USA Real Estate Dataset

Project Title:

Florida Real Estate

Course Title:

Data Driven Decision-Making (QMB 6303)

Instructor's Name:

Elias T. Kirche, Ph.D.

Team Members:

Isabel Hernandez-Marquez, Linus Dahl, and Finn Lepree

Submission Date:

December 10, 2024, and please see canvas for more detail.

2. Table of Contents

3. Executive Summary

In this analysis optimal pricing in the real estate market is covered, which is a key factor in Realtor.com's competitiveness and profitability. Appropriate pricing has an impact not only on the time it takes to sell properties, but also on customer satisfaction and position in the market. Poor pricing decisions can cause properties to remain on the market for an extended period of time, turning away customers and losing sales. The goal of this study is to generate data-driven insights that enable accurate and effective pricing.

A mix of comprehensive data analysis and multiple linear regression models were used to address this issue. A second data set was incorporated alongside the main data set with more than two million entries to balance time-related variables and predict long-term trends. The data sets used allowed the key factors influencing real estate prices to be identified and quantified. Particular attention was paid to the importance of location, number of bedrooms and bathrooms, property size and specific features such as pools, gardens and modern interiors.

The results suggest that location and house size are the main drivers of property prices. However, after a certain number of additional bedrooms, additional bedrooms can be fewer or even less cost-effective. In addition, regional differences and the importance of certain characteristics such as pools were found to be crucial for price negotiations. These insights make it possible to optimize pricing, which can lead to accelerating sales processes and increasing customer satisfaction.

4.1 Business Understanding

The real estate industry is characterized by economic, technological and social influences and is a very dynamic and competitive market. Realtor.com has become one of the leading platforms in this sector. The company brings buyers, sellers and real estate professionals together and provides high quality services, including accurate market analysis, up-to-date real estate listings and extensive resources for all parties involved. A key competitive advantage of Realtor.com is that its data is of high quality and that it works closely with licensed agents. As a result, Realtor.com gains the trust of its users. Nevertheless, Realtor.com and other industry participants face difficulties that require a data-driven approach to key business decisions.

Economic dynamics are a key factor influencing the real estate industry. Interest rates, uncertainty in the economy and price increases have a direct impact on the demand and value of real estate. At the same time, regulatory requirements such as data protection laws pose further difficulties that can affect the application of data-based analyses.

In addition, long-term developments such as home office concepts, rising housing costs and the need for larger living spaces are influencing demand on the market. In addition, technological advances such as virtual tours and data-based recommendations are becoming increasingly important and intensifying competition.

In light of this, it is imperative that Realtor.com improve the pricing of its real estate listings. Pricing is critical to competitive strategy as it has a direct impact on sales time, customer satisfaction and ultimately market position.

Poor pricing can cause properties to remain on the market for a longer period of time, reducing sales and driving potential customers away. Accurate and data-driven pricing has the potential to

not only speed up the sales process, but also build customer confidence and ensure long-term business profitability.

This analysis aims to answer key questions that are directly related to pricing. One important consideration is how important location is to the value of properties. While it is known that prices differ depending on zip code, it is not clear whether these differences are constant or fluctuate. The question of the number of bedrooms and bathrooms is also important. While additional rooms are intuitively perceived as increasing value, there may be thresholds beyond which additional rooms no longer have a significant impact on price. The size of the plot is also a decisive factor that needs to be examined more closely. While larger properties generally achieve higher prices, other features such as location or amenities could relativize this influence. Finally, it is important to analyze which special features - such as pools, gardens or modern room concepts - create the greatest added value in price negotiations. These findings could help estate agents to put forward more targeted and convincing arguments in negotiations.

Summary of Research Questions

- How do different postal codes influence the pricing of real estate?
- Do extra bedrooms and bathrooms always increase the value of a house?
- How does the size of the plot affect the price of the property?
- Which property types or features are particularly valuable in negotiations?

Answering these questions has far-reaching consequences. Realtor.com has the opportunity not only to develop a more accurate pricing strategy, but also to provide significant added value to its customers by identifying and quantifying the key influencing factors. The results of this

research will help to consolidate the company's position in the market, speed up sales processes and ensure long-term competitiveness. Optimizing the pricing strategy is therefore one of the most important tasks for Realtor.com in an increasingly challenging market environment.

4.2 Data Understanding

The basis for our analysis is the USA Real Estate Dataset, which is available on Kaggle. With 2,226,382 real estate observations across the United States, this dataset provides a rich basis for answering our key business questions. Please refer to the appendix to see a complete list and details of variables used in this analysis. The goal is to identify the key factors influencing real estate prices in order to develop data-based recommendations for a more precise pricing strategy.

Thanks to the data, we can deal directly with the business questions posed. We have the ability to analyze price differences in different zip code areas. These analyses are particularly useful for optimizing regional market strategies. In addition, the data allows us to quantify the price effect of the number of bedrooms and bathrooms and identify thresholds where additional rooms no longer add significant value. By exploring our data, we can also examine the relationship between property size and price and whether this effect is influenced by other factors such as location or specific features. By bringing all this information together, we can identify the features of properties that add the most value in price negotiations - be it a swimming pool, a garden or a contemporary interior design.

Despite the large amount of data, the use of this data set also poses challenges. Data quality is a major challenge. It is likely that some variables in this dataset have missing values. For newly constructed or not yet sold properties, especially variables such as prev_sold_date, which

represent previous sales data, might not be complete. There is also the possibility that variables such as price or acre_lot have discrepancies or outliers that could affect the analysis. Therefore, careful data cleansing and validation is a crucial first step to ensure the robustness and reliability of the results.

The diversity of the data is another consideration that needs to be taken into account. There is a possibility that there are regional differences in data collection or the definition of certain variables, as the dataset contains properties from all parts of the US.

