# FLORIDA REAL ESTATE

Isabel Hernandez-Marquez, Linus Dahl, and Finn Lepree

# BUSINESS UNDERSTANDING

Industry Dynamics

- Influenced by economic, technological and social factors
- Competitive Market – Realtor.com leading platform connecting buyers/sellers/agents

Importance of Pricing

- Impacts sales time, customer satisfaction, and market position
- Main risks: Longer market time for properties – Reduced sales and customer attrition

Research objectives

- Investigate key factors influencing real estate pricing (size, location, etc.)

→ Develop a more accurate pricing strategy – Enhanced market position, long-term competitiveness

# DATA UNDERSTANDING

USA Real Estate Dataset with over 2 million observations

- Rich variables: price, bedrooms, bathrooms, lot size, house size, acre lot

Florida real estate dataset added to address missing time-related variables

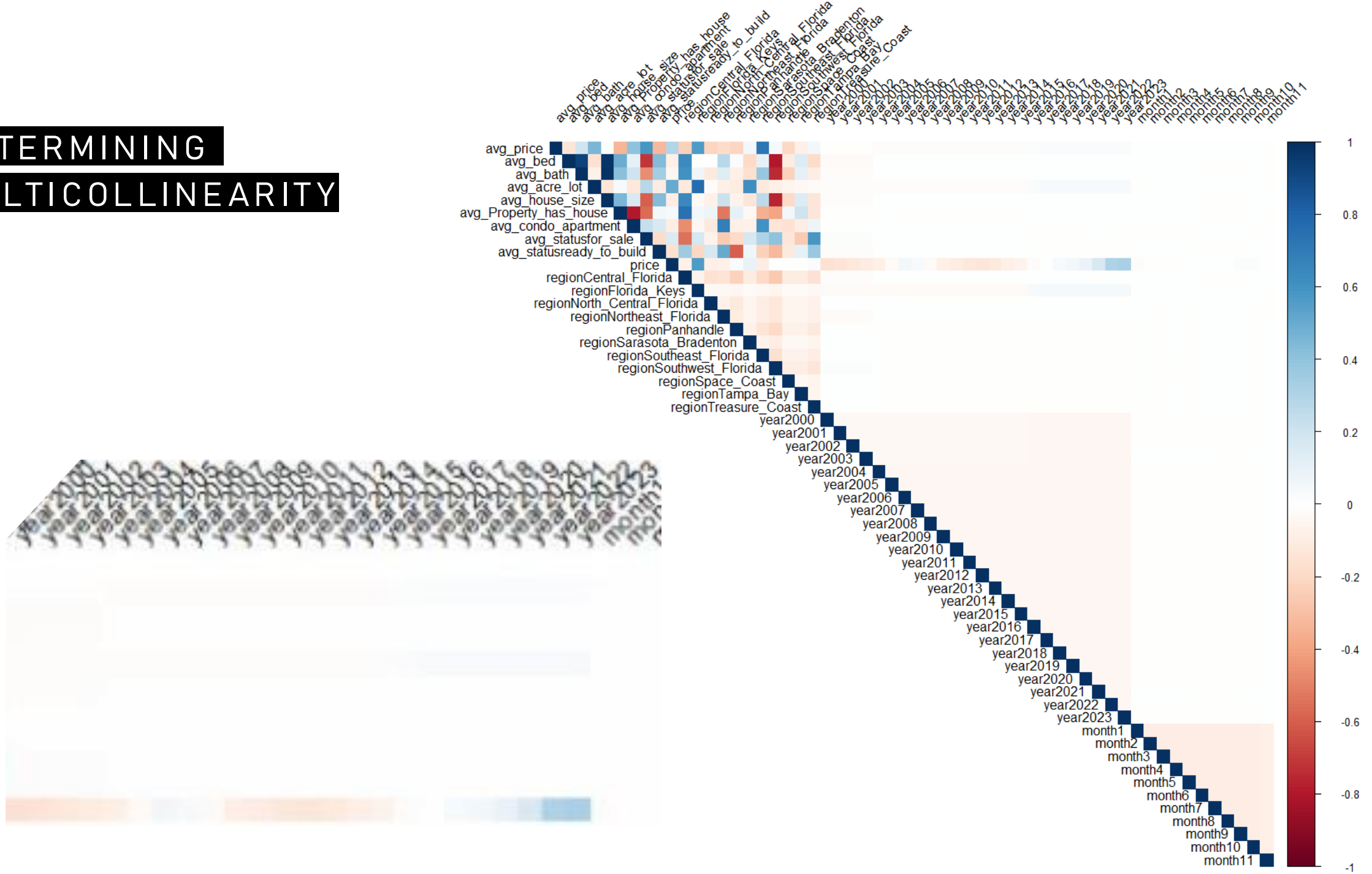
Key Insights:

