

---

# ISyE 6740 – Fall 2022

## Project Final Report

---

Team Member Names: Namrata Buxani and Chonel Chase

Project Title: Predicting the Value of a House and Identifying the Top Macroeconomic & Microeconomic Value Drivers

### **Problem Statement**

The act of purchasing a home is a major decision that typically requires the consideration of both qualitative and quantitative factors. For example, insights on factors such as livable area, neighborhood, and the actual building quality impact the sticker price of the home. Other larger factors dictated by the national economy, like GDP and mortgage rate, affect the full amount a buyer pays to complete the sale.

We look at two datasets of all major factors that are considered in buying a house to understand which factors make the biggest impact in a buyer's decision, and therefore raise or lower the price of the house as well. In this model, we look at the accuracy of using various regression modeling techniques to predict housing price index (at the macro level) and the sale price of houses (at the micro level). This will help us determine the top value drivers for a house price from the macro and micro perspective. Researching this problem serves a great value to buyers across the country that are looking to make an informed house purchase.

### **Data Set Sources**

Our sources come from CoStar and Kaggle to understand and inform ourselves on both the quantitative and qualitative factors that are important within the real estate market and to a potential buyer. The macroeconomic factors dataset has 13 columns and 425 entries in it. These predictors are a mix of date, integer and float data types. Each record captures the value of each predictor on the first day of every month from January 1987 to May 2022. The first data source of this project is referenced in the appendix <sup>[7]</sup>.

In regards to the micro factors, we utilized a data source from Kaggle <sup>[8]</sup>. The microeconomic factors dataset has 81 columns and 1460 entries. The values are of integer, float and string data types. Each entry represents a house and its characteristics in Ames, Iowa.

### **Methodology**

The goal of this project is to identify the most influential macroeconomic and microeconomic factors that impact the price of housing. In our first data set, we use the Housing Price Index (HPI) as our dependent variable. HPI measures price changes of houses from a specific start date. Hence, we can make the assumption that the HPI would increase year over year if the average sale price for those same years is increasing as well. The macroeconomic factors included follow the U.S. economy and are the following: Year, Population, House Supply,

GDP, Mortgage Rate, Employment Rate, Permit New, Producer Price Index involved with Residential Construction, M3 Money Supply, Consumer Confidence Index, Delinquency Rate, and Housing Credit Availability Index.

Micro factor data is collected for Ames, Iowa. We acknowledge that this is a small city within the United States and that every community's micro factors will differ based on the subjectivity of what is considered important. The dependent variable for this data is the Sale Price of the home and variables are representative of the characteristics of the house. Micro factors are qualitative and quantitative data that describe and provide the value proposition to the house in question. Some examples of the variables are the following: Square Foot, Year Built, Number of Bedrooms, Number of Bathrooms, Floors, and Neighborhood.

Before running our models and statistical tests, we plan to standardize and clean the data via scaling, dropping NA's, and ensuring data types of columns are corrected. During our data exploration phase, we plan to seek out trends and correlations between various attributes that are within our datasets.

Principal Component Analysis (PCA) is utilized for further data cleaning by determining if there are specific years that the housing market attributes were more correlated or similar to each other. PCA will also be used to understand which factors are closely related to each other both on the micro and macro level. Variable Inflation Factor (VIF) and Lasso Regression will be attempted methods for variable selection to reduce the number of variables. Finally, we compare various regression model outputs to predict the price of a house and HPI within the specific years of the data provided. These models include Linear, Ridge, and Random Forest. During the implementation of the regression models, the alpha parameter will be tuned to the ideal value for the model size and data types. In our final results, we will determine which models are optimal for micro and macro housing factors, respectively. Metrics that will be looked at include R-Squared, Mean Squared Error, Root Mean Squared Error, and Error Rate.

## **Data Cleaning and Standardization**

Both the macro and micro factor datasets have null values for some of the entries. For the macro data set, we opt to drop these values from our data frame. This leaves us with 285 entries to use for our model. This dataset has a date column of the format 'yyyy-mm-dd'. Since entries of the same year did not have big variances, we group the data by year and average out the other predictors. This is a simple process, as all the predictors were numeric. This allows us to explore the data further based on year-to-year comparisons.

