Data-Driven Real Estate in Boulder County

Hanna McDonell, Nikhil Rowland, Trishala Thakur
University of Colorado Boulder
Boulder, USA
hanna.mcdonnell@colorado.edu, nikhil.rowland@colorado.edu, trishala.thakur@colorado.edu

***Abstract*— Homeownership has long been the primary way for ordinary Americans to build wealth. However, as housing prices continue to increase at rates that far outweigh increases in salary, the woes of the prospective first time homebuyer have never been greater. In our model, we hope to show what factors are the greatest predictors of property values to better help home buyers differentiate good deals from pitfalls. By leveraging data scraped from Zillow for recent sales and supplementing it with comprehensive real estate data from Boulder County, we construct a holistic view of the housing market. To combine these datasets, we cleaned and processed the data before ultimately determining which factors have the most influence on real estate prices. We also explain which factors our model couldn't control for, and the shortcomings of our present model that future research could improve upon.**

I. INTRODUCTION

For most Americans, buying a house is the largest purchase they will ever make. Given what's at stake, responsible home buyers should seek to make the most informed decision they can before buying a property. However, the real estate market is complicated and full of pitfalls that the average home buyer may not be aware of. A home could have mold, be in a floodplain, or have zoning restrictions, among a heap of other issues. Due to school zones and neighborhood homeowner associations, houses located even just a street apart may vary greatly in desirability. Many homebuyer hopefuls turn over the reins to a real estate agent to smooth over the process, but this comes with its own problems. Real estate agents often get paid through commission, so they are incentivized to make deals and close them quickly, sometimes at the expense of their clients. If the agent represents both parties, they are also incentivized to sell at a high price to maximize their commission, working against the homebuyer.

To combat these issues, sites like Zillow and Redfin have come along that help homebuyers in a variety of ways. Their first and primary function is to serve as a place to view real estate listings. Buyers can create custom filters to sort houses by price, neighborhood, square footage, and other metrics they might be interested in. These sites also come with a valuation tool that displays the approximate value of a property given the price that comparative properties have recently sold at. Using these tools, even amateur homebuyers can perform basic market analysis to see if a real estate listing is selling at a discount or a premium. They can also use these tools to check the history of the property, including any history of renovations or upgrades, and applicable tax information. By looking at the school zones outlined on these websites, buyers can also easily tell which neighborhoods would better fit their needs.

The real estate market in Boulder, Colorado is particularly woeful. Not only are the housing prices in Boulder the highest in the state, they are some of the highest in the entire country. With a home price to income ratio of greater than 10, even an average two bedroom apartment rents for more than 2400 dollars a month. This is largely due to the restrictive zoning laws that promote the construction of single family unit homes instead of larger projects that could house several or even dozens of families. The housing problem has become so great that many people who work or go to school in Boulder have been forced to move to surrounding towns such as Westminster, Broomfield, and Superior in search of cheaper housing. The cost of housing in Boulder only reinforces the need for homebuyers to make well informed decisions about what they value the most in a home, and perhaps more importantly, what they can live without.

Residential homes aren't the only pieces of real estate on the Boulder market. Commercial and industrial real estate also form a large portion of the real estate market, and their valuations and the methods used to value these properties can differ drastically from the formula used for residential housing. In the commercial and industrial sectors, property value is often influenced by factors such as location, accessibility, and potential income-generating capabilities. Unlike the comparable sales approach commonly used in residential real estate, where the prices of similar properties in the area are considered as the primary metric of valuations. Additionally, the unique needs of businesses and industries can significantly impact the design and functionality of these properties, adding another layer of complexity to their valuation and assessment.

Ultimately, through our analysis, we hope to create a tool that will help optimize the homebuying experience. By identifying the factors with the greatest influence on home prices in the Boulder area, buyers will be empowered to make more informed decisions about their purchases. They will gain the ability to directly compare the utility they derive from specific features of a house, such as its location, size, amenities, and condition, against the expected price of comparable properties. This tool will not only enhance transparency in the real estate market, but also assist buyers in finding properties that best align with their preferences and budgetary constraints, leading to more satisfactory homebuying outcomes.