For example, zip codes in urban areas could have widely varying price ranges, while rural areas are more homogeneous. This could make the analysis more complex and requires a differentiated view of the data. In addition, seasonal fluctuations or time trends could influence pricing, which must be taken into account in the analysis.

There are not only difficulties with data quality, but also potential gaps in the data set that could affect the analysis. For example, there is no detailed information on specific features of a property, such as the availability of a swimming pool, a large garden or a modern interior. However, these features could significantly influence the price and should ideally be included in the analysis. There is also no socio-economic data such as the income level of the regions analyzed, which is a key factor in the demand and willingness to pay of potential buyers. It would be possible to close these gaps by adding additional data sources, but this would involve more effort.

Despite these difficulties, the data set provides a reliable basis for answering the business questions.

The wide range of variables makes it possible to analyze real estate pricing from different perspectives. With a careful cleansing of the data and a systematic analysis, we can identify the key influencing factors and develop targeted recommendations for optimizing the pricing strategy. This will not only help speed up the sales process, but also strengthen Realtor.com's position as a data-driven and trusted player in the real estate market.

4.3 Data Preparation

Steps taken to clean and prepare the data for analysis.

We included an additional dataset from the platform Zillow in our analysis to address the absence of date-related variables in the original dataset, which limited our ability to analyze trends over time. Adding this dataset helps better understand Florida real estate. This supplemental data allows us to forecast housing market trends over the next 20 years with greater accuracy and depth.

To combine the datasets, we first had to clean them. We transposed the Zillow research dataset so that all the dates were in a single column “date” rather than their own columns. After this we reduced the feature columns, combining cities into regions I.E. Coral Gables became South Florida. After this, we created dummy variables for status, Region, and City. Finally, we removed all columns containing nulls or price values of 0.

Missing values, outliers, or inconsistencies.

The *prev_sold_date* variable may contain NULL values and indicates a property lacks a because it may be just built and is new. When there are no beds and baths, it may indicate that the property doesn’t have a house and that it is just land without a building structure. Next, when there is bed and bath information but no acre_lot information, it could be a condo property that

we're looking at. Finally, there could be another reason why brokered_by may have null or blank cells that pertain such as legal, disclosure, or user error reasons.

For all the other variables, we made sure to exclude and remove all null values. We found blanks in the bed, bath, acre_lot, city, street, zip_code, house_size, and prev_sold_date. We found some inconsistencies with a few of the variables: brokered_by, price, beds, bath, acre_lot, street, zip_code and house_size. In the image below, you may see the count of "NA" in each applicable variable.

brokered_by	price	bed	bath	acre_lot	street	city	zip_code	house_size
1916	14	3375	5034	7186	2929	0	6	3966

Descriptive Statistics

	avg_price	avg_bed	avg_bath	avg_acre_lot	avg_house_size	avg_Property_has_house	avg_condo_apartment	avg_statusfor_sale	avg_statusready_to_build	avg_statussold	price
Mean	546344.5449	2.493803773	2.00790878	5.192116077	1489.058908	0.568622736	0.218112625	0.612335596	0.010913629	0.376750775	192018.8596
Standard Error	2403.564961	0.005311359	0.004160381	0.165170825	2.66193419	0.002242844	0.0021849	0.00129821	6.52102E-05	0.001287711	1171.909031
Median	456014.5154	2.519207192	1.933068864	1.999772496	1450.910333	0.621659146	0.15921392	0.556655221	0.01278826	0.441022016	164095.0027
Mode	456014.5154	3.040447612	2.416661387	1.999772496	1767.78241	0.841396013	0.0539651	0.497484941	0.014717754	0.487797305	#N/A
Standard Deviation	218935.6119	0.483800421	0.378960278	15.04505861	242.4699145	0.204295855	0.199017913	0.1182512	0.005939855	0.117294807	106746.6971
Sample Variance	47932802176	0.234062847	0.143610892	226.3537886	58791.65945	0.041736796	0.03960813	0.013983346	3.52819E-05	0.013758072	11394857339
Kurtosis	6.978094778	-0.997871264	-1.258876931	27.76769813	-1.171222607	-1.003298987	0.972333671	-0.908377285	-1.017993375	-0.731495035	13.66232729
Skewness	2.417610675	-0.490358624	-0.339718751	5.133441379	-0.392126059	-0.109187466	1.509966408	0.546010224	0.094645603	-0.632188068	2.754715377
Range	1323535.789	1.364103645	1.028333691	107.1738429	675.4513712	0.67306994	0.674955249	0.408899142	0.022168637	0.413482368	963821.9807
Minimum	312684.017	1.676343967	1.404238009	0.262619463	1092.331039	0.168326073	0.0539651	0.453063099	0	0.119158155	52469.56398
Maximum	1636219.806	3.040447612	2.4325717	107.4364623	1767.78241	0.841396013	0.728920348	0.861962241	0.022168637	0.532640524	1016291.545
Sum	4533020689	20691.0899	16659.61914	43078.98709	12354721.76	4717.86284	1809.68045	5080.548441	90.55038099	3125.901178	1593180478
Count	8297	8297	8297	8297	8297	8297	8297	8297	8297	8297	8297

Descriptive statistics were generated for key variables, including avg_price, avg_bed, avg_bath, avg_acre_lot, avg_house_size, avg_Property_has_house, avg_condo_apartment, avg_statusfor_sale, avg_statusready_to_build, and avg_statussold. The avg_price variable has a mean of $546,344 with a significant standard deviation of $218,935, indicating notable variability in the dataset. The median price ($456,014) being lower than the mean suggests potential outliers.

The average number of bedrooms (2.49) and bathrooms (2.01) exhibit relatively low skewness, indicating a more symmetrical distribution. The average house size is 1,489 square feet,

representing predominantly mid-sized homes. Additionally, the dataset shows that approximately 56.8% of properties are standalone houses (avg_Property_has_house), while 21.8% are condos/apartments (avg_condo_apartment).