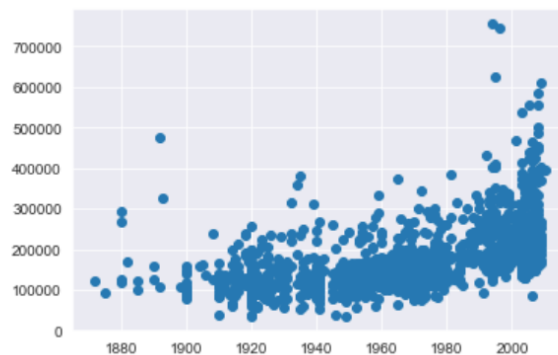
The micro dataset is more complex, because there are many categorical variables. Also, due to the wide range of features to describe a house, every data point has at least one NaN value for some attribute. This means we could not drop the entry like we are doing for the macro dataset. Instead, if the attribute was categorical, we use one-hot encoding. This means each categorical variable is replaced with a series of columns that represent the possible categories <sup>[4]</sup>. These new categories are then assigned a value of 1 if the original data point has a value for that category and zero otherwise. We prefer this method over assigning the categories a numerical value because this prevents the algorithm from seeing one category as "more important" than the others. The downside of choosing this method is that it increases the

number of columns, and therefore variables, to be followed by the model. Now, the micro dataset has 282 variables. If the attribute was numerical, we replaced the NaN value with a zero.

As part of the data exploring, we create a standardized version of our datasets. We utilize the StandardScaler package to preprocess and scale the data as needed for different processes at hand. To compare the variables fairly, it is important to ensure that the variances are compared at a magnitude that is similar to one another via scaling.

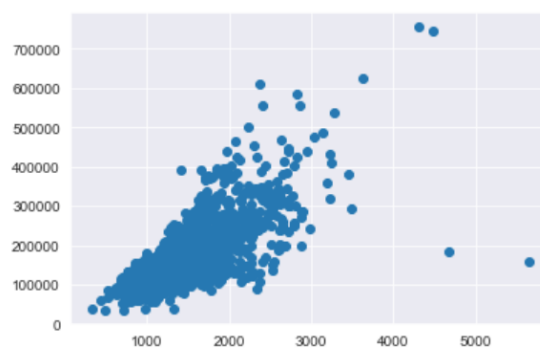
## Data Exploration

Once the data sources are finalized, cleaned, and standardized, we explore the data to understand if there are any correlations between variables, specific outliers, or other clear trends at a more surface level. Some examples of exploration are below:



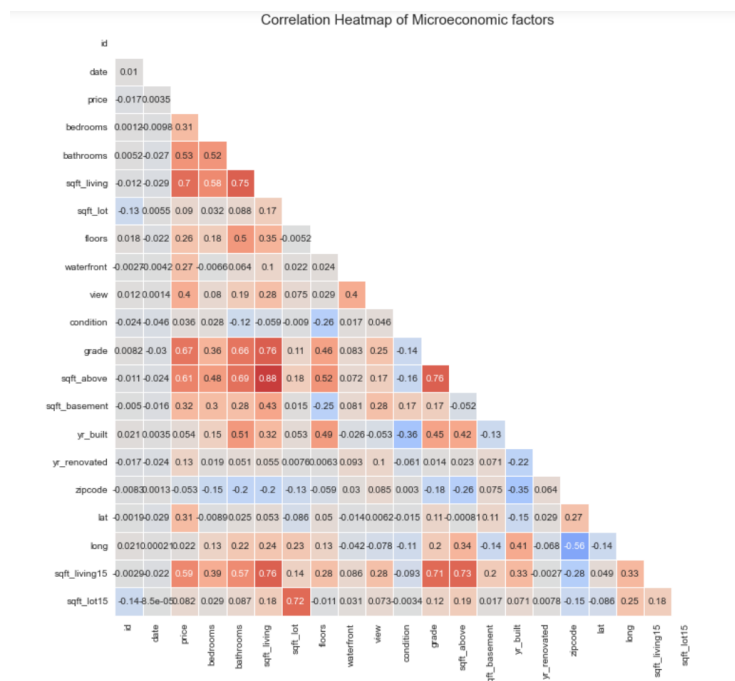
**Trend of Year Built versus Sale Price**

**Figure 1:** Comparison of year built to cost of house. In the above, Year is on the X-Axis and the Sale Price on the Y-Axis. One can notice that even some houses that were built in the 1800's still have high selling prices. There is not a direct linear correlation with the year built and the price of house. This is a common misconception that the year a house is built has a large impact on the price of the house. While it does have an impact, we assume other factors like location, refurbishment, and macroeconomic factors at the time have more of an impact. We remove the outliers that were over two standard deviations away from the mean for that given year.