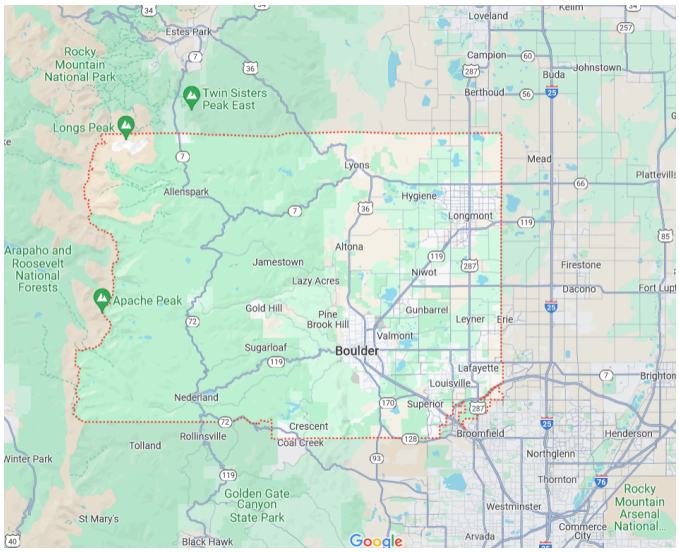

Fig. 1. Boulder County Map

II. RELATED WORK

Our study delves into the challenges confronting prospective first-time homebuyers in the United States, aligning with existing research that underscores homeownership as a pivotal strategy for wealth accumulation among ordinary Americans. The widening gap between housing prices and income growth accentuates the urgency of our investigation into factors influencing property values.

In contributing to the literature, our aim is to uncover the primary determinants of real estate values, empowering homebuyers to navigate the intricate and evolving housing market landscape with confidence. By amalgamating data sourced from recent Zillow sales and comprehensive real estate records from Boulder County, our methodology offers a comprehensive view of housing market dynamics. Through meticulous data cleaning and processing, we strive to identify the fundamental drivers shaping real estate prices.

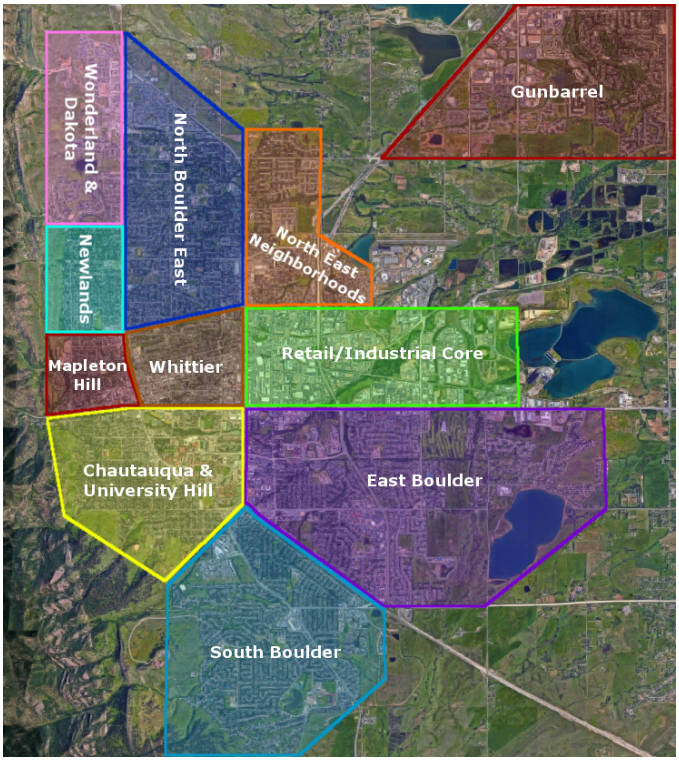

Fig. 2. Boulder County Map 2

Moreover, we acknowledge the constraints of our model in capturing certain variables and pinpoint avenues for future research to mitigate these limitations. This acknowledgment underscores our commitment to advancing our comprehension of the complexities of the housing market and fortifying decision-making support for prospective homebuyers.

Drawing from the realms of real estate and machine learning, prior studies have employed diverse methodologies, including traditional regression models, spatial analyses, and machine learning algorithms like random forests and neural networks, to forecast property values. While some investigations have centered on specific geographic locales or property types, others have explored broader market trends and patterns. Building on this foundational research, our study endeavors to introduce innovative insights and methodologies to the discipline, thereby enhancing our capacity to comprehend and forecast real estate dynamics in a swiftly evolving environment.

