Group Project Final Report

Combining Disparate Approaches to Better Predict Future Residential Housing Prices

Team #60

Marc Presume ~ mpresume3

Joy Kakkanad ~ jkakkanad3

Evan Swain ~ eswain7

Russell Dawkins ~ rdawkins6

MGT 6203 Spring 2024

Professor Jonathan Fan

April 2024

1. INTRODUCTION

The United States housing market plays a crucial role in the economy, with a market cap of $47 trillion (about $140,000 per person in the US) and anticipated growth of 4.5% per year in the coming decade (Rosen, 2023). Given its significant contribution to economic activity and wealth generation, precise forecasting of residential real estate prices is vital for stakeholders.

Accurately forecasting residential property prices is pivotal in the real estate sector, as it unlocks numerous business opportunities and guides critical decision-making processes (Adetunji, et al., 2022). From assisting first-time homebuyers in securing fair deals to enabling large-scale investments in residential properties, the implications of precise price prediction are manifold. However, it remains a challenge for machine learning engineers to accurately predict housing prices (Manjula, Srivastava, Kher, & Jain, 2017).

Traditionally, price prediction in real estate has been approached through two primary methodologies. The first involves forecasting real estate prices in aggregate across regions or markets over time, while the second focuses on predicting individual property prices by comparing them to recent sales within the same market (Adetunji, et al., 2022). These approaches have historically relied heavily on historical sales data, which, while informative, often overlooks crucial factors such as neighborhood amenities and market sentiment.

Factor-based models offer an alternative approach, aiming to predict the price of a particular property within a specific market or neighborhood. These models utilize attributes like lot size and number of bedrooms to identify "comparable properties" and make predictions based on this comparison (Manjula, Srivastava, Kher, & Jain, 2017). Some even use clustering models to group houses in different categories and estimate their values (Li, Pan, Yang, & Guo, 2016). While effective in the short term, factor-based models tend to lack predictive power beyond the immediate future, providing limited insights into pricing trends over three, six, or twelve months. This restricts their usefulness for stakeholders requiring comprehensive, long-term price forecasts.

Data provided by the National Association of Realtors showed that 97% of home buyers rely on the internet in their home search, and 51% of buyers have found their purchased homes from the internet (Christopherson, 2021). Due to easy access to technology, online search tools are becoming more common and are used by homebuyers, real estate firms, and realtors. Recent studies have explored innovative approaches to housing price prediction, including the use of aggregated internet search data associated with machine learning algorithms (Guo, Chiang, Liu, Yang, & Gou, 2020). Other authors have demonstrated the potential of such methods in predicting housing prices on a large scale across time (Choi & Varian, 2012; Wu & Brynjolfsson, 2015; Veldhuizen, et al., 2016). However, while these approaches have shown promise in forecasting trends in major markets and broad indices, their applicability to smaller markets and individual properties remains undetermined.

In this project, we examine whether aggregated internet search data can improve a factor-based model in predicting the future sales price of individual residential properties. If aggregated internet search data can improve factor-based models in predicting the sales price of single properties, it presents an opportunity to improve existing price prediction models.

2. METHODOLOGY

2.1 Data Collection and Characteristics

Our project combines sales price (closing price) and detailed property characteristics from home sales in the Houston metro area from the past five years and internet search data from Google for the same period. The housing data (HAR data) was downloaded from the agent section of the Houston Association of Realtors'(HAR) website and spans the period December 2018 to February 2024. The aggregated internet search statistics on the Houston housing market (GT data) was downloaded from Google Trends (Google Trends, 2019).

The HAR data contains over 120,000 observations with 363 columns of continuous, discrete, and categorical variables. These variables include closing date, closing price, and characteristics of all single-family home sales in the Houston market. The GT data is aggregated statistics on specific queries to the Google search engine. Specifically, it is the ratio of searches for a specific query to all searches in a given time frame. For the geographic region from which to pull the query statistics, we chose the United States. This is because Houston's population has rapidly expanded in the past decade, with many new residents immigrating from out of state (McCullough & Ura, 2016; Hoque, 2017).

For the specific search queries, we used seven based on the recommendations of a Houston area realtor. These are "HAR house for sale," "HAR Houston," "HAR real estate," "houses for sale Houston," "Houston home prices," "Houston real estate," and "Houston realtor." By inputting these parameters to the Google Trends query site, we retrieved time-series data of monthly search statistics for each query (Google Trends, 2019). These data represent the ratio of searches for each query relative to all queries in the United States for the month specified.

2.2 Data Cleaning

2.2.1 Houston Association of Realtors Data

The HAR data required considerable cleaning due to missing values. We undertook this process in five steps. First, all blank values were converted to "NA" and the percentage of rows missing data were calculated for each column. When columns with more than 50% of the missing data were removed, the number of columns was reduced to 147 (data_cleaned_stage1). Next, we examined each column while consulting a Houston area realtor and removed columns unlikely to significantly contribute to our model. This reduced our dataset to 34 columns (data_cleaned_stage2).

We then created dummy variables for categorical variables, creating a new column for each categorical variable value and recording one or zero in the column. This expanded the variables to 184 (data_cleaned_stage3). We continued by further eliminating columns with more than 15% of observations missing. This resulted in the elimination of 6 columns (data_cleaned_stage4).

Finally, we added a new column called HouseAge which represents the number of years since the house was built. We also eliminated observations where the property is for lease instead of sale. We further removed "PropertyType" (unary value) and "PostalCode" (redundant) and converted the categorical columns of "Geomarket area," "county," and "schooldistrict" into dummy variables (data_cleaned_stage5).

2.2.2 Google Trends data

The Google Trends data, on the other hand, was provided in a single time series and needed minimal cleaning. We simply converted the monthly row names to a distinct column and eliminated a descriptive row. The data was manipulated in several ways to incorporate it into our analysis. First, we summed the time series data for the seven different queries into one vector as a representation of aggregated search interest in the Houston real estate market for each month.