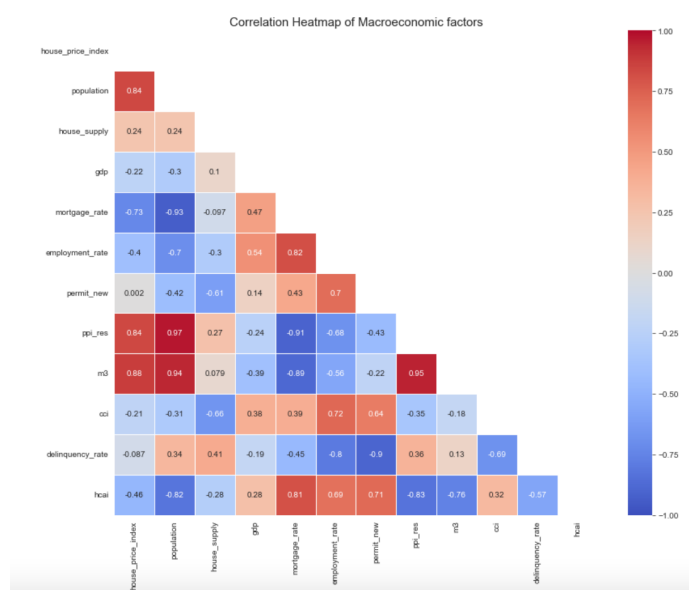


**Trend of Square Foot versus Sale Price**

**Figure 2:** Comparison of total square foot to cost of house. We can see that on average, there is a strong correlation between the square footage and the price of the house, aligning with expectations. However, there are a few anomalies that have very high square footage but low sale prices. This could be due to the condition of the house and other various factors. However, since there were so few scenarios like this, we remove these outliers as well.



**Figure 3:** Micro Factor Data Set Correlation Matrix: In this correlation matrix, we do not see a lot of strong correlation. Most of the strong positive correlation is representative of square footage of the homes based on the above or below ground level. It is logical that there is also a relationship between the number of baths and bedrooms. The more rooms in the home, the more likely that the square footage would increase as well. There is a negative correlation with zip code and longitude. This brought to light that although zip code is numerical, it solely operates in our data as a categorical variable because the zip codes are just a label for the locations and therefore do not represent a numerical value.

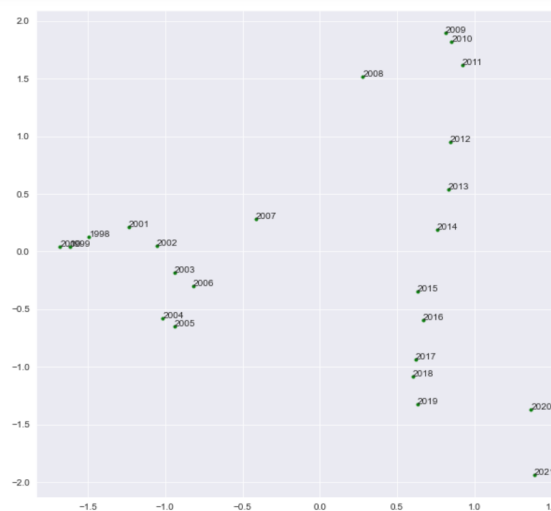


**Figure 4:** Macro Factor Data Set Correlation Matrix: Some factors are very closely correlated such as the M3 Money Supply and Producer Price Index. Mortgage rate is also correlated to employment rate which is expected. Population is also closely related to M3 Money Supply and Producer Price Index. Inversely, the CCI value decreases as the delinquency rate goes up which also makes logical sense. There are other correlations that can be analyzed as seen above.