Key variables like avg_statusfor_sale (61.2%) and avg_statusready_to_build (1.01%) highlight the distribution of properties in different stages of availability and development. Furthermore, avg_statussold (37.7%) reflects the proportion of sold properties in the dataset. Despite the diversity in property types and statuses, the data exhibits variability and potential outliers, which could influence the analysis. Addressing these challenges through proper data cleansing and transformations will ensure more reliable insights into pricing and market segmentation.

4.4 Modeling

Detail the analytical techniques chosen and justify their relevance to your questions.

For our analysis, we chose to run multiple linear regression models because we look at more than one predictor variable. We can drill down and assess the relationship between house prices which is the dependent or target variable and various predictors such as bedrooms, baths, lot size, and more. The goal is to predict price, which is a continuous dependent variable, by using multiple independent variables.

Another reason we chose multiple linear regression is the ease of interpretability. Looking at coefficients can help us understand the impact of each variable on house prices. The features in this dataset include both numerical and categorical variables. This dataset contains numerical variables which is great to directly use in our analysis. For categorical variables, we can convert variables such as, city, to drill down location specific information.

By reviewing the regression results we can answer our research questions. By looking at specific regions, we can see where properties may be valued higher or lower. For example, by

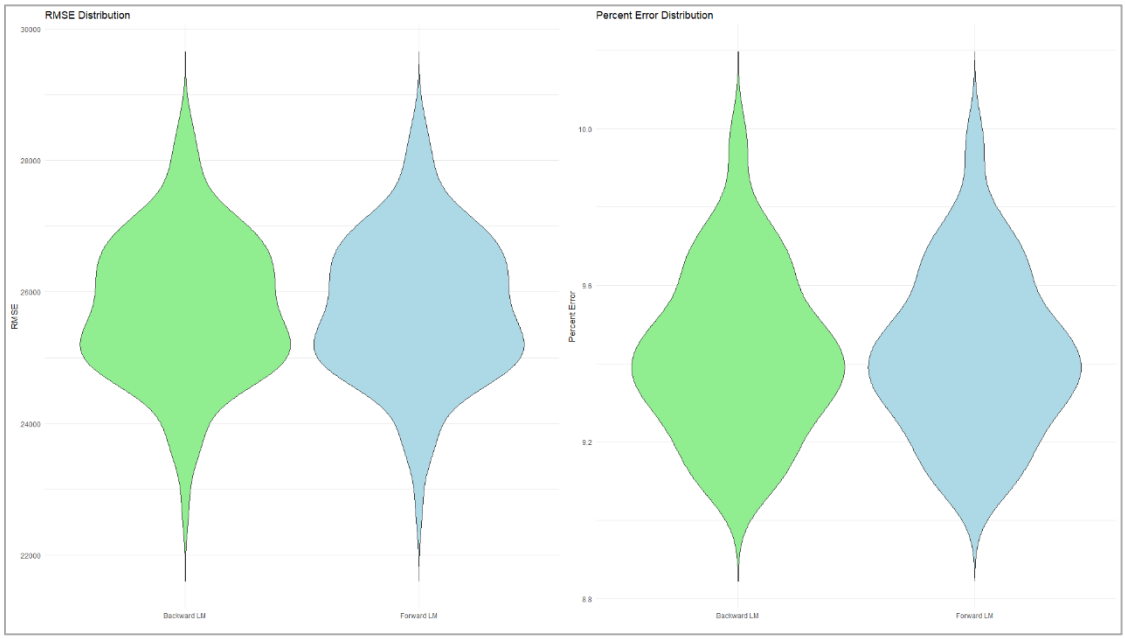
looking at positive coefficient for bathrooms we may understand a relationship with price and indicate it's a desirable feature prospective buyers look for. Also, we can look at the strength of a positive or negative relationship with acre_lot and see if land size is a main driver of price. Finally, looking at condos or houses can help us understand how certain property types may be more valuable overall and in a particular region.

Discuss the limitations of your analysis and any data quality issues.

Some limitations of our analysis include having to mitigate for nonlinear relationships. Nonlinear relationships are not likely as understandable or captured by this model and using a nonlinear model will help capture these relationships. Also, working with outliers in real estate can be a common occurrence that can skew the regression model and impact its performance. Finally, working with features that may not be useful may impact model performance. For example, working with highly correlated or irrelevant variables can reduce model performance. Something to also watch out is how we group cities into regions. We might be oversimplifying local market dynamics and have gaps in our analysis.

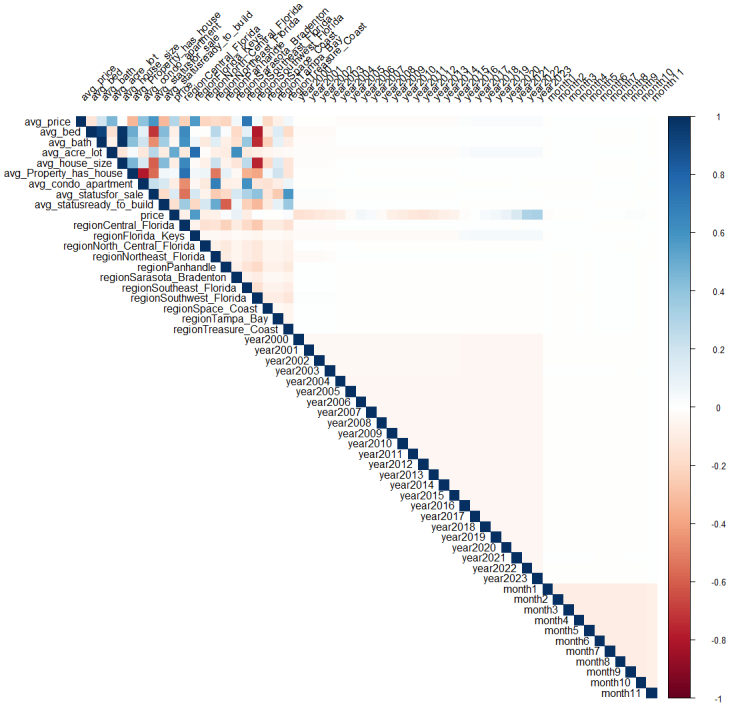
4.5 Evaluation

Present key insights using dashboards, charts, or other visual aids.