Second, to explore any delays in search interest impacting closing price, we lagged the summed data at one-, two-, three-, six-, and twelve-months. We also considered whether moving averages of the summed data may be good predictors. To explore this aspect, we calculated the simple moving averages for trailing three-, six-, and twelve-months.

Finally, we considered the possibility that momentum and trend in the summed data may be predictive of closing price. For a momentum indicator, we chose the Relative Strength Index (RSI) with weighted moving averages. For trend, we chose Moving Average Convergence/Divergence (MACD). We calculated the MACD of the summed data using exponential moving averages (EMAs). For our parameters, we chose the standard form, with a “fast” 12-month EMA, a “slow” 26-months EMA, and an oscillator signal line at a 9-month simple moving average.

2.3 Analysis Approach

We conducted our analysis in four steps. First, we conducted exploratory data analysis to understand the characteristics of the HAR data better. Second, we created multiple factor-based models using the HAR data, exploring different approaches to model and factor selection. We found multiple linear regression to be the best approach and chose the model with the most predictive power as measured by adjusted R-squared. Third, we created a model for predicting the monthly mean closing price of the HAR data from the Google Trends data. We took this step to better judge whether any effect found in the last step was meaningful or random. Finally, we added the Google Trends data as additional factors to our best HAR model and measured improvement.

2.4 Model selection

We built and fit three multiple linear regression models in our second step of analysis.

2.4.1 Lasso regression model

A lasso regression model was built using the “cv.glmnet” function in R with 10-fold cross validation and a lambda value of 20.

2.4.2 Linear regression

A linear regression model was built with the variables selected from the lasso regression model as predictor variables and closing price as the response variable.

2.4.3 PCA-RegEx-Forward Stepwise Regression Model

A linear regression model was built incorporating principal component analysis, regular expression factor consolidation, and forward stepwise factor selection. In this model, we reintroduced principal component analysis to reduce the correlation between the four remaining continuous factors before further variable selection.

We then used regular expression analysis to simplify the remaining categorical factors. We did this by extracting unique terms from each original factor variable column and creating new dummy variables for each term. For example, a property with a "Zoned, Central Electric, Solar Assisted" cooling system, originally recorded as a positive observation in one dummy variable "CoolSystem.Zoned.Central.Electric.Solar.Assisted" was reassigned positive observations in new dummy variables "Cooling_Zoned," "Cooling_Central.Electric," and "Cooling_Solar.Assisted."

Finally, we used forward stepwise regression for factor selection within the new dummy variables. Starting with a model containing only the four principal component vectors, we added the new dummy variables to the factor set of the model one at a time. For each new dummy variable, we refit the model and calculated the new model's adjusted R-squared. If the addition of the dummy variable improved the overall model's adjusted R-squared by one percent or more, we retained the dummy variable. If not, we removed the variable.

2.5 Google Trends data in predicting mean monthly closing price

In incorporating the Google Trends data to our model, we first investigated its predictive power of the monthly mean price of our HAR data. To accomplish this, we calculated the monthly mean closing price for all observations in the cleaned HAR data. We then fit linear regression models with our Google Trends data as independent variables and the mean closing price as the dependent variable.

2.6 Incorporating Google Trends data

We incorporated the Google Trends data to our best performing HAR model in two ways and measured changes in the model predictive power as measured in adjusted R-squared. First, we looked at each of the GT factors individually by adding them one-by-one to the HAR model. Next, we performed principal component analysis on the GT factors to remove correlation and created a new model with all the resulting principal component vectors added to the factors from our HAR model.

3. RESULTS AND DISCUSSION

3.1 Data Exploration

We found strong correlation between continuous variables in the HAR data in our exploratory analysis (Fig. 1). Since correlation between factors in linear regression models can lead to overestimation of predictive power, we minimized this correlation in our modeling approaches through variable reduction or principal component analysis.

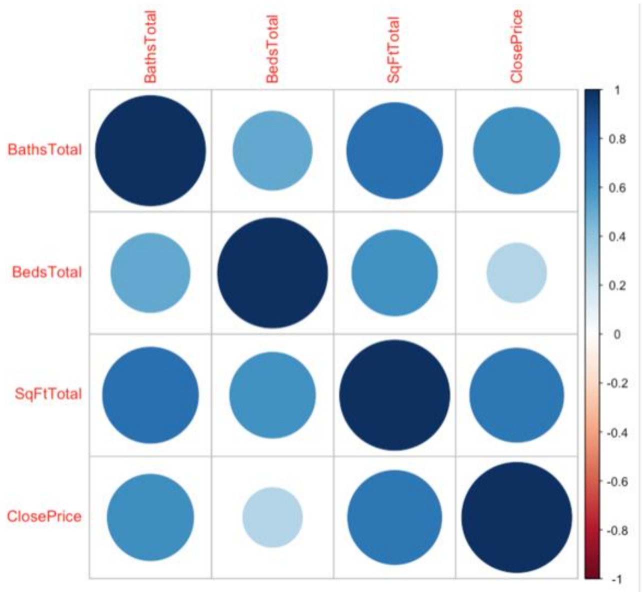


Fig 1. Correlation matrix of continuous variables

In Figure 2, results from the scatterplot showed a positive correlation between square footage and closing price. This means housing prices increased with increasing square footage. A similar result was observed for total baths. Closing price slightly increased with newer houses. However, no significant relationship was observed between closing year and closing price.