## Data Cleaning Process via PCA

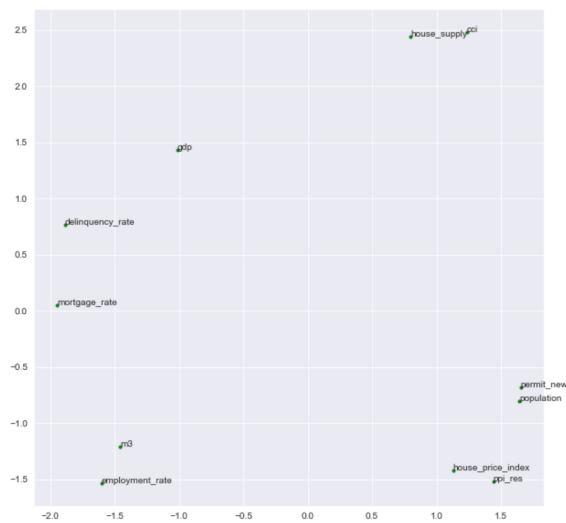
PCA is used to throw out outliers of data. The main data points that will be removed are 2007 and 2008 as pointed out in the PCA figure 5 as distinct from all other years. This is satisfying to see as we can acknowledge this was the great recession that was faced in the U.S.

We acknowledge that PCA is generally used as a variable selection technique. We did not use it for this in our project, but rather as an exploratory tool to find similar features as well as outliers that are in the dataset. Similar to the Food Consumption homework, we learn that features that are closer together are more closely correlated and share similar feature variable values.



**PCA Analysis of Macro Data Grouped by Year**

**Figure 5:** This process helps show that the years 2007 and 2008 are distinct from the other years. This makes sense because of the major recession that occurred during those years. Additionally, we see similarities between 2020 and 2021 which could be attributed to COVID-19.



## PCA Analysis of Micro Features

**Figure 6:** This process shows that the PPI and HPI are strongly correlated and associated with each other. You can see that the house supply and consumer confidence index are also pretty close features as well.

## Modeling and Predictions

### Variable Selection Procedure Method 1 - Variance Inflation Factor (VIF)

VIF score represents how well one variable is explained by another variable, detecting multicollinearity in regression models. Multicollinearity is when there is correlation between two features that are independent variables in the model. While R-Squared values will lie between 0 and 1, the formula for the VIF allows it to be greater than 1. As the R-Squared value increases, the greater the VIF score. Generally, a VIF score above 5 (or 10) can indicate a high multicollinearity. Too many variables that are highly correlated in this way can make it difficult to interpret results and can also cause overfitting.

- *VIF for Micro Data:*


	feature	VIF
0	Id	1.206750
1	MSSubClass	34.540139
2	LotArea	3.411319
3	OverallQual	5.700854
4	OverallCond	2.736926
5	YearBuilt	15.384162
6	YearRemodAdd	3.685736
7	MasVnrArea	3.093365
8	BsmtFinSF1	inf
9	BsmtFinSF2	inf
10	BsmtUnfSF	inf

16	BsmtFullBath	2.960097
17	BsmtHalfBath	1.462938
18	FullBath	4.151755
19	HalfBath	3.122439
20	BedroomAbvGr	3.475027
21	KitchenAbvGr	4.415801
22	TotRmsAbvGrd	6.626126
23	Fireplaces	2.096384
24	GarageYrBlt	2134.571810
25	GarageCars	8.129213
26	GarageArea	8.033068
27	WoodDeckSF	1.511033

28	OpenPorchSF	1.624476
29	EnclosedPorch	1.608341
30	3SsnPorch	1.216181
31	ScreenPorch	1.326141
32	PoolArea	1.544878
33	MiscVal	1.393516
34	MoSold	1.237829
35	YrSold	1.319821
36	SalePrice	14.487323

**Figure 7:** VIF Micro Analysis - We can see that most factors are less than 10 which indicates that there is not much multicollinearity, a positive takeaway. However, there is a lengthy list of variables that had very high VIF scores (not listed) that could indicate multicollinearity for variable selection. Before removing variables, we decided to use the LASSO regression model to assess further which variables to remove.

- *VIF for Macro Data:*



	feature	VIF
0	house_price_index	213.664246
1	population	99.625544
2	house_supply	39.141969
3	gdp	10.160276
4	mortgage_rate	54.372646
5	employment_rate	50.381939
6	permit_new	136.450740
7	ppi_res	200.654674
8	m3	166.850245
9	cci	9.240747
10	delinquency_rate	125.714466
11	hcai	59.595155

**Figure 8:** VIF Macro Analysis: Majority of the VIF scores are above 10 for the macro data values. This indicates that there may be multilinear collinearity that we need to be aware of. Since all the VIF scores are fairly high, using this as a variable selection method would remove almost all variables, leaving us with little to no variables to build our models with. Hence, we will need to try other methods like LASSO regression model to assess further what variables to remove.