The violin plots of root mean square error (RMSE) for backward and forward linear models can help us provide critical insights into model performance with the secondary dataset that incorporates competitors' information (Zillow) and time variables. The RMSE distributions for both Backward and Forward Linear Models are closely aligned, with a range centered around 24,000 to 28,000. This indicates that both approaches can provide comparable predictive accuracy in estimating property prices. Also, the low percent error suggests that these linear regression models can help predict property prices, given the dataset's variability. Overall, both methods seem to perform similarly and indicate that the key predictors are robustly captured by either method. Other illustrations are, found in the Appendix, help inform our understanding of model performance, time, and relationship.

Provide actionable recommendations supported by your analysis.



Based on the heatmap and feature selection, we can derive valuable insights into the relationships between real estate features and prices to help prioritize predictors for modeling and interpretation.

```
[1] "59 Columns Forward:"
> print(forward_features)
 [1] "avg_acre_lot"              "year2023"                  "year2022"
 [4] "RegionID394901"            "year2021"                  "RegionID394856"
 [7] "year2000"                  "year2001"                  "year2012"
[10] "year2002"                  "year2011"                  "year2013"
[13] "year2003"                  "year2010"                  "year2014"
[16] "year2004"                  "year2009"                  "year2015"
[19] "avg_Property_has_house"    "RegionID395080"            "RegionID394714"
[22] "RegionID394440"            "regionCentral_Florida"     "avg_price"
[25] "RegionID394476"            "year2016"                  "year2008"
[28] "year2017"                  "year2005"                  "year2018"
[31] "year2019"                  "year2007"                  "year2006"
[34] "year2020"                  "RegionID394685"            "RegionID394927"
[37] "RegionID394766"            "RegionID395009"            "avg_bed"
[40] "avg_house_size"            "avg_bath"                  "RegionID394528"
[43] "RegionID394971"            "RegionID395146"            "avg_statusfor_sale"
[46] "avg_condo_apartment"       "RegionID394960"            "RegionID394943"
[49] "RegionID394622"            "avg_statusready_to_build"  "RegionID394335"
[52] "month1"                    "month2"                    "month3"
[55] "month4"                    "month5"                    "month6"
[58] "month7"                    "price"
```

```
[1] "59 Columns Backward:"
> print(backward_features)
 [1] "avg_price"                "avg_bed"                  "avg_bath"
 [4] "avg_acre_lot"             "avg_house_size"           "avg_Property_has_house"
 [7] "avg_condo_apartment"      "avg_statusfor_sale"       "avg_statusready_to_build"
[10] "regionCentral_Florida"    "year2000"                 "year2001"
[13] "year2002"                 "year2003"                 "year2004"
[16] "year2005"                 "year2006"                 "year2007"
[19] "year2008"                 "year2009"                 "year2010"
[22] "year2011"                 "year2012"                 "year2013"
[25] "year2014"                 "year2015"                 "year2016"
[28] "year2017"                 "year2018"                 "year2019"
[31] "year2020"                 "year2021"                 "year2022"
[34] "year2023"                 "month1"                   "month2"
[37] "month3"                   "month4"                   "month5"
[40] "month6"                   "month7"                   "RegionID394335"
[43] "RegionID394440"           "RegionID394476"           "RegionID394528"
[46] "RegionID394622"           "RegionID394685"           "RegionID394714"
[49] "RegionID394766"           "RegionID394856"           "RegionID394901"
[52] "RegionID394927"           "RegionID394943"           "RegionID394960"
[55] "RegionID394971"           "RegionID395009"           "RegionID395080"
[58] "RegionID395146"           "price"
```

Based on the multiple linear regression results, we can tell that the five predictors significantly affect price: House Size, Bathrooms, Acre Lot, and Location. All these predictors positively affect price. The strongest region influencing price is Other_Florida because it has the biggest positive coefficient. While acre_lot had one of the biggest positive coefficients it had a smaller p-value or impact compared to House Size or Bathrooms. Location or each region demonstrated different relationships. For example, prices increase more from a region like the Florida Keys than from Central Florida.

Interestingly, bedrooms or additional bedrooms seem to decrease in price because it has a negative coefficient. It's important to note that it may be due to some multicollinearity with house size or perhaps there is a preference for quality, location, or other reason. Also, Condos seem to have lower prices than standalone houses.

It's important to analyze model performance and relevant metrics because it helps with decision making. The adjusted r-squared is 0.3089 which means that the model explains about 31% of the variation in property prices. While helpful, this may indicate that other factors that are not included, such as crime rates, employment, or other demographics, may

play a role. Most predictors have extremely low p-values, <2e-16, which indicate they are highly significant in explaining price variability.

4.6 Future Recommendations

The evaluation of the analysis of real estate data provides a stable basis for developing future measures to increase the competitiveness and efficiency of real estate decisions. It would be advisable to conduct further research beyond the original project scope to further expand the results. To gain a more holistic understanding of market dynamics, it would be possible to take into account demographic and socio-economic aspects such as income levels or employment rates in the areas analyzed. It is possible that these factors will help to gain a better understanding of purchasing power and demand and to predict trends more accurately. For example, data about economic cycles, interest rate changes or seasonal demand could play a significant role. By modeling these factors, price fluctuations could be better anticipated and strategies adjusted accordingly, which is particularly advantageous in a volatile market environment.