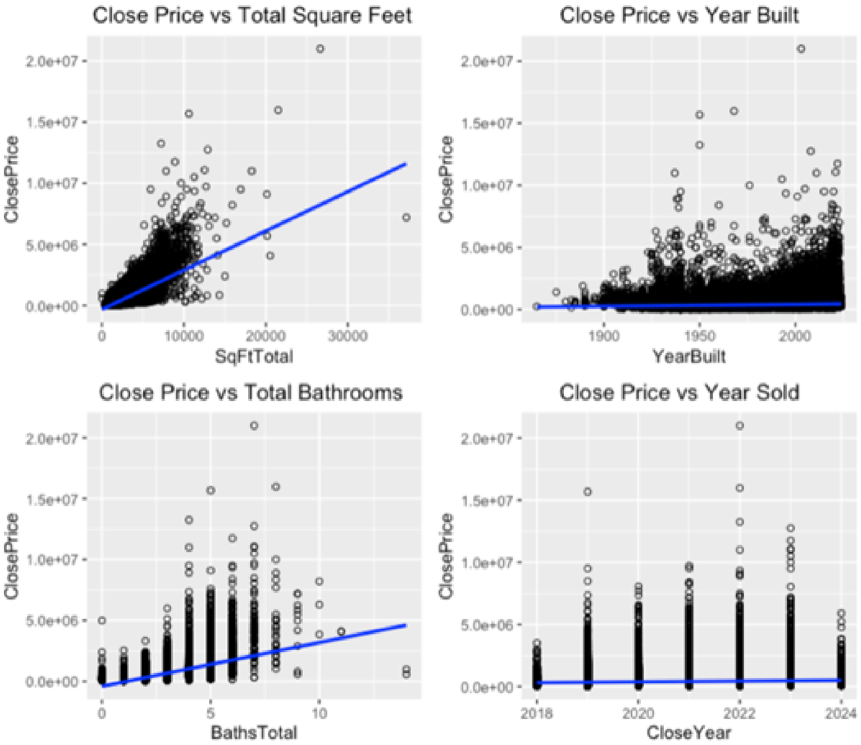


Fig 2. Scatterplot of Close price and square footage, year built, total bath, and close year.

3.2 Lasso Regression Model

By setting the penalty variable lambda at 20, we dropped all predictors that were not statistically significant. The resultant sparse matrix revealed that the most statistically significant predictors were HOAMandatory, NewConstruction, SqFtTotal, and PoolPrivate. These predictors were also checked against our correlation threshold of 0.5 to verify that there was no overfitting in our model. The linear regression model created by these factors produced an R squared value of 0.5792. While this value is lower than result of our full regression model (0.8036), we believe that this value is more accurate because the full regression model was overfit due to multicollinearity in the predictors.

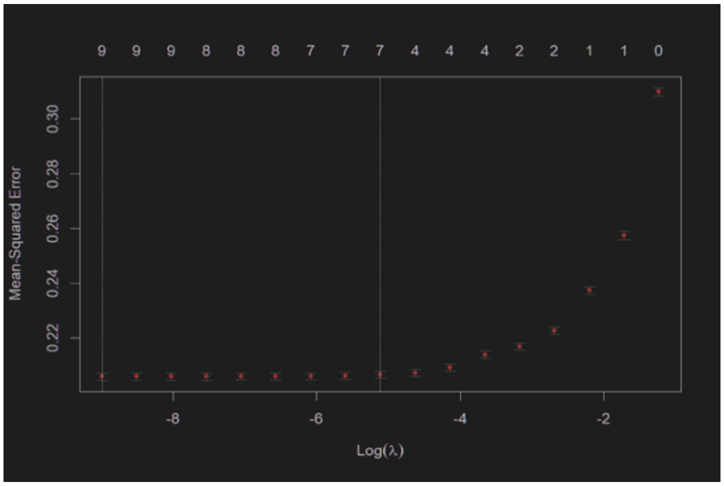
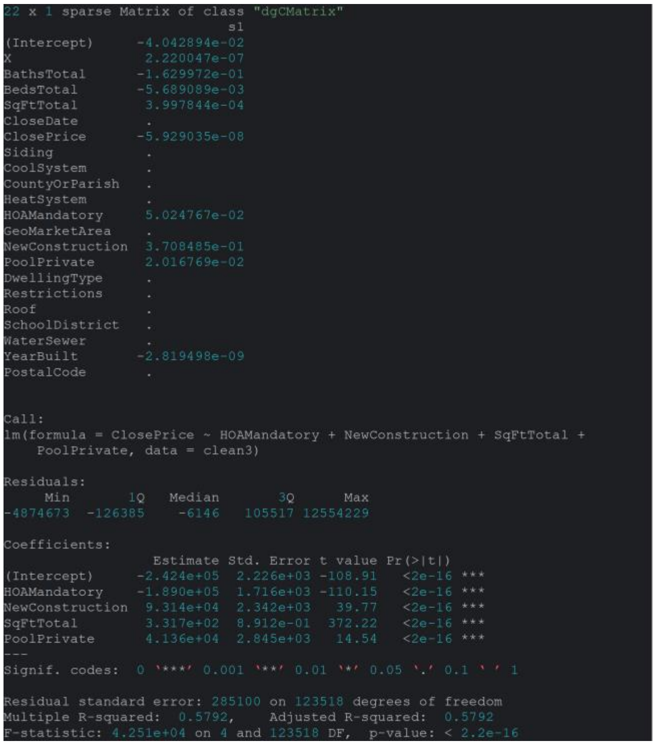