- ❑ Regional differences in prices based on zip codes
- ❑ Bedroom and bathroom threshold impact value
- ❑ Lot size and specific features influence pricing

# DATA PREPARATION

- The code begins with two datasets
  - A Zillow dataset with housing prices over time by city
  - A housing dataset with features like bedroom, bathrooms, lot size, etc.
- We first remove outliers where price is more or less than two standard deviations
- We then group like regions
  - Cities in Florida are grouped into larger regions like "southeast_Florida"
  - This helps simplify the analysis and get rid of multicollinearity.
- We then calculate the averages for each region (average price, bedroom, bathroom, lot size, etc)
- We then join the two datasets together by region
- Finally, we remove missing values and create dummy variables.

# DETERMINING MULTICOLLINEARITY




Price ->

# BASE METRICS PRE FEATURE REDUCTION

```
Residual standard error: 25680 on 8233 degrees of freedom
Multiple R-squared:  0.9426,    Adjusted R-squared:  0.9421
F-statistic:  2144 on 63 and 8233 DF,  p-value: < 2.2e-16
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 4.415e+06 | 9.699e+04 | 45.518 | < 2e-16 | *** |
| avg_price | -2.540e+00 | 5.323e-02 | -47.717 | < 2e-16 | *** |
| avg_bed | -4.835e+05 | 1.897e+04 | -25.487 | < 2e-16 | *** |
| avg_bath | 7.879e+06 | 1.680e+05 | 46.910 | < 2e-16 | *** |
| avg_acre_lot | 1.227e+04 | 1.948e+02 | 62.969 | < 2e-16 | *** |
| avg_house_size | -8.992e+03 | 1.755e+02 | -51.253 | < 2e-16 | *** |
| avg_Property_has_house | -3.895e+06 | 1.120e+05 | -34.780 | < 2e-16 | *** |
| avg_condo_apartment | -3.476e+06 | 1.023e+05 | -33.979 | < 2e-16 | *** |
| avg_statusfor_sale | -1.516e+06 | 4.314e+04 | -35.138 | < 2e-16 | *** |
| avg_statusready_to_build | 3.502e+06 | 2.388e+05 | 14.665 | < 2e-16 | *** |
| regionCentral_Florida | -3.733e+05 | 1.071e+04 | -34.860 | < 2e-16 | *** |
| year2000 | -2.511e+05 | 2.123e+03 | -118.264 | < 2e-16 | *** |
| year2001 | -2.436e+05 | 2.123e+03 | -114.743 | < 2e-16 | *** |
| year2002 | -2.343e+05 | 2.123e+03 | -110.341 | < 2e-16 | *** |
| year2003 | -2.223e+05 | 2.129e+03 | -104.438 | < 2e-16 | *** |