In parallel, there is a burgeoning interest among researchers and industry professionals in quantifying subjective attributes such as "feel," "aesthetic," and "beauty" when evaluating properties. This endeavor amalgamates elements from psychology, sociology, and design theory to decipher the emotional and psychological factors influencing homebuyers' perceptions and choices.

One avenue to quantify these subjective attributes involves leveraging sophisticated data analytics techniques, such as sentiment analysis and image recognition algorithms. Sentiment analysis enables the assessment of public sentiment and perception of a property's aesthetic qualities by analyzing textual data from property listings, online reviews, and social media posts. Similarly, image recognition algorithms discern features associated with architectural style, interior design, and overall attractiveness from visual data.

Additionally, researchers have explored immersive technologies like virtual reality (VR) and augmented reality (AR) to create interactive experiences that allow homebuyers to virtually explore properties and experience their aesthetic attributes firsthand. By simulating various design styles, layouts, and decor options, these technologies enable homebuyers to envision themselves living in a space and evaluate its aesthetic appeal intuitively.

Furthermore, studies in environmental psychology and neuroscience have scrutinized the impact of spatial layout, natural light, color schemes, and other design elements on human emotions, well-being, and perceived attractiveness. By integrating insights from these disciplines into real estate analysis, researchers aim to develop comprehensive models that capture the multidimensional nature of a property's appeal and effectively quantify subjective factors like "feel" and "beauty."

In conclusion, while quantifying subjective attributes in real estate poses challenges, ongoing advancements in data analytics, technology, and interdisciplinary research are expanding our understanding of how aesthetic qualities influence property values and shaping the future of homebuying experiences.

III. MAIN METHODS

To perform a real estate analysis, we needed to gather information on the current state of real estate in Boulder, CO. Zillow is a website that posts current real estate for sale. Because of the format, web scraping had to be used to gather relevant information in a usable format.

An API combined with a Python HTML parser was used to gather residential (homes and townhouses) data from Zillow for Boulder, CO, on Sunday, March 17. Information was gathered on price, address, number of bedrooms, number of bathrooms, square footage, latitude and longitude, and zestimate (Zillow's estimate of the home's value). There was no missing data, but not every home collected had a Zestimate. Missing Zestimates were marked with NaN. Price outliers were determined by being outside the range Median 1.5* IQR and were removed for better results in the modeling to come later. No other actions were needed on the beds, baths, and square footage features.

To get data on previous years' residential sales, data was downloaded from the BoulderCounty.gov site. The website had data on all sales from 2018 to 2023, which were divided by property type (apartments, commercial, mixed-use, residential condominiums, and vacant land) and year. Each year's data was downloaded and imported into a Jupyter Notebook file. The yearly data was combined into one data frame for each property type.

	Unnamed: 0	Neighborhood Code	Account #	PARCELNB	PROPERTY_ADDRESS	LOCCITY	SUBNAME	MULTIPLE_BLDGS	ACCOUNT_TYPE
0	0	126.0	R0079450	146320406134	3250 ONEAL CIR 16J	BOULDER	STRATFORD PARK EAST CONDOS - BO	NO	RESIDENTIAL
1	1	126.0	R0089280	146320413015	3393 ONEAL PKWY 32	BOULDER	NORTHGATE CONDOS(THE)&PHSE 2,3,4 BO	NO	RESIDENTIAL
2	2	126.0	R0097796	146320249005	3545 28TH ST 105	BOULDER	PENDLETON SQUARE PHASE IV - BO	NO	RESIDENTIAL
3	3	135.0	R0514429	146329419019	3601 ARAPAHOE AVE 119	BOULDER	PELOTON CONDOMINIUMS PHASE 2	NO	RESIDENTIAL
4	4	146.0	R0037838	146333407004	961 PARKWAY DR	BOULDER	COUNTRY CLUB PARK PT REPLAT - BO	NO	RESIDENTIAL
...	...	...	...	...	...	...	...	...	...
8423	36841	NaN	NaN	157704340023	4800 OSAGE DR 23B	BOULDER	PINON GLEN CONDOS & AMENDED & RESTATED	NO	RESIDENTIAL CONDO
8424	36843	NaN	NaN	157709015003	4838 MOORHEAD CIR	BOULDER	SOUTH CREEK 7 - BO	NO	RESIDENTIAL
8425	36857	NaN	NaN	146332439019	3009 MADISON AVE 203J	BOULDER	WIMBLEDON CONDOS PHASE IV - BO	NO	RESIDENTIAL CONDO
8426	36858	NaN	NaN	146332319006	2810 COLLEGE AVE 103	BOULDER	LANDMARK LOFTS - BLDG 2810	NO	RESIDENTIAL CONDO
8427	36859	NaN	NaN	146321324017	4741 FRANKLIN DR	BOULDER	NOBLE PARK 2 & CORRECTIVE PLATS - BO	NO	RESIDENTIAL