## Variable Selection Procedure Method 2 - Lasso Regression

We use the VIF values to analyze multicollinearity and recognize the correlations between some of the variables but could not remove variables from the process. Lasso regression is therefore introduced now for variable selection because it has a penalty factor that helps determine the coefficient weights in the final equation for all features (variables). In order to create stronger models, we remove variables with a relatively low or zero value coefficient when training the other regression models.

LASSO regression requires some tuning as compared to the VIF method. There is an alpha parameter which is the penalty factor that represents the amount of shrinkage used by the model equation. This is defined prior to running the model, so we use cross validation in order to determine which alpha value gives us the strongest model for each regression technique. After running the model and comparing the coefficients, we remove the variables with the lower coefficient absolute value (macro) or the variables with coefficients of a 0 value due to the penalty factor (micro). For the macro dataset, this shifted our number of variables from 13 to 8. For micro, the number of variables reduces from 282 to 58.

Macro Factors to be removed: GDP, Mortgage Rate, Consumer Confidence Index, and Year

Micro Factors to be removed (summarized, not fully listed due to space): Total Rooms Above Ground, Exterior Covering on house, Neighborhood, Garage Condition, Building Type, Sale Condition, Type of Heating, Lot shape, Paved Driveway, Roof Material, Electrical System

*\*Note: due to one-hot encoding for categorical variables, some variables of the above categories remained in the dataset and were still used in the models - (ex. Exterior Covering is a categorical variable with several labels that were converted into multiple binary variables. Some of these variables were removed due to a 0 value coefficient and some remained in the final dataset)*

## Comparison of Regression Techniques

Now that we have the variables that we will remove from each data set, we will run all the regression models with all variables (Pre VS) and with remaining variables after variable selection (Post VS). The four models we used were Linear Regression, LASSO regression, Ridge Regression and Random Forest. All were run using the Scikit Learn library. To evaluate each of the models we use R-Squared, (Root) Mean Squared Error and Error Rates as our metrics. Error Rate is a test to see if the predictions are within 5% (@5) and 10% (@10) of the actual value.

Macro								
	Parameters		R2		MSE		Error	
	Pre VS	Post VS	Pre VS	Post VS	Pre VS	Post VS	Pre VS	Post VS
Linear	N/A	N/A	0.80 0.78	0.92 0.91	234	104	0.15@5 0.15 @10	0.14@5 0.13@10
Lasso	0.03	0.03	98.63 98.07	98.50 97.82	28.40	32.10	0.0854@5 0.0201@10	.0854@5 .0201@10
Ridge	0.5	1.0	98.62 98.11	98.50 97.88	27.80	31.20	0.0804@5 0.005@10	.0804@5 .0151@10
Forest	244	466	100.0 99.75	99.97 99.63	3.75	5.39	0.0@5 0.0@10	0.0@5 0.0@10

**Figure 8:** Output Metrics for Various Model Performance on Macroeconomic Factors

Micro								
	Parameters		R2		RMSE		Error	
	Pre VS	Post VS	Pre VS	Post VS	Pre VS	Post VS	Pre VS	Post VS
Linear	N/A	N/A	-1e-24 -1e-24	0.985 0.98	1.36e+17 1.36e+17	10671 9987	0.16@5 0.06@10	0.1@5 0.03@10
Lasso	1694.85	90.45	90.59 89.85	91.88 90.02	24387	23185	0.28@5 0.16@10	.28@5 .16@10

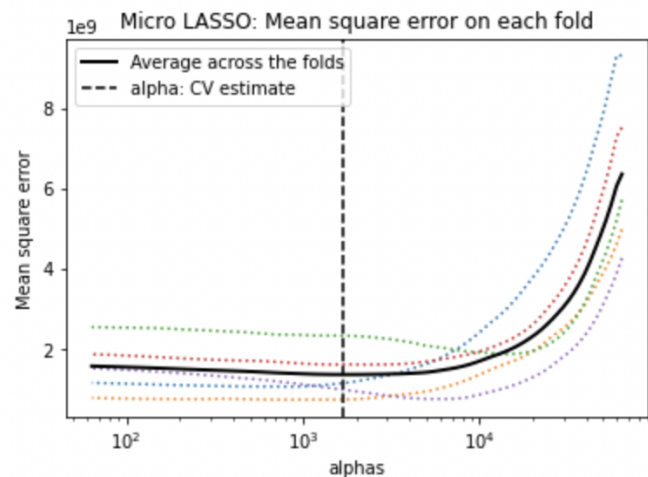
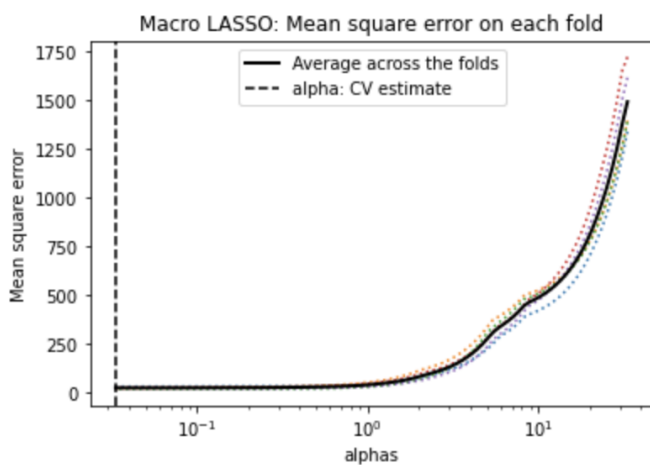