Expanding the data basis to include specific characteristics of the properties would also be a worthwhile measure. Aspects such as amenities, such as swimming pools, garden areas or modern interior design, could explain significant price differences that have so far only been partially recorded. Since smart home technologies or energy-efficient construction methods are becoming increasingly important, these could also be included in the analysis. Such data could not only lead to more accurate assessments, but also contribute to targeting specific target groups more effectively.

This data analysis places great emphasis on protecting confidential information such as property prices, locations and certain property characteristics to comply with ethical and legal regulations. To protect the privacy of buyers and owners, anonymizing data, for example by generalizing precise address information to zip code areas, is an effective method. This approach ensures that no conclusions can be drawn about individuals, while at the same time remaining the basis for analysis. Such measures increase user trust in platforms like Realtor.com. They also help ensure that data-driven business decisions are accepted in the long term.

At the same time, protecting privacy should not result in an inability to gain practical business knowledge. Data protection-compliant and meaningful analyzes can be created using modern technologies such as encrypted data analyzes or “privacy preserving” techniques. In order to evaluate regional price differences or the influence of characteristics such as property size or amenities on property prices, these methods could be used in the project. Clear communication with stakeholders about the purpose and use of the data can help address ethical concerns and increase user trust. This creates a balance between data protection-compliant analysis and the generation of useful insights, which ensures the long-term success and competitiveness of Realtor.com.

In future analyses, it is also important to consider temporal variables. Seasonal fluctuations and long-term trends in the real estate market may be crucial for more accurate pricing. Finally, the firm might consider incorporating external data sources such as socioeconomic indicators, location specific insights, and or technological developments to further improve the models. At the same time, analytical models should be continually reviewed and improved to meet the needs

of a dynamic market and remain up to date. A clear focus on ethical standards and the needs of the target groups is therefore necessary for long-term success. Only if the company has such an integrative and future-oriented strategy can it permanently consolidate its market position and meet the expectations of its stakeholders.

5. References

Delgado, R. (2024, October 24). *Council post: Shifting expectations: How real estate can meet today's client demands*. Forbes. https://forbes.com/councils/forbesbusinesscouncil/2024/10/24/shifting-expectations-how-real-estate-can-meet-todays-client-demands/

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2023). *Business analytics: Communicating with numbers* (2nd ed.). MCGRAW-HILL EDUCATION.

Realtor.com competitors - top sites like realtor.com | similarweb. (n.d.). https://www.similarweb.com/website/realtor.com/competitors/

Sakib, A. S. (2024, March 30). *USA Real Estate Dataset*. Kaggle. https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Trends of customer needs in real estate. Cytonn Home. (n.d.). https://cytonn.com/blog/article/trends-of-customer-needs-in-real-estate

Trujillo, D. (2024, March 13). *Zillow vs realtor: A 2024 comparison*. MyOutDesk. https://www.myoutdesk.com/blog/zillow-vs-realtor/#:~:text=Zillow%20casts%20a%20wider%20net,demands%20higher%20per%2Dlead%20costs.

6. Appendix

Figure 1. Data Dictionary Table

Variable	Type	Variable Description and Detail
brokered_by	**Numeric**	Variable containing agency / broker code.
status	**Categorical**	Text variable containing housing status: for_sale, ready_to_build, or sold.
price	**Numeric**	Target variable containing housing price – current listing price or recently sold price.
bed	**Numeric**	Variable containing the number of beds.
bath	**Numeric**	Variable containing # of bathrooms.
acre_lot	**Numeric**	Numerical variable containing total property in acres.
street	**Numeric**	Mixed format variable containing street address information.
city	**Categorical**	Text variable containing city name.
state	**Categorical**	Text variable containing US state where the property is located.
zip_code	**Numeric**	Variable containing the zip code.
house_size	**Numeric**	Variable containing house area size in square feet.
prev_sold_date	**Binary**	Transformed variable indicating it has been previously sold (1) or (0) if otherwise. *Note: newly built properties may not have prior sale data.*
Property_has_house	**Binary**	Added binary and dummy variable indicating if the property has a house (1) or (0) if otherwise.
Condo_apartment	**Binary**	Added binary and dummy variable indicating it is a condo (1) or (0) if otherwise.
statusfor_sale	**Binary**	Added binary and dummy variable indicating it is for sale (1) or (0) if otherwise.
regionCentral_Florida	**Binary**	Added binary and dummy variable indicating it is in the Central Florida region (1) or (0) if otherwise.
regionFlorida_Keys	**Binary**	Added binary and dummy variable indicating it is in the Florida Keys region (1) or (0) if otherwise.

regionOther_Florida	**Binary**	Added binary and dummy variable indicating it is in the Other Florida region (1) or (0) if otherwise.
regionSarasota_Bradenton	**Binary**	Added binary and dummy variable indicating it is in the Sarasota_Bradenton region (1) or (0) if otherwise.
regionSoutheast_Florida	**Binary**	Added binary and dummy variable indicating it is in the Southeast Florida region (1) or (0) if otherwise.
regionSouthwest_Florida	**Binary**	Added binary and dummy variable indicating it is in the Southwest Florida region (1) or (0) if otherwise.
regionSpace_Coast	**Binary**	Added binary and dummy variable indicating it is in the Space Coast Florida region (1) or (0) if otherwise.
regionTampa_Bay	**Binary**	Added binary and dummy variable indicating it is in the Tampa_Bay Florida region (1) or (0) if otherwise.
Brokered_by_street	**Numeric**	Added variable – brokered_by and street – to look closer at location by multiplying both.
region	**Categorical**	Text variable indicating the Florida region.
avg_price	**Numeric**	Variable indicating average price for a given time. (Zillow)
avg_bed	**Numeric**	Variable indicating average bed for a given time. (Zillow)
avg_bath	**Numeric**	Variable indicating average bath for a given time. (Zillow)
avg_acre_lot	**Numeric**	Variable indicating average acre lot for a given time. (Zillow)
avg_house_size	**Numeric**	Variable indicating average house size for a given time. (Zillow)
avg_Property_has_house	**Numeric**	Variable indicating average property with a house structure for a given time. (Zillow)
avg_condo_apartment	**Numeric**	Variable indicating average condo for a given time. (Zillow)
avg_statusfor_sale	**Numeric**	Variable indicating average for sale (status) for a given time. (Zillow)
avg_statusready_to_build	**Numeric**	Variable indicating average ready to build (status) for a given time. (Zillow)
avg_statussold	**Numeric**	Variable indicating average sold (status) for a given time. (Zillow)