Fig3. Plot of Lasso Regression Model

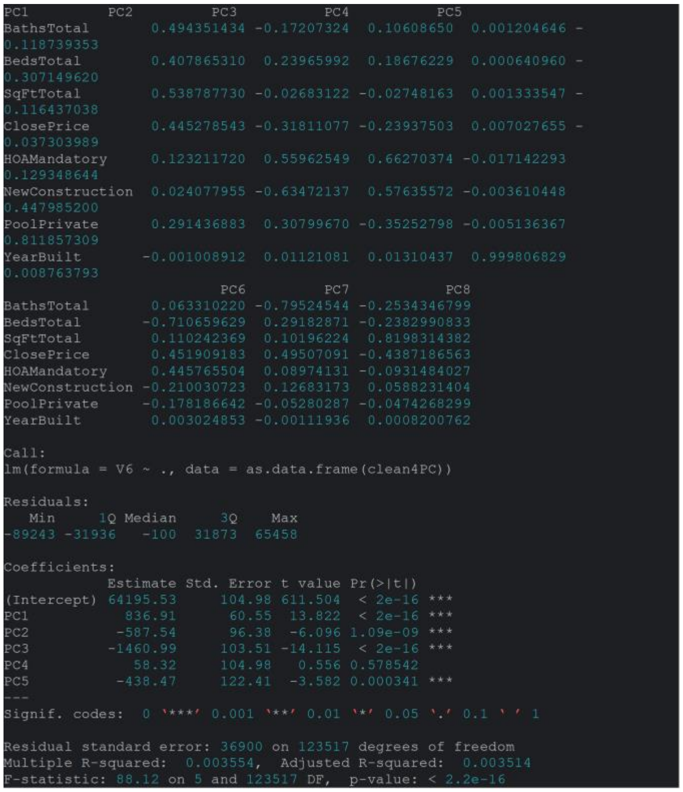


```
22 x 1 sparse Matrix of class "dgCMatrix"
                           s1
(Intercept)     -4.042894e-02
X                2.220047e-07
BathsTotal      -1.629972e-01
BedsTotal       -5.689089e-03
SqFtTotal        3.997844e-04
CloseDate        .
ClosePrice      -5.929035e-08
Siding           .
CoolSystem       .
CountyOrParish   .
HeatSystem       .
HOAMandatory     5.024767e-02
GeoMarketArea    .
NewConstruction  3.708485e-01
PoolPrivate      2.016769e-02
DwellingType     .
Restrictions     .
Roof             .
SchoolDistrict   .
WaterSewer       .
YearBuilt       -2.819498e-09
PostalCode       .


Call:
lm(formula = ClosePrice ~ HOAMandatory + NewConstruction + SqFtTotal +
    PoolPrivate, data = clean3)

Residuals:
     Min       1Q   Median       3Q      Max
-4874673  -126385    -6146   105517 12554229

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.424e+05  2.226e+03 -108.91   <2e-16 ***
HOAMandatory    -1.890e+05  1.716e+03 -110.15   <2e-16 ***
NewConstruction  9.314e+04  2.342e+03   39.77   <2e-16 ***
SqFtTotal        3.317e+02  8.912e-01  372.22   <2e-16 ***
PoolPrivate      4.136e+04  2.845e+03   14.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 285100 on 123518 degrees of freedom
Multiple R-squared:  0.5792,    Adjusted R-squared:  0.5792
F-statistic: 4.251e+04 on 4 and 123518 DF,  p-value: < 2.2e-16
```

Fig 4. Summary of Lasso Regression Model

3.3 Principal Component Analysis

In order to compute the Principal Component Analysis, a new dataframe was created that only included the numerical columns: BathsTotal, BedsTotal, ClosePrice, YearBuilt, HOAMandatory, NewConstruction, SqFtTotal, PoolPrivate, and YearBuilt. Based on the PCA plot, it appeared that most of the variance in the data could be captured in the first 5 principal components. A linear regression model built using the first 5 components revealed that all of them except for PCA 4 were statistically significant at the 0.001 level (Fig. 5). However, the adjusted R^2 value of 0.003514 reveals that this model poorly predicts home prices. The poor fit of this model is likely because each factor was adjusted for correlation prior to the model selection. Therefore, it can be concluded that PCA by itself isn’t a good model for variable selection, but it can be used to reduce correlation in a more advanced model.



```
PC1          PC2            PC3             PC4             PC5
BathsTotal       0.494351434 -0.17207324  0.10608650  0.001204646 -
0.118739353
BedsTotal        0.407865310  0.23965992  0.18676229  0.000640960 -
0.307149620
SqFtTotal        0.538787730 -0.02683122 -0.02748163  0.001333547 -
0.116437038
ClosePrice       0.445278543 -0.31811077 -0.23937503  0.007027655 -
0.037303989
HOAMandatory     0.123211720  0.55962549  0.66270374 -0.017142293
0.129348644
NewConstruction  0.024077955 -0.63472137  0.57635572 -0.003610448
0.447985200
PoolPrivate      0.291436883  0.30799670 -0.35252798 -0.005136367
0.811857309
YearBuilt       -0.001008912  0.01121081  0.01310437  0.999806829
0.008763793
                         PC6         PC7          PC8
BathsTotal       0.063310220 -0.79524544 -0.2534346799
BedsTotal       -0.710659629  0.29182871 -0.2382990833
SqFtTotal        0.110242369  0.10196224  0.8198314382
ClosePrice       0.451909183  0.49507091 -0.4387186563
HOAMandatory     0.445765504  0.08974131 -0.0931484027
NewConstruction -0.210030723  0.12683173  0.0588231404
PoolPrivate     -0.178186642 -0.05280287 -0.0474268299
YearBuilt        0.003024853 -0.00111936  0.0008200762

Call:
lm(formula = V6 ~ ., data = as.data.frame(clean4PC))

Residuals:
   Min     1Q Median     3Q    Max
-89243 -31936   -100  31873  65458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64195.53     104.98 611.504  < 2e-16 ***
PC1           836.91      60.55  13.822  < 2e-16 ***
PC2          -587.54      96.38  -6.096 1.09e-09 ***
PC3         -1460.99     103.51 -14.115  < 2e-16 ***
PC4            58.32     104.98   0.556 0.578542
PC5          -438.47     122.41  -3.582 0.000341 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36900 on 123517 degrees of freedom
Multiple R-squared:  0.003554,  Adjusted R-squared:  0.003514
F-statistic: 88.12 on 5 and 123517 DF,  p-value: < 2.2e-16
```