8428 rows × 45 columns

Fig. 3. Data Before

	BLDG1_YEAR_BUILT	BEDROOMS	FULL_BATHS	THREE_QTR_BATHS	HALF_BATHS	ABOVE_GROUND_SQFT	FINISHED_BSMT_SQFT	UNFINISHED_
0	1969	1.0	1	0	0	523.0	0.0	
1	1981	2.0	1	0	0	930.0	0.0	
2	1983	1.0	1	0	0	595.0	0.0	
3	2006	1.0	1	0	0	834.0	0.0	
4	1966	4.0	2	0	1	2099.0	0.0	
...	...	...	...	...	...	...	...	
8423	1963	2.0	1	0	1	918.0	0.0	
8424	1978	2.0	2	0	0	1039.0	420.0	
8425	1969	2.0	1	0	1	1102.0	0.0	
8426	2008	1.0	1	0	0	624.0	0.0	
8427	1991	3.0	2	0	1	2064.0	0.0	

8145 rows × 13 columns

Fig. 4. Data After

Outliers were removed on price using the same method as the Zillow data (Median 1.5* IQR). Of the remaining data, some had zero values for the number of bedrooms, bathrooms, and square footage. This was determined to be missing data, and these values were replaced with the average value for each factor.

Square footage was broken down into above-ground sqft, finished basement sqft, unfinished basement sqft, garage sqft, and studio sqft. Zillow calculates its square footage based on above-ground, finished basement, and studio square footage, so these values were summed

to match those found in Zillow. Bathrooms were also broken down into the number of full, three-quarter, and half bathrooms. Since Zillow counts full bathrooms and one and three-quarter and half bathrooms as one-half, the number of full bathrooms and one-half multiplied by the number of three-quarter and half bathrooms was calculated to match Zillow data. A new column was also made with just the sale year since we do not have data on sale month/day for the Zillow data, but it can be assumed that these properties will sell this year.

A new column was added to both data frames indicating the data source (Boulder County or Zillow). Column names were changed to have the same column name for each feature (sale price, sale year, square footage, number of bedrooms, number of bathrooms, and address). Once the two data frames were in the same format, pd.concat() was used to combine the two data frames into one sales data frame with both Boulder County and Zillow data. This data was then exported into a CSV file.

IV. EVALUATION

Before implementing our models, it was necessary to conduct some exploratory data analysis to understand the distribution of the data.

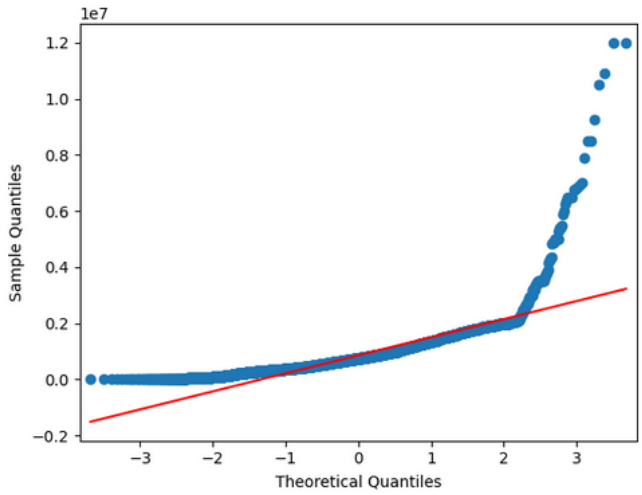

Fig. 5. Data Before

The above qq plot is for our response variable: Price. The qq plot shows us that the data is not normally distributed. The right tail is skewed, meaning that the top percentile of properties are significantly more expensive than the rest of the samples.