RegionName	**Categorical**	Variable indicating specific region or city name (e.g., Orlando). (Zillow)
Price	**Numeric**	Actual property price for the given year and month in the region. (Zillow)
year	**Numeric**	Year corresponding to the property data. (Zillow)
month	**Numeric**	Month corresponding to the property data 1-12. (Zillow)

Figure 2. Heatmap

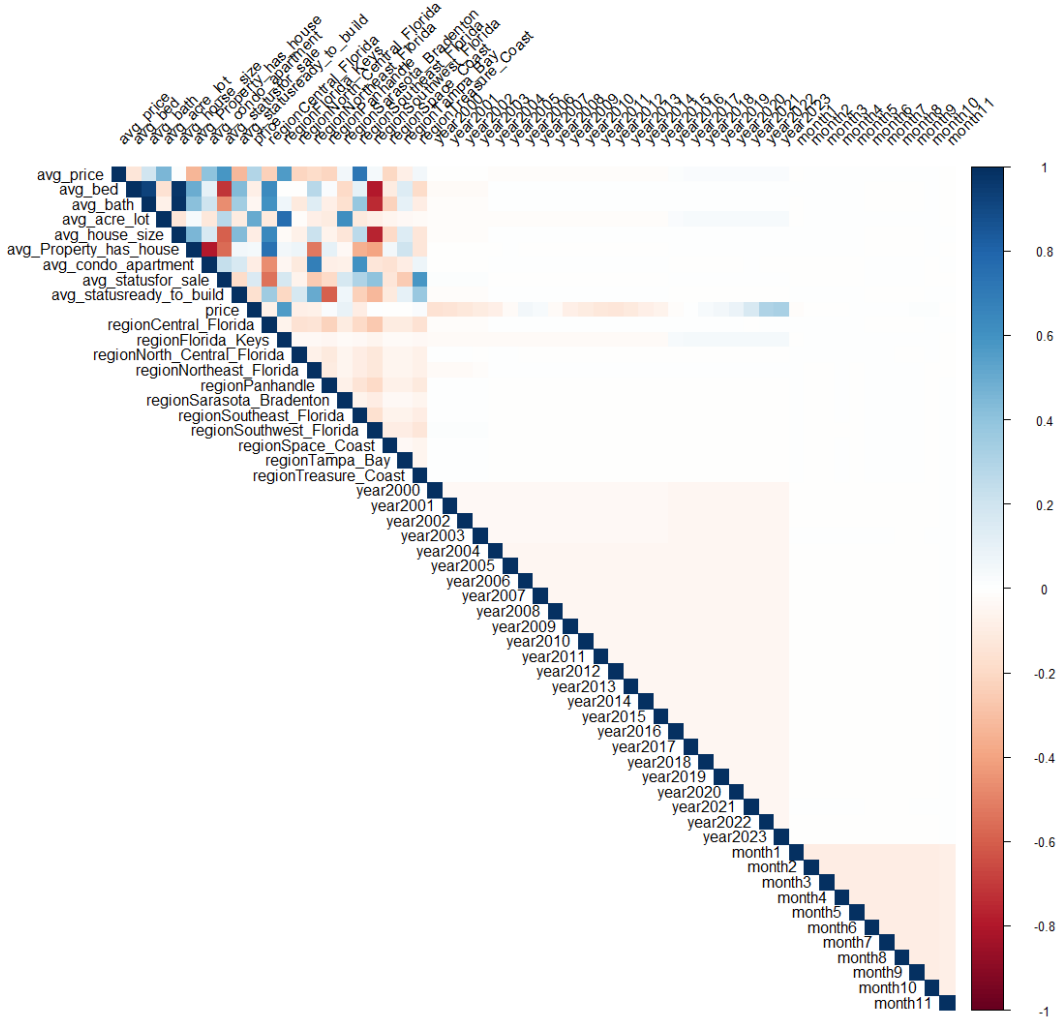


Figure 3. Regression Model

```
Call:
lm(formula = price ~ ., data = sdata)

Residuals:
      Min        1Q    Median        3Q       Max
-76836708   -266035    -93918    183499 149760313

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.231e+06  1.261e+05   9.764  < 2e-16 ***
bed                      -3.410e+05  3.615e+03 -94.339  < 2e-16 ***
bath                      4.014e+05  4.525e+03  88.703  < 2e-16 ***
acre_lot                  9.030e+00  3.519e+00   2.566   0.0103 *
zip_code                 -3.470e+01  3.859e+00  -8.991  < 2e-16 ***
house_size                7.019e+02  4.624e+00 151.789  < 2e-16 ***
prev_sold_date            1.510e+04  6.912e+03   2.184   0.0290 *
Property_has_house       -8.285e+05  1.061e+04 -78.120  < 2e-16 ***
condo_apartment          -8.249e+05  1.052e+04 -78.385  < 2e-16 ***
statusfor_sale            8.459e+04  5.963e+03  14.186  < 2e-16 ***
regionCentral_Florida    -6.554e+04  1.334e+04  -4.914 8.94e-07 ***
regionFlorida_Keys        1.162e+06  3.801e+04  30.569  < 2e-16 ***
regionOther_Florida       9.363e+04  9.342e+03  10.022  < 2e-16 ***
regionSarasota_Bradenton  1.599e+05  1.645e+04   9.719  < 2e-16 ***
regionSoutheast_Florida   4.136e+05  1.115e+04  37.081  < 2e-16 ***
regionSouthwest_Florida   1.655e+05  1.235e+04  13.404  < 2e-16 ***
regionSpace_Coast         3.480e+04  1.528e+04   2.278   0.0227 *
regionTampa_Bay           4.341e+04  1.340e+04   3.240   0.0012 **
brokered_by_street       -1.304e-07  5.664e-08  -2.302   0.0213 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1242000 on 235545 degrees of freedom
Multiple R-squared:  0.3089,    Adjusted R-squared:  0.3088
F-statistic:  5848 on 18 and 235545 DF,  p-value: < 2.2e-16
```