Fig 5. Summary of initial Principal Component Analysis and the resultant Linear Model

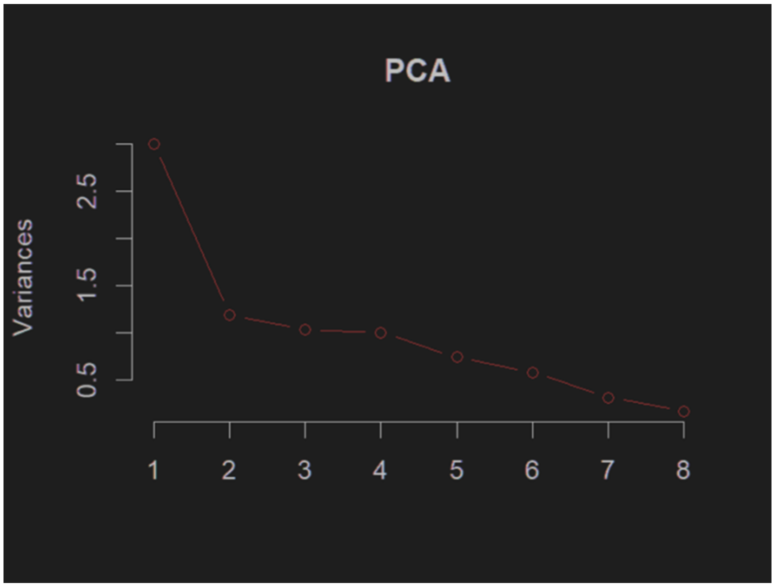
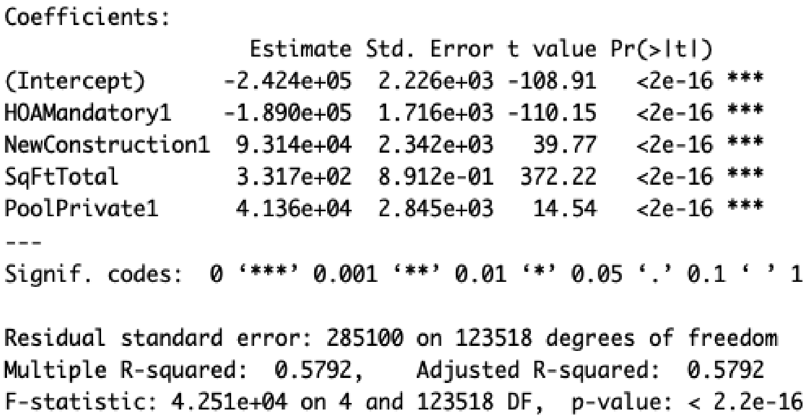


Fig 6. Screeplot for initial Principal Component Analysis

3.4 Linear regression model

The linear regression model was built with the variables selected from the lasso regression. We were looking for the relationship between 4 independent variables (HOA Mandatory, New construction, Total square footage, and Private pool) and closing price, our response variable. All 4 variables were statistically significant with significance level less than 0.001. The model gave an adjusted r-squared value of 0.5792, which means nearly 58% of variability in the closing price was explained by the independent variables in this model (Fig. 7).