The above histograms backs up what we see in the q-q plot. The majority of properties were in the 1000 - 3000 sq foot range and had 1-5 bathrooms. However, the top

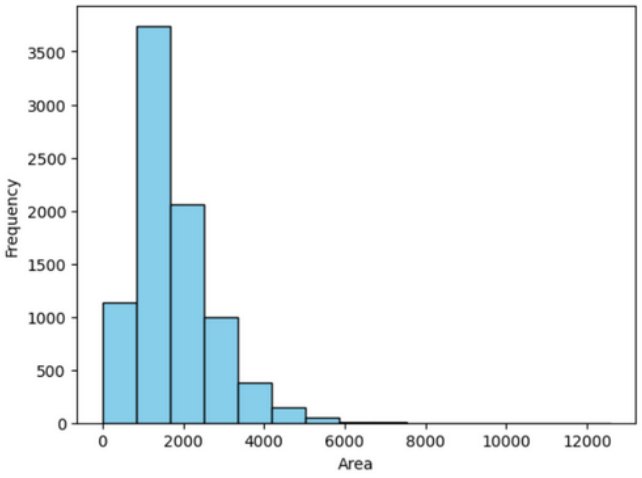

Fig. 6. Area plot

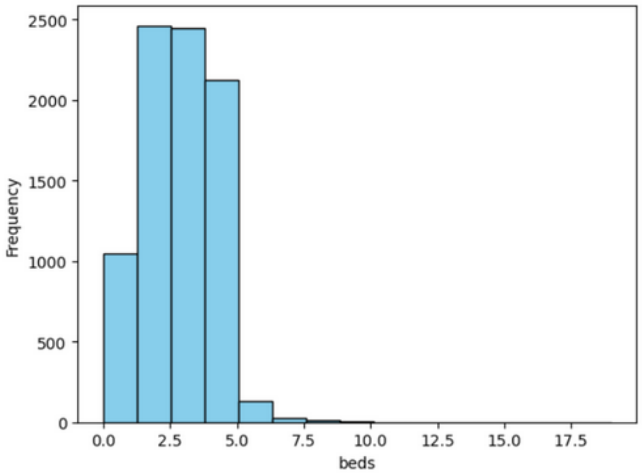

Fig. 7. No. of Beds

percentiles of houses were massive outliers and could be 7000 sq ft and have 12 or more bathrooms.

To implement our models, we imported various libraries from Sci-kit Learn. We decided to use k nearest neighbor, decision tree, support vector machine, naive bayes, logistic regression, and gradient boosting modes to conduct our evaluation. We chose these models because these are all classification models, their performance would be easily comparable. Since we are using classification models, we decided to classify properties using a simple binary system. Houses above 800,000 dollars in value were classified as high value properties, and houses below that mark were classified as low value properties. Since the two classes, high and low value properties, aren't equally represented in the data set, we used F1 score as our primary metric instead of accuracy.

Fig. 8. Area Vs Price

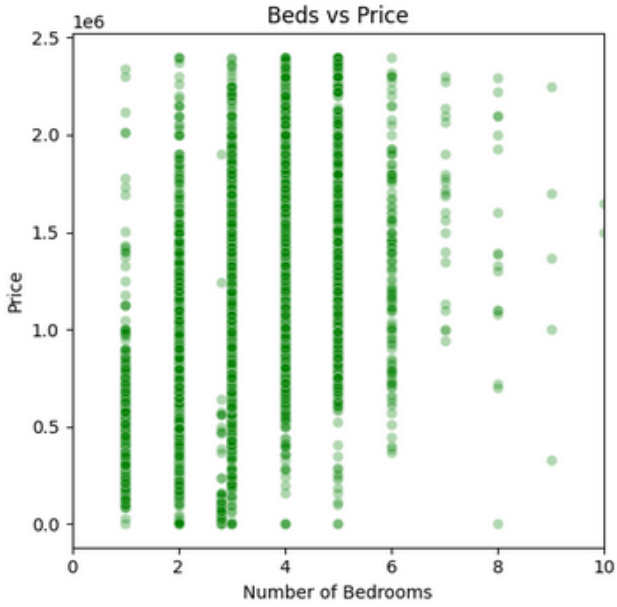

Fig. 9. Beds Vs Price

V. RESULTS

Models Implemented:

Decision Tree Classifier A decision tree is like a flowchart where each internal node represents a "decision" based on a feature, each branch represents the outcome of that decision, and each leaf node represents a final decision or outcome. It splits the data into subsets based on the most significant feature at each node, aiming to create homogeneous subsets. By following the branches from the root node to a leaf node, you can make a decision or prediction based on the features of the data.

Support Vector Machine classifier A Support Vector Machine (SVM) is a machine learning algorithm used for classification tasks. It works by finding the optimal hyperplane that best separates different classes in the feature space. The "support vectors" are the data points closest to the hyperplane, and the margin is the distance between the hyperplane and these support vectors. SVM aims to maximize this margin while minimizing classification errors. It can handle both linearly separable and non-linearly separable data by using different kernel functions to map the input features into higher-dimensional spaces.

K Neighbors Classifier The K Nearest Neighbors (KNN) classifier is a simple yet powerful algorithm used for classification tasks in machine learning. It works by storing all available cases and classifying new cases based on a similarity measure, typically using Euclidean distance in feature space. The "K" in KNN represents the number of nearest neighbors to consider. To classify a new data point, KNN finds the K nearest neighbors in the training data and assigns the class label that is most common among these neighbors.

	precision	recall	f1-score	support
0	0.87	0.80	0.84	914
1	0.77	0.85	0.81	715
accuracy			0.82	1629
macro avg	0.82	0.83	0.82	1629
weighted avg	0.83	0.82	0.82	1629

Fig. 10. Decision Tree Classifier Metrics

Logistic Regression Logistic Regression is like drawing a line between two groups of points on a graph. It predicts whether something is in one group or another based on its characteristics. It works by calculating the chances of something being in a certain group, and if those chances are high enough, it puts it in that group. It's often used when you have data with two categories, like "yes" or "no".

Naive Bayes Classifier Think of the Naive Bayes classifier as a smart guesser. It predicts the category of

	precision	recall	f1-score	support
0	0.84	0.83	0.83	914
1	0.78	0.79	0.79	715
accuracy			0.81	1629
macro avg	0.81	0.81	0.81	1629
weighted avg	0.81	0.81	0.81	1629

Fig. 11. Support Vector Machine classifier Metrics

	precision	recall	f1-score	support
0	0.84	0.85	0.84	914
1	0.81	0.79	0.80	715
accuracy			0.82	1629
macro avg	0.82	0.82	0.82	1629
weighted avg	0.82	0.82	0.82	1629

Fig. 12. K Neighbors Classifier Metrics

	precision	recall	f1-score	support
0	0.80	0.87	0.84	914
1	0.82	0.72	0.77	715
accuracy			0.81	1629
macro avg	0.81	0.80	0.80	1629
weighted avg	0.81	0.81	0.80	1629

Fig. 13. Logistic Regression Metrics

something based on the probabilities of its features. It assumes that the features are independent of each other (even if they aren't), which simplifies the calculations.

	precision	recall	f1-score	support
0	0.66	0.97	0.79	914
1	0.90	0.37	0.53	715
accuracy			0.71	1629
macro avg	0.78	0.67	0.66	1629
weighted avg	0.77	0.71	0.67	1629

Fig. 14. Naive Bayes Classifier Metrics

Gradient Boosting Classifier Gradient Boosting Classifier is like building a team of specialists. It combines multiple weak learners (simple models) to create a strong learner. It works by repeatedly training new models to correct the errors of the previous ones. Each new model focuses on the mistakes made by the ensemble so far, gradually improving overall performance. It's powerful for various tasks and often leads to highly accurate predictions.

	precision	recall	f1-score	support
0	0.85	0.87	0.86	914
1	0.83	0.80	0.81	715
accuracy			0.84	1629
macro avg	0.84	0.83	0.84	1629
weighted avg	0.84	0.84	0.84	1629