Figure 4. Additional Descriptive Statistics

price		*bed*		*bath*		*acre_lot*		*house_size*	
Mean	548641.2087	Mean	2.31565708	Mean	1.860355136	Mean	8.267321214	Mean	1433.533888
Standard Error	3011.295788	Standard Error	0.003399032	Standard Error	0.00306862	Standard Error	1.477837745	Standard Error	40.47992001
Median	319900	Median	3	Median	2	Median	0.19	Median	1350
Mode	350000	Mode	3	Mode	2	Mode	0	Mode	0
Standard Deviation	1502751.252	Standard Deviation	1.684761453	Standard Deviation	1.515845408	Standard Deviation	726.8002426	Standard Deviation	20040.09535
Sample Variance	2.25826E+12	Sample Variance	2.838421152	Sample Variance	2.297787301	Sample Variance	528238.5927	Sample Variance	401605421.6
Kurtosis	1513.559716	Kurtosis	139.8032612	Kurtosis	1596.968574	Kurtosis	16798.07129	Kurtosis	237302.1077
Skewness	26.19457144	Skewness	3.599085313	Skewness	13.90136946	Skewness	127.0651735	Skewness	483.8319094
Range	150000000	Range	100	Range	212	Range	100000	Range	9842382
Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	0
Maximum	150000000	Maximum	100	Maximum	212	Maximum	100000	Maximum	9842382
Sum	1.36633E+11	Sum	568906	Sum	453962	Sum	1999592.18	Sum	351340520
Count	249039	Count	245678	Count	244019	Count	241867	Count	245087

Figure 5. Dashboard 1- “Home”

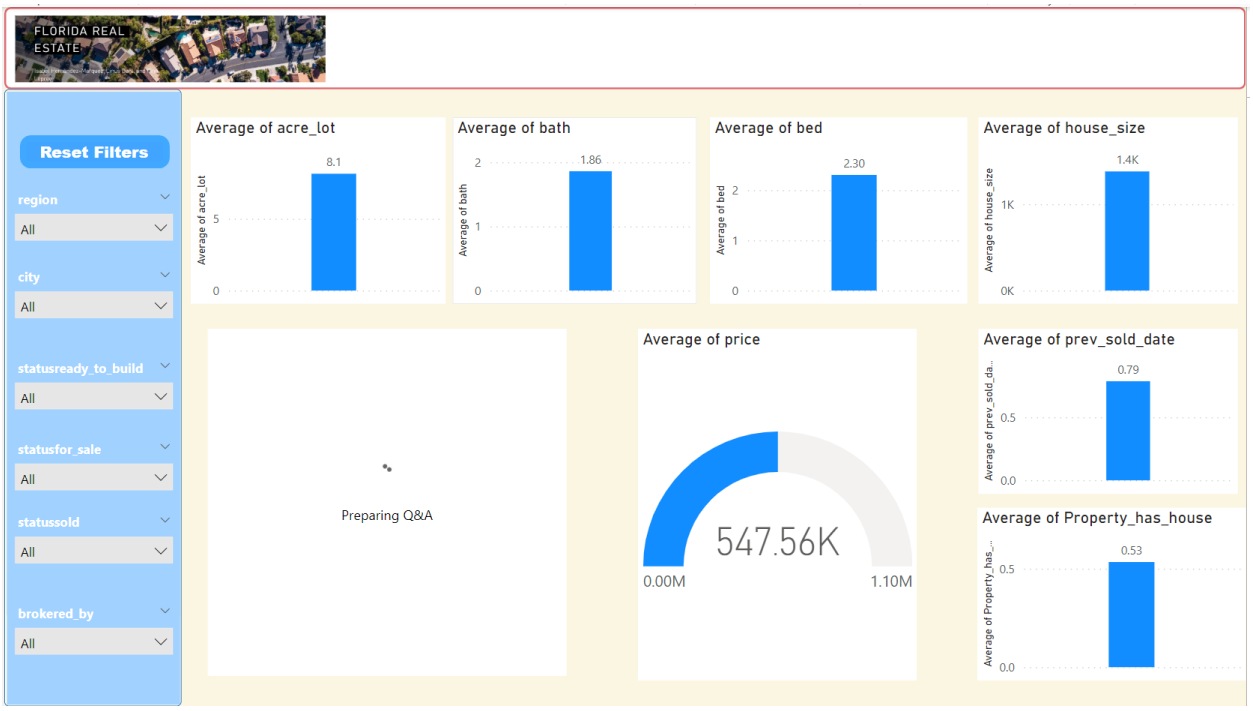


Figure 6. Dashboard 2 – “Time”