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.424e+05  2.226e+03 -108.91   <2e-16 ***
HOAMandatory1    -1.890e+05  1.716e+03 -110.15   <2e-16 ***
NewConstruction1  9.314e+04  2.342e+03   39.77   <2e-16 ***
SqFtTotal         3.317e+02  8.912e-01  372.22   <2e-16 ***
PoolPrivate1      4.136e+04  2.845e+03   14.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 285100 on 123518 degrees of freedom
Multiple R-squared:  0.5792,	Adjusted R-squared:  0.5792 
F-statistic: 4.251e+04 on 4 and 123518 DF,  p-value: < 2.2e-16
```

Fig 7. Summary of the linear regression model

3.5 PCA-RegEx-Forward Stepwise Model

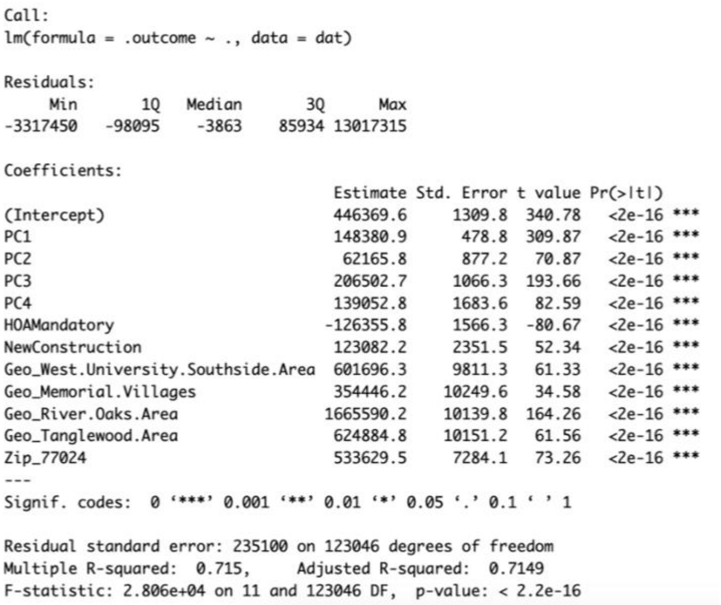
In this model, we found all principal component vectors resulting from PCA to have significant coefficients when fitted to the data using multiple linear regression. We retained all four vectors.

In simplifying dummy variables using regular expression analysis, we found finer parsing of the text of categorical variables reduced the number of necessary dummy variables in many cases. This is because variations in the data entry of each factor had resulted in a proliferation of similar dummy variables in earlier models. In addition, this produced a model which better accounted for the effects of single features. We attribute this to observations for a key term not becoming dispersed among several dummy variables. For example, after regular expression analysis, the effect of central electric cooling on closing price was no longer dispersed between observations of "zoned central electric cooling," "central electric cooling," and "solar assisted central electric cooling."

In some cases, we further reduced dummy variables by identifying significant terms within a single categorical variable. For siding materials, which had 383 unique terms after regular expression analysis, we created a linear regression with closing price as the dependent variable and the parsed siding material dummy variables as independent variables. We then retained the dummy variables with model coefficients significant at the $\alpha = 0.05$ level and removed the remaining terms. This is because we expected certain terms (e.g. "asbestos," "stucco") to have a substantial impact in predicting price while most would not.

In other cases where we anticipated the relative relationship of variables to be important, we chose to retain all the dummy variables. For example, we retained all 144 variations of different postal codes after reducing all entries to five digits and discarding inaccurate entries using regular expression analysis.

Our forward stepwise variable selection of the resulting dummy variables resulted in seven of these variables being chosen: "HOAMandatory," "NewConstruction," "Geo_West.University.Southside.Area," "Geo_Memorial.Villages," "Geo_River.Oaks.Area," "Geo_Tanglewood.Area," and "Zip_77024." Combined with the four principal component vectors, this resulted in a multiple regression model with eleven factors. Regressing these factors against closing price yielded our best performing multiple linear regression model for the HAR data. This model had an R-squared of 0.715 and an adjusted R-squared of 0.7149. We chose this model as the base model for incorporating the Google Trends data.



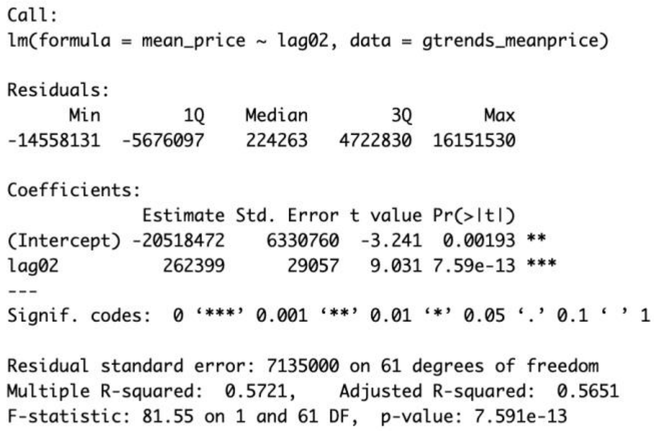
```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-3317450   -98095    -3863    85934 13017315

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       446369.6     1309.8  340.78   <2e-16 ***
PC1                               148380.9      478.8  309.87   <2e-16 ***
PC2                                62165.8      877.2   70.87   <2e-16 ***
PC3                               206502.7     1066.3  193.66   <2e-16 ***
PC4                               139052.8     1683.6   82.59   <2e-16 ***
HOAMandatory                     -126355.8     1566.3  -80.67   <2e-16 ***
NewConstruction                   123082.2     2351.5   52.34   <2e-16 ***
Geo_West.University.Southside.Area  601696.3     9811.3   61.33   <2e-16 ***
Geo_Memorial.Villages             354446.2    10249.6   34.58   <2e-16 ***
Geo_River.Oaks.Area              1665590.2    10139.8  164.26   <2e-16 ***
Geo_Tanglewood.Area               624884.8    10151.2   61.56   <2e-16 ***
Zip_77024                         533629.5     7284.1   73.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 235100 on 123046 degrees of freedom
Multiple R-squared:  0.715,     Adjusted R-squared:  0.7149
F-statistic: 2.806e+04 on 11 and 123046 DF,  p-value: < 2.2e-16
```

Fig 8. Summary of the PCA-RegEx- Forward Stepwise Model

3.6 Google Trends data in predicting mean monthly closing price

We took the step of modeling the relationship of the mean monthly closing price of the HAR data to the Google Trends data because several authors found the latter to be less predictive of housing prices for smaller markets, even at the state level (Wu & Brynjolfsson, 2015; Veldhuizen, et al., 2016). In attempting to evaluate this data in predicting the closing price of individual houses in a single city, we wanted to get some context in judging whether any observed effects were meaningful or random.

We found the lagged GT data and moving averages to be good predictors of the mean price, with the two month-lagged GT data (lag02) being the best predictor as measured by R-squared. The two month-lagged GT data in a simple linear regression explained 57% of the variation in mean monthly closing price. This was quite remarkable given that our mean price is a very crude measurement unadjusted for the number or category of houses sold in each month.



```
Call:
lm(formula = mean_price ~ lag02, data = gtrends_meanprice)