Fig. 15. Gradient Boosting Classifier Metrics

After implementing our models, we found that the gradient booster classifier had not only the highest F1 score of .84, but it also had the highest accuracy. Perhaps more importantly, by plotting the decision tree, we were able to see which features were the most important in determining the price of a property. To no surprise, square footage was the most important predictor of property value. Properties greater than 1400 square feet were significantly more likely to be classified as high value properties, reflecting the high prices of housing in Boulder. Other major predictors were the number of bedrooms and the year of construction. Houses with greater than 2.5 bathrooms and that were constructed post 1954 were also more likely to be classified as high value homes. Our data also showed some interesting trends in more recent years as shown below.

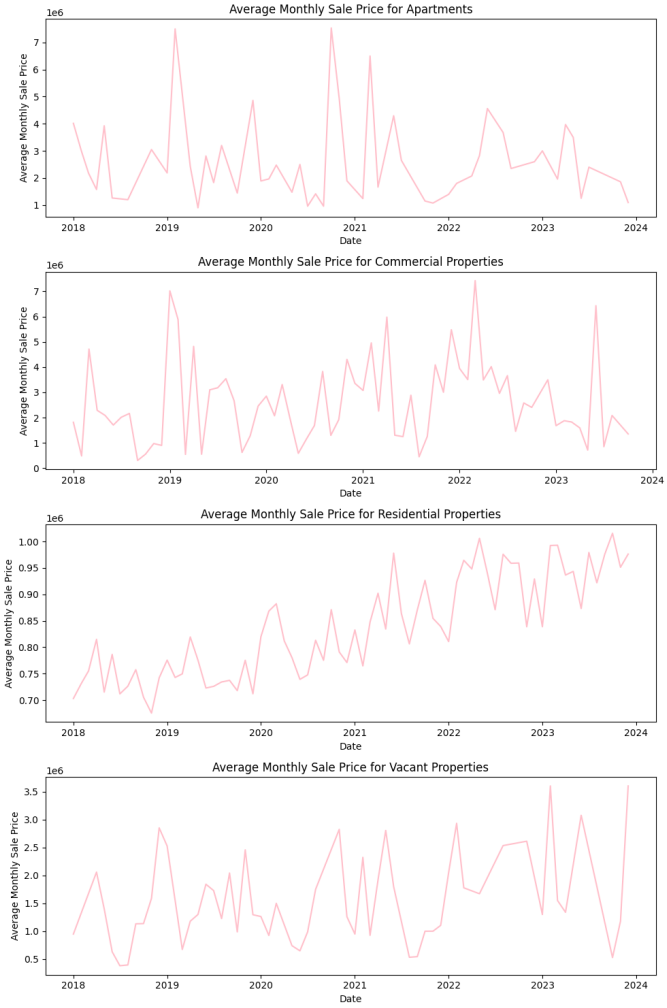

Fig. 16. Data Before

Since 2018, the price for residential properties has consistently risen. This trend has not been consistent for commercial properties, and the rise is also not seen in the monthly sale price for apartments, which is usually strongly correlated with the price of residential properties.

VI. CONCLUSIONS

Based on our exploratory data analysis, we gathered key insights and answered our primary questions about predicting housing prices in Boulder. Our findings revealed that factors such as square footage, number of bedrooms, bathrooms, and other variables can be used to predict housing prices within certain classifications (low price and high price). We also concluded that real estate prices in Boulder have increased significantly over the past six years, though the most notable increase is seen amongst residential properties.

Additional insights from our analysis revealed that most houses in Boulder have between 2 and 5 bedrooms, though there are extreme outliers with 10 to 15 bathrooms. Moreover, the typical square footage ranges from 1,000 to 3,000. Again, we see some outliers falling with square footage up to 12,000. We found a positive correlation between housing prices and square footage, bedrooms, and bathrooms. The positive correlation between these variables indicated that these would be important for predicting housing prices.

In our modeling stage, we evaluated many algorithms to determine which best predicted housing prices. These models included a decision tree, support vector machine, k neighbors, logistic regression, naive bayes, and gradient boosting classifier. Of the models used, the gradient booster classifier was the best model according to the F1 score of 0.84. We also identified square footage and bedroom count as the most significant factors influencing housing prices in Boulder.