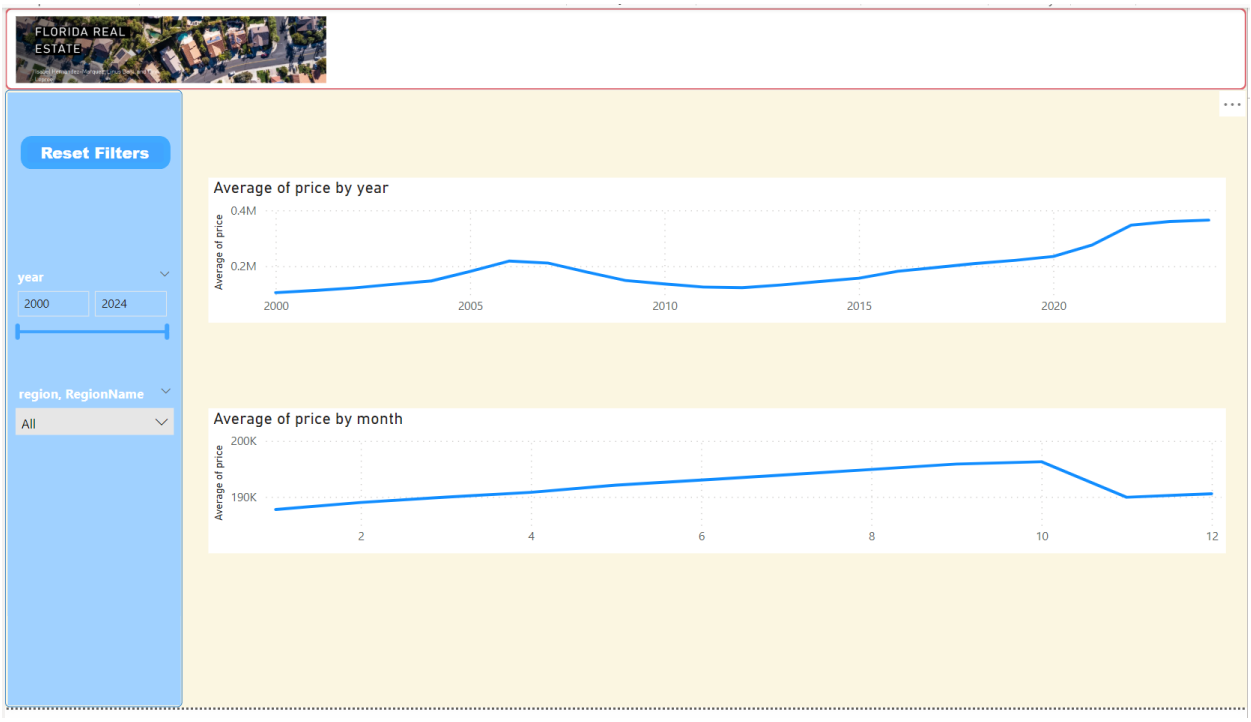


Figure 7. Initial Data Exploration of Price and Acre_Lot

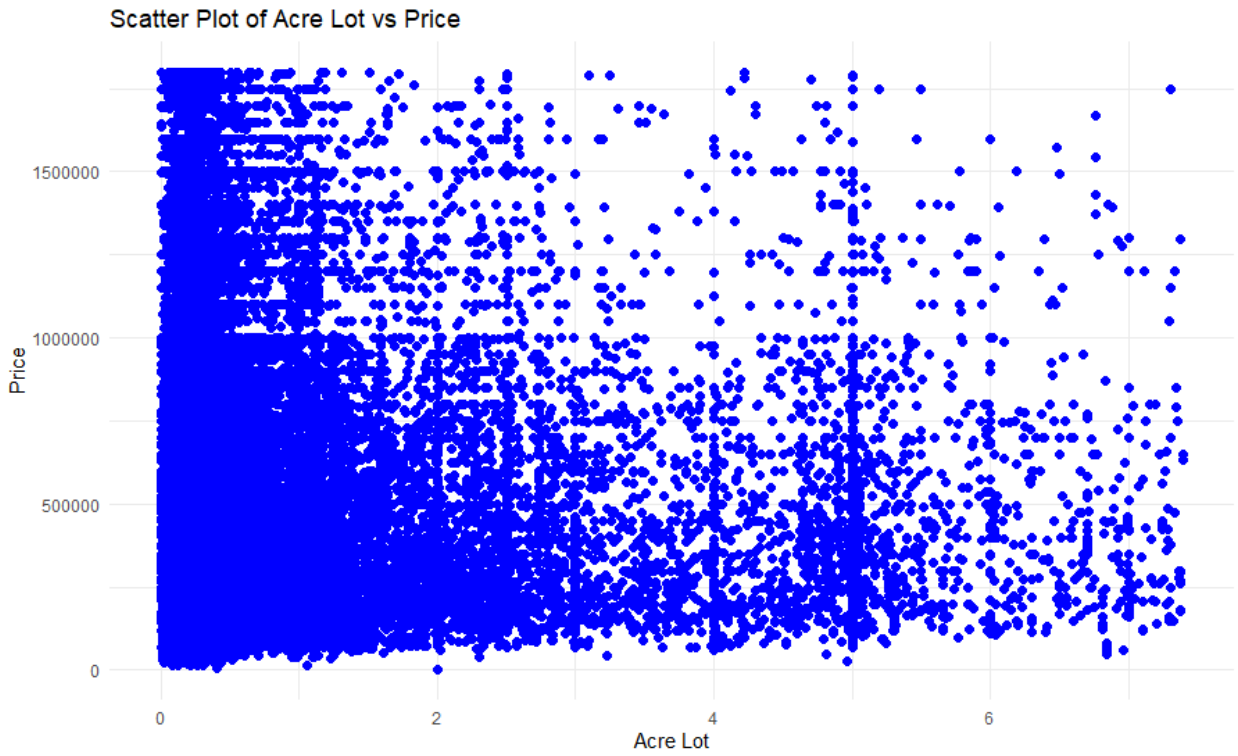


Figure 8. Initial Data Exploration of Price and Bed