Residuals:
      Min        1Q    Median        3Q       Max
-14558131  -5676097    224263   4722830  16151530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -20518472    6330760  -3.241  0.00193 **
lag02          262399      29057   9.031 7.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7135000 on 61 degrees of freedom
Multiple R-squared:  0.5721,    Adjusted R-squared:  0.5651
F-statistic: 81.55 on 1 and 61 DF,  p-value: 7.591e-13
```

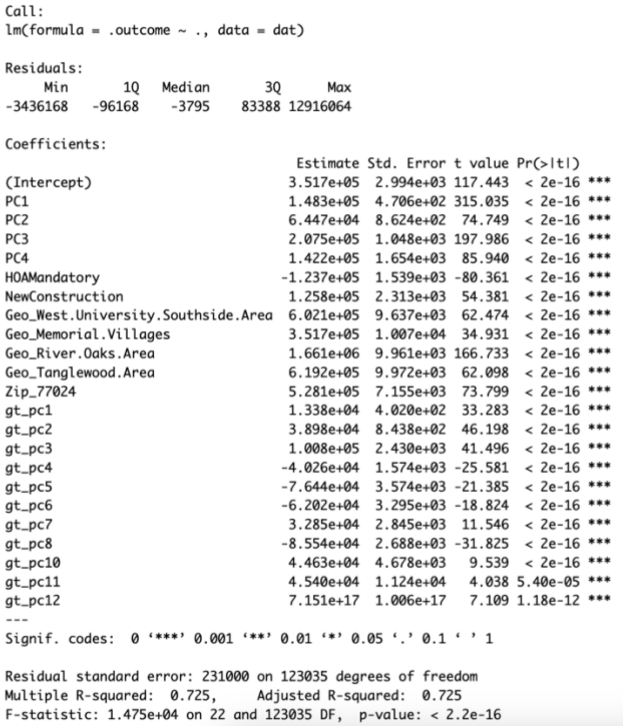
Fig 9. Summary of the Google Trends Model

3.7 Impact of GT data on predicting individual housing prices

When GT factors were individually added to the base HAR model, we found all the factors except the three-month moving average increased the predictive power slightly as measured by adjusted R-squared. Of these, the 12-month lagged sum of the GT data was most additive, raising the adjusted R-squared to 0.7167.

When all GT factors were PCA-transformed and added to the base HAR model, we found two of the resulting principal component vectors had coefficients which were insignificant (gt_pc9 and gt_pc13). Removing these and refitting the model using ten-fold cross-validation, we found the addition of GT factors to the baseline model resulted in a model with R-squared and adjusted R-squared of 0.725. This is a modest, but significant, 1.41% increase in the adjusted R-squared, despite the addition of 11 factors.

Although the increase in predictive power was modest, taken in the context of our findings in modeling the mean monthly closing price, we conclude the Google Trends data does contribute to a better prediction model for Houston single-family home prices.