| year2004 | -2.024e+05 | 2.066e+03 | -97.970 | < 2e-16 | *** |
| year2005 | -1.677e+05 | 2.066e+03 | -81.157 | < 2e-16 | *** |
| year2006 | -1.311e+05 | 2.064e+03 | -63.501 | < 2e-16 | *** |
| year2007 | -1.383e+05 | 2.064e+03 | -66.986 | < 2e-16 | *** |
| year2008 | -1.700e+05 | 2.064e+03 | -82.374 | < 2e-16 | *** |
| year2009 | -2.003e+05 | 2.064e+03 | -97.007 | < 2e-16 | *** |
| year2010 | -2.128e+05 | 2.064e+03 | -103.089 | < 2e-16 | *** |
| year2011 | -2.239e+05 | 2.064e+03 | -108.459 | < 2e-16 | *** |
| year2012 | -2.264e+05 | 2.064e+03 | -109.653 | < 2e-16 | *** |
| year2013 | -2.166e+05 | 2.064e+03 | -104.928 | < 2e-16 | *** |
| year2014 | -2.041e+05 | 2.064e+03 | -98.854 | < 2e-16 | *** |
| year2015 | -1.923e+05 | 2.064e+03 | -93.154 | < 2e-16 | *** |
| year2016 | -1.825e+05 | 2.047e+03 | -89.150 | < 2e-16 | *** |
| year2017 | -1.700e+05 | 2.046e+03 | -83.104 | < 2e-16 | *** |
| year2018 | -1.560e+05 | 2.046e+03 | -76.248 | < 2e-16 | *** |
| year2019 | -1.449e+05 | 2.046e+03 | -70.843 | < 2e-16 | *** |
| year2020 | -1.307e+05 | 2.046e+03 | -63.897 | < 2e-16 | *** |
| year2021 | -8.942e+04 | 2.050e+03 | -43.615 | < 2e-16 | *** |
| year2022 | -1.898e+04 | 2.046e+03 | -9.278 | < 2e-16 | *** |
| year2023 | -5.910e+03 | 2.046e+03 | -2.889 | 0.00388 | ** |
| month1 | -9.674e+03 | 1.393e+03 | -6.947 | 4.02e-12 | *** |
| month2 | -9.047e+03 | 1.392e+03 | -6.499 | 8.54e-11 | *** |
| month3 | -8.102e+03 | 1.392e+03 | -5.820 | 6.11e-09 | *** |
| month4 | -7.036e+03 | 1.393e+03 | -5.052 | 4.45e-07 | *** |
| month5 | -5.946e+03 | 1.392e+03 | -4.271 | 1.97e-05 | *** |
| month6 | -4.864e+03 | 1.393e+03 | -3.493 | 0.00048 | *** |
| month7 | -3.918e+03 | 1.393e+03 | -2.813 | 0.00491 | ** |
| month8 | -3.123e+03 | 1.393e+03 | -2.242 | 0.02501 | * |
| month9 | -2.430e+03 | 1.393e+03 | -1.744 | 0.08118 | . |
| month10 | -1.775e+03 | 1.392e+03 | -1.275 | 0.20221 | |
| month11 | -7.752e+02 | 1.406e+03 | -0.551 | 0.58137 | |
| RegionID394335 | 1.416e+04 | 2.104e+03 | 6.732 | 1.78e-11 | *** |
| RegionID394440 | 1.112e+05 | 2.104e+03 | 52.835 | < 2e-16 | *** |
| RegionID394476 | -1.568e+04 | 2.104e+03 | -7.452 | 1.01e-13 | *** |