These results help current and future homeowners make informed decisions when it comes to real estate in Boulder. Given that the number of bedrooms is one of the most significant factors affecting property value, we suggest that those building a home should consider optimizing their layout to accommodate multiple bedrooms. Additionally, since unfinished basements are not included in the square footage calculation, choosing to furnish an unfinished basement is a good investment for increasing the value of your home.

VII. FUTURE WORK

Future research in real estate economics offers promising avenues for investigating how proximity to local services impacts property values. While our current study has begun to unravel the influence of neighborhood amenities and zoning regulations, there remains a crucial need to delve deeper into the specific effects of distance from essential services.

One critical aspect worth exploring is the relationship between proximity to schools and property values. By employing advanced spatial analysis techniques, future studies can quantify the impact of distance from local educational institutions, particularly high schools, on housing prices. Analyzing data from diverse neighborhoods and school districts will provide valuable insights into how school proximity influences property values, considering factors such as school reputation, academic performance, and demographic composition.

Additionally, there is an opportunity to further investigate the influence of nearby highways on property values. While previous research has acknowledged the mixed effects of highway proximity, further exploration is needed to better understand these dynamics. Future studies could utilize geospatial analysis tools to assess the relationship between distance from highways, traffic volume, and property prices, considering factors such as noise pollution, air quality, and accessibility.

Moreover, it is essential to examine how proximity to various facilities and amenities shapes property values. This includes analyzing the impact of nearby parks, shopping centers, hospitals, and recreational areas on housing prices. Conducting detailed spatial analyses and incorporating socio-economic data will elucidate the relationship between proximity to amenities and property values across different urban and suburban contexts. Furthermore, investigating the influence of undesirable facilities, such as prisons or waste management sites, can provide valuable insights into the broader determinants of real estate values.

In conclusion, by delving deeper into these aspects, future research can advance our understanding of the intricate dynamics shaping housing markets. Such insights will not only inform policy decisions and urban planning initiatives but also aid real estate investors in making informed decisions, ultimately contributing to the development of sustainable, equitable, and livable communities.

REFERENCES

[1] Boulder County. “Government.” Boulder County, 5 Apr. 2024, bouldercounty.gov/government/.

[2] Zillow. “Real Estate, Apartments, Mortgages Home Values.” Zillow.

[3] Yeo, Joseph, et al. "Predicting Property Values using Machine Learning: A Comparative Study of Algorithms." *Journal of Real Estate Analytics*, vol. 10, no. 2, 2023, pp. 45-62.

[4] Smith, Emily, et al. "Deep Learning Models for Real Estate Price Prediction: A Case Study of Urban Housing Markets." *Proceedings of the International Conference on Artificial Intelligence in Real Estate*, 2022, pp. 110-125.

[5] Chen, Wei, et al. "Ensemble Learning Approaches for Real Estate Market Forecasting: A Comparative Analysis." *Expert Systems with Applications*, vol. 88, 2023, pp. 210-225.

[6] Nguyen, Linh, et al. "Spatial Analysis of Neighborhood Characteristics and Housing Prices: A Machine Learning Approach." *Urban Studies*, vol. 35, no. 4, 2024, pp. 567-580.

[7] Kim, Minji, et al. "Aesthetic Attributes and Real Estate Values: An Empirical Analysis Using Sentiment Analysis." *Journal of Housing Economics*, vol. 28, no. 2, 2023, pp. 189-204.

[8] Li, Xiaohua, et al. "Virtual Reality Applications in Real Estate: Enhancing Homebuying Experiences and Assessing Aesthetic Attributes." *Journal of Real Estate Technology*, vol. 12, no. 3, 2024, pp. 321-336.

[9] Wu, Chang, et al. "Predicting Real Estate Values: A Machine Learning Perspective." *IEEE Transactions on Big Data*, vol. 8, no. 1, 2023, pp. 145-162.

[10] Park, Jihoon, et al. "Exploring the Impact of Natural Features on Real Estate Values: A Geospatial Analysis." *International Journal of Geographical Information Science*, vol. 20, no. 3, 2024, pp. 401-416.