```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-3436168   -96168    -3795    83388 12916064 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        3.517e+05  2.994e+03 117.443  < 2e-16 ***
PC1                                1.483e+05  4.706e+02 315.035  < 2e-16 ***
PC2                                6.447e+04  8.624e+02  74.749  < 2e-16 ***
PC3                                2.075e+05  1.048e+03 197.986  < 2e-16 ***
PC4                                1.422e+05  1.654e+03  85.940  < 2e-16 ***
HOAMandatory                      -1.237e+05  1.539e+03 -80.361  < 2e-16 ***
NewConstruction                    1.258e+05  2.313e+03  54.381  < 2e-16 ***
Geo_West.University.Southside.Area 6.021e+05  9.637e+03  62.474  < 2e-16 ***
Geo_Memorial.Villages              3.517e+05  1.007e+04  34.931  < 2e-16 ***
Geo_River.Oaks.Area                1.661e+06  9.961e+03 166.733  < 2e-16 ***
Geo_Tanglewood.Area                6.192e+05  9.972e+03  62.098  < 2e-16 ***
Zip_77024                          5.281e+05  7.155e+03  73.799  < 2e-16 ***
gt_pc1                             1.338e+04  4.020e+02  33.283  < 2e-16 ***
gt_pc2                             3.898e+04  8.438e+02  46.198  < 2e-16 ***
gt_pc3                             1.008e+05  2.430e+03  41.496  < 2e-16 ***
gt_pc4                            -4.026e+04  1.574e+03 -25.581  < 2e-16 ***
gt_pc5                            -7.644e+04  3.574e+03 -21.385  < 2e-16 ***
gt_pc6                            -6.202e+04  3.295e+03 -18.824  < 2e-16 ***
gt_pc7                             3.285e+04  2.845e+03  11.546  < 2e-16 ***
gt_pc8                            -8.554e+04  2.688e+03 -31.825  < 2e-16 ***
gt_pc10                            4.463e+04  4.678e+03   9.539  < 2e-16 ***
gt_pc11                            4.540e+04  1.124e+04   4.038 5.40e-05 ***
gt_pc12                            7.151e+17  1.006e+17   7.109 1.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 231000 on 123035 degrees of freedom
Multiple R-squared:  0.725,	Adjusted R-squared:  0.725 
F-statistic: 1.475e+04 on 22 and 123035 DF,  p-value: < 2.2e-16
```

Fig 10. Summary of the PCA-Transformed Google Trends Model

3.8 Additional Variables

We did additional analysis to explore factors such as Federal Interest rate, Inflation, Money Supply, and Migration trends. Based on the graph above, there is an inverse relationship between the closing price of a house and its age, which makes intuitive sense. The relationship between the monthly interest rate and monthly inflation rate on the closing price did not seem to have a clear relationship, at least visually. We found a faint increasing trend of the closing price with the M2 money Supply.

When running a basic linear regression model with BathsTotal, BedsTotal, SqFtTotal and House_Age instead of Year_Built, we found the R-squared and adjusted R-squared to give a marginally better score. When we added monthly interest rate, monthly inflation rate or Money Supply variable to this linear model, we observed that the R-squared and adjusted R-squared did not improve but remained the same. This is not to say that these variables are not significant, they are, it is just that the R squared and adjusted R squared is already at the upper limit.

We were also interested in exploring the migration trends for Houston and its surrounding areas to check its effect on the price. We think that additional exploration for these variables (Federal Interest rate, Inflation, Money Supply, and Migration trends) along with running suitable models can enhance the understanding of the impacts on the housing market prices.

Additional code and explanation are available on our GitHub site.

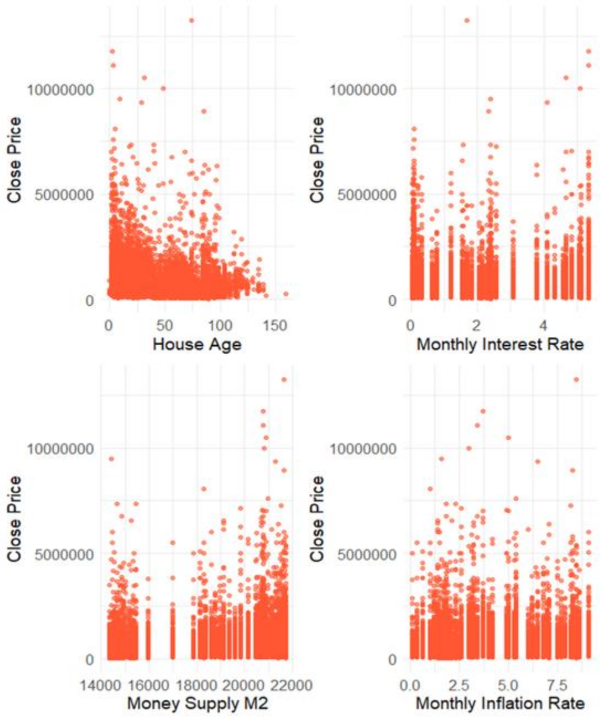


Fig 11. Grid plot of Close Price vs. House Age, Monthly Interest Rate, M2, and Monthly Inflation Rate respectively

4. CONCLUSION

In this project, we examined the effect of adding aggregated internet search data as a factor to a housing price prediction model using data from the Houston, TX market. Our approach was to find the most predictive model for our housing data, add the aggregated internet search data, and measure the change in predictive power as measured in adjusted R-squared.

For our base housing data model, we found correlation reduction in variables key to wrangling our large HAR dataset. We used principal component analysis to reduce correlation and various approaches to variable selection, including global approaches such as lasso regression and classical approaches such as stepwise regression. All models performed well, allowing us to explain 57-71% of the variability in closing prices based on the characteristics of individual houses and their location. Of our housing data models, our PCA-Regular Expression-Forward Stepwise Regression Model had the most predictive power with an adjusted R-squared of 0.7149. We chose it as our baseline housing data model from which to evaluate the addition of the Google Trends aggregated internet search data.

In incorporating the Google Trends data, we first modeled it against the mean monthly closing price of the HAR data to gain insight into the broader relationship of the two datasets. We found the two-month lagged Google Trends data explained 57% of the variation in the mean closing price. This finding stands out in the extent of its predictive power and adds to previous findings of aggregated internet search data's efficacy in predicting mean price of larger state-wide markets.

Transitioning to predicting closing prices of individual houses, we found adding the Google Trends data to our baseline model increased adjusted R-squared by 1.4%. Although a modest gain, this improvement suggests aggregated internet search data can lead to better predictions of individual residential housing prices, especially in the context of our findings in modeling the mean closing price.

Moving forward, we believe further investigation of search terms and geographic constraints in the Google Trends data can increase the additive effect of such data to housing price prediction models. Further, our analysis of additional time series factors such as inflation, M2 money supply, Federal interest rate, and migration trends suggest these too can each modestly improve housing price prediction models. Cumulatively, we believe these underutilized factors have the potential to substantially improve predictive power of housing price prediction models.

5. REFERENCES

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Elsevier*, 806-813.

Choi, H., & Varian, H. (2012). Predicting the Present wwith Google Trends. *Economic Record*, 2-9.

Christopherson, M. (2021, September). *Real Estate in a Digital Age*. Retrieved from National Association of Realtors: https://cdn.nar.realtor//sites/default/files/documents/2021-real-estate-in-a-digital-age-10-05-2021.pdf?_gl=1*lml1aa*_gcl_au*MjAxNjQzMTYzNS4xNzEwNjQzODM0

Google Trends. (2019). *Google Trends*. Retrieved from Google.

Guo, J.-q., CHIANG, S.-h., Liu, M., Yang, C.-C., & Gou, K.-Y. (2020). CAN MACHINE LEARNING ALGORITHMS ASSOCIATED WITH TEXT MINING FROM INTERNET DATA IMPROVE HOUSING PRICE PREDICTION PERFORMANCE? *International Journal of Strategic Property Management*.

HAR. (2024, 02 24). *Texas Real Estate - Homes for Sale and Rent*. Retrieved from Houston Association of Realtors.

Hoque, N. (2017). Projections of the Population of Texas and Counties in Texas by Age, Sex, and Race/Ethnicity from 2010 to 2050. *University of Houston School of Public Affairs White Paper Series*.

Li, Y., Pan, Q., Yang, T., & Guo, L. (2016). Reasonable price recommendation on Airbnb using Multi-Scale clustering. *Chinese Control Conference.* Chengdu.

Manjula, R., Srivastava, S., Kher, P. R., & Jain, S. (2017). Real estate value prediction using multivariate regression models. *IOP Conf. Series: Materials Science and Engineering* (p. 263). IOP Publishing.

McCullough, J., & Ura, A. (2016, 04 20). *Texas Drawing Millions Moving from Other States*. Retrieved from The Texas Tribune.

Rosen, P. (2023, 08 11). *Business Insider*. Retrieved from https://markets.businessinsider.com/news/commodities/housing-market-inventory-shortage-home-prices-value-real-estate-property-2023-8

Veldhuizen, S. v., Vogt, B., & Vogt, B. (206). Internet searches and transactions on the Dutch housing market. *Applied Economics Letters*, 1321-1324.

Wu, L., & Brynjolfsson, E. (2015). The Future of Prediction: How Google Search Foreshadow Housing Prices and Sales. In A. Goldfarb, S. M. Greenstein, & C. E. Tucker, *Economic Analysis of the Digital Economy* (pp. 89-118). Chicago: National Bureau of Economic Research.