| RegionID394528 | -5.665e+04 | 2.109e+03 | -26.855 | < 2e-16 | *** |
| RegionID394622 | 3.876e+04 | 2.109e+03 | 18.377 | < 2e-16 | *** |
| RegionID394685 | -9.292e+04 | 2.109e+03 | -44.052 | < 2e-16 | *** |
| RegionID394714 | 9.363e+04 | 2.206e+03 | 42.435 | < 2e-16 | *** |
| RegionID394766 | -8.147e+04 | 2.111e+03 | -38.590 | < 2e-16 | *** |
| RegionID394856 | 1.118e+05 | 2.104e+03 | 53.115 | < 2e-16 | *** |
| RegionID394901 | 2.253e+05 | 2.106e+03 | 106.971 | < 2e-16 | *** |
| RegionID394927 | -9.568e+04 | 2.212e+03 | -43.259 | < 2e-16 | *** |
| RegionID394943 | -3.189e+04 | 2.109e+03 | -15.117 | < 2e-16 | *** |
| RegionID394960 | -5.563e+04 | 2.104e+03 | -26.443 | < 2e-16 | *** |
| RegionID394971 | -8.553e+04 | 2.104e+03 | -40.652 | < 2e-16 | *** |
| RegionID394995 | -2.070e+03 | 2.104e+03 | -0.984 | 0.32519 | |
| RegionID395009 | 7.024e+04 | 2.104e+03 | 33.383 | < 2e-16 | *** |
| RegionID395080 | -1.231e+05 | 2.212e+03 | -55.660 | < 2e-16 | *** |
| RegionID395146 | -7.912e+04 | 2.104e+03 | -37.604 | < 2e-16 | *** |

# MODELING

- We use two methods to choose important features
  - Forward selection
    - Start with an empty model
    - Temporarily adds each feature to the model
    - Trains model to see if the feature is statistically significant
    - If it is then the feature is added to the final model
  - Backward Selection
    - Start with a full model
    - Remove features one by one starting with the highest first
    - Remove until no features with p-value greater than the threshold are left
- We then train both models 100 times with 100 random splits to determine which model performs the best on average

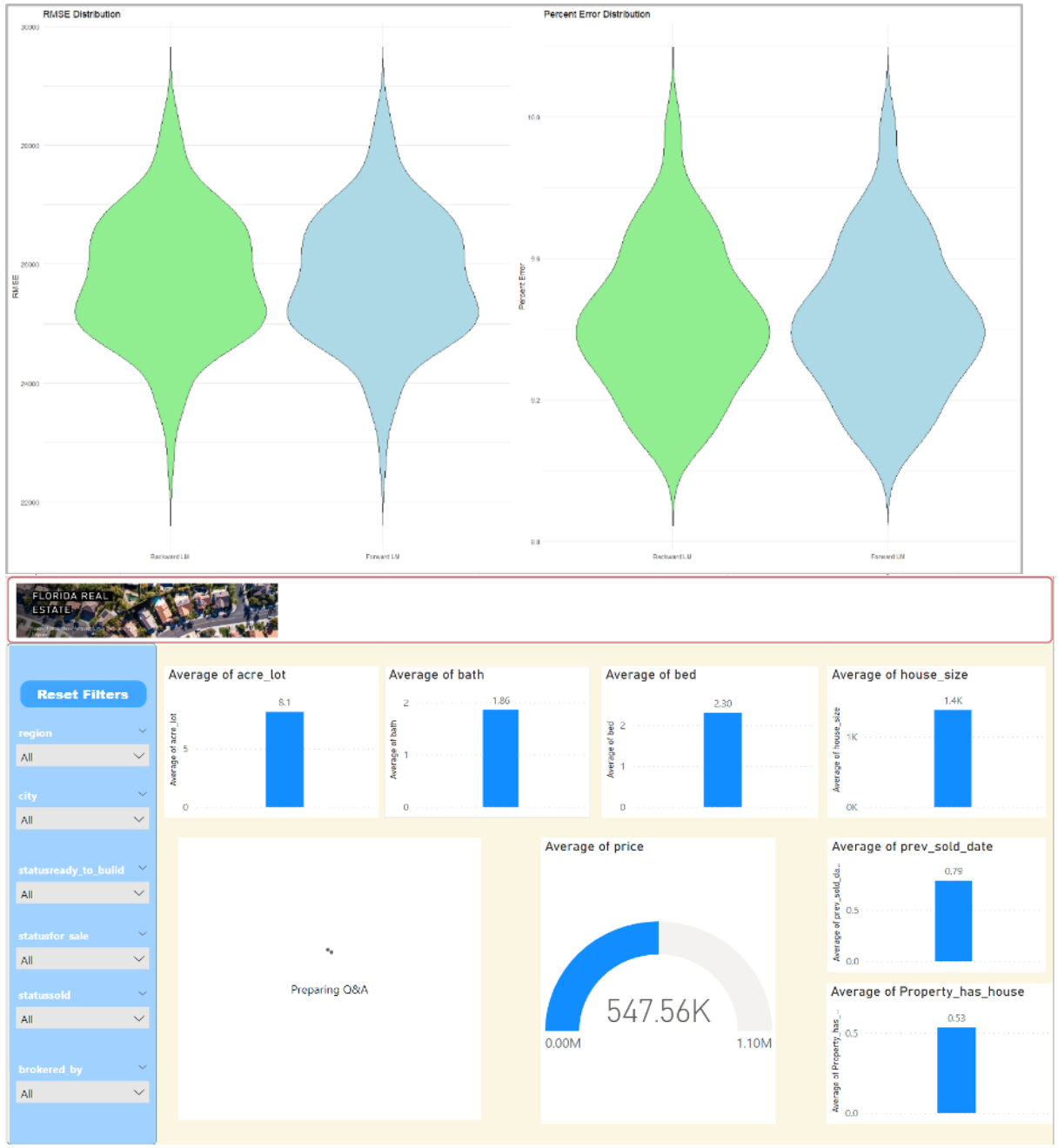
# AFTER FEATURE REDUCTION

*4 features removed

```
[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"
```

# EVALUATION

- Model:
  - RMSE error: 24,000-28,000
  - Low p-values : <2e-16
- Positive Predictors
  - House Size
  - Bathroom
  - Acre Lot
  - Location
- Negative Predictors
  - Bedroom



# FUTURE RECOMMENDATIONS

Enhance Data Scope


Improve Predictive Models

Strengthen Data Privacy

Continuous Model Review

# 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.