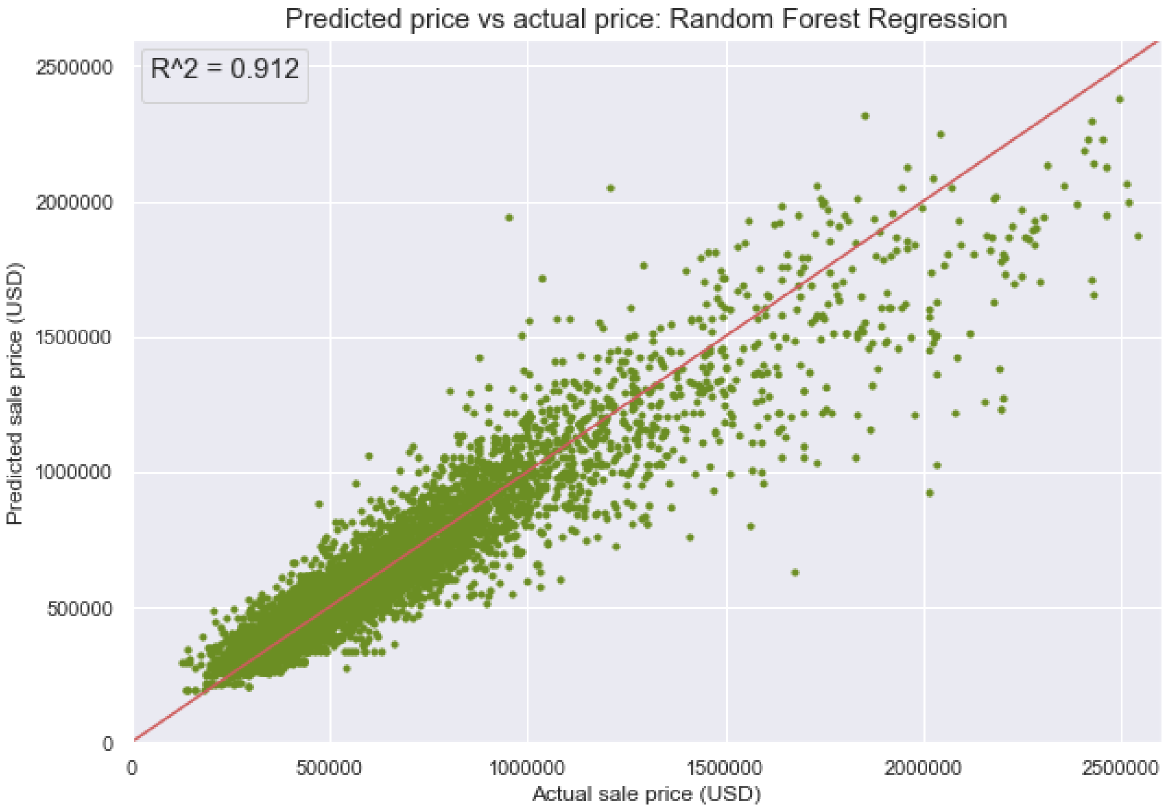
When we used the default model on the training data, we found that there was some overfitting going on, with the training data R^2 being 0.98 and the test data R^2 being 0.91. To mitigate this, we performed hyperparameter optimization using random search. Since default parameters lead to fully grown unpruned trees and likely lead to overfitting, as we have seen, we want to search parameter values that are more restrictive. We searched between 10 and 100 trees, considered all or only 30% of the features, checked max number of levels per tree between 10 and 110, allowed for 2,5 or 10 Minimum number of samples required to split a node, allowed for 1,2,4 Minimum number of samples required at each leaf node, and used or not use bootstrap.

The resulting optimal model had n_estimators=70, min_samples_split=5, min_samples_leaf=1, max_features=0.3, max_depth=80, and bootstrap=False. The optimized model showed an accuracy improvement of 0.55% over the default model.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

So, according to our hyperparameter optimization, we can gain 0.55% in accuracy improvement. Overall, the selected optimal model is much more restrictive yet yields similar performance as the base model, and hence we recommend using the optimal model.

Upon evaluating the final optimized model, we found that the Random Forest test set R^2 was 0.885 and the RMSE was 155176.301.



The model worked well for houses under 1,000,000 USD, but as the price increased, the performance decreased. We observed more severe undervaluing by the model as opposed to overvaluing, and some of the more expensive houses over 3 million were undervalued.

Conclusion

Table. Model Metrics

	R squared	RMSE	Parameters

Multiple linear regression	0.839	136589	
Polynomial regression	0.856	129436	
Multiple linear regression model with log transformed response	0.857	129093	
XGBoost	0.887	114524	{'objective': 'reg:squarederror', 'max_depth': 4, 'reg_lambda': 1, 'min_split_loss': 0, 'learning_rate': 0.5}
Random Forest	0.912	100904.277	{'n_estimators': 70, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 0.3, 'max_depth': 80, 'bootstrap': False} Random Forest Train set R^2: 0.993

The three regression models presented show an improvement in the R squared value as we move from the Multiple Linear Regression model to the Polynomial Regression model, and finally to the Multiple Linear Regression model with log transformed response. This suggests that incorporating more complex terms and transformations into the regression model can improve its ability to explain the variance in the response variable.

The Polynomial Regression model shows a slightly better R squared value of 0.856 compared to the Multiple Linear Regression model with an R squared value of 0.839. However, the Multiple Linear Regression model with log transformed response has the highest R squared value of 0.857 among the three models.

The RMSE values for the Polynomial Regression and the Multiple Linear Regression model with log transformed response are quite close to each other, with the Polynomial Regression model having an RMSE of 129436 and the Multiple Linear Regression model with log transformed response having an RMSE of 129093. The Multiple Linear Regression model has a slightly higher RMSE value of 136589.

Overall, based on the R squared and RMSE values, it seems that the Multiple Linear Regression model with log transformed response is the best-performing model among the three models presented.

Compared to the three regression models discussed previously, XGBoost shows a substantial improvement in the R squared value, achieving an R squared value of 0.887, which is higher than the best R squared value of 0.857 achieved by the Multiple Linear Regression model with log transformed response. Additionally, XGBoost has a lower RMSE value of 114524.052, indicating that it is better at predicting the response variable than the three regression models.

XGBoost is a powerful ensemble learning algorithm that combines multiple weak decision trees to create a strong predictive model. The model is trained iteratively, with each new tree attempting to correct the errors of the previous trees, resulting in a final model that is capable of making accurate predictions on unseen data.

The hyperparameters used to train the XGBoost model are also optimized, with a learning rate of 0.5 and a maximum depth of 4. These hyperparameters likely contributed to the high performance of the XGBoost model.

Overall, XGBoost is a very strong performer among the models discussed, with a higher R squared value and lower RMSE value than the three regression models. It is worth noting, however, that XGBoost may require more computational resources to train and may be more complex to interpret than the simpler regression models.

Based on the information provided, it appears that Random Forest is the best performing model among those listed, with an R squared value of 0.912 and the lowest RMSE value of 100904.277. The model was trained using a set of hyperparameters that were tuned to maximize performance. It is important to note that the Random Forest model was also able to achieve a very high R squared value of 0.993 on the training set, indicating that it is likely overfitting to the data.