Ridge	514	1.9	90.79 89.36	91.89 90.18	24970	23876	.29@5 .17@10	.28@5 .16@10
Forest	511	333	100.0 88.23	100.0 88.96	26259	25426	0.26@5 0.15@10	.24@5 .14@10

**Figure 9:** Output Metrics for Various Model Performance on Microeconomic Factors - We take the Root Mean Square Error (RMSE) as it is easier to interpret with the magnitude of house price errors.

## Validation Method Definitions

- 1) **R-Squared** - Indicates how much variation of one variable is dependent on the other variable. A high R-Squared value is preferred overall.
- 2) **MSE/RMSE** - The average squared difference of estimated values compared to actual values. As seen above, we take the root mean squared error for the micro factor prediction of sale price. The resource <sup>[1]</sup> supports the idea of using the RMSE for house price predictions.
- 3) **Model Parameters:**
  - a) **Alpha Tuning via Cross Validation** - Lasso, Linear, and Ridge were all tested with the CV packages and tuning to cross validate the alpha parameter.
  - b) Below is the example of Lasso CV Folds over different alpha values to find the correct parameters. This same approach was applied to the various other models.



- c) **Random Forest Parameters via Cross Validation** - The SciKit Learn Random Forest Regressor has several parameters that can be tuned to create a well performing model. Using the Randomized Search CV method, a number of parameter settings are sampled from specified distributions. The parameters sampled were number of estimators (trees), maximum number of features, maximum tree depth, minimum samples to split, minimum samples per leaf, and bootstrap <sup>[5]</sup>.
- 4) **Error Rate** - Error Rate is a test to see if the predictions of HPI and Sale Price are within 5% and 10% of the actual value instead of measuring only if the prediction is *exactly* the

same as the prediction. This allows us to see how close the predictions are to the actual values.

## Evaluation and Final Results

The validation methods above are used to determine which regression model is best for predicting house price, micro factor values, and macro factor values. We will also compare PCA and other component methods as needed to determine the most influential variables for housing value.

One initial analysis made is that all of the R-Squared values are quite high for the models. This could indicate that there is overfitting of the data. Another potential cause could be having variables at different scales but this can be ruled out in our case as we have normalized and scaled data as needed. For models that have a higher MSE but lower error rate, this could be attributed to precision versus accuracy. Precision, which is the error rate, checks to see if the prediction is within an acceptable (precise) threshold and is black and white in its output (0 if in threshold, 1 if it is without the threshold, considering it an error). Accuracy, which shows how close the value is to the actual value, can be analyzed with the MSE/RMSE. Hence, if the error rate is higher but the MSE is lower, we can assume that the errors were outside the threshold we allow, but the differences are smaller overall.

## Output Analysis

The variable selection methods that are performed prove to make an impact on the error rates that are measured for both the macro and micro factors. In regards to the MSE/RMSE metric, it is more useful with the micro factor analysis as the magnitude of the error decreases with this process. On the flip side, with the macro factors, the magnitude of the MSE metric actually increases slightly. This is further supported by the error rate having no change or little change after the variables are removed. We can assume that the change in the MSE value is due to change in variance based on the factors that were kept but the overall accuracy of the predictions was still the same. In summary, removing the variables simplifies the model slightly but does not provide a large improvement to the model fit.

In regards to changes in the error rates, we can see that all the models had either no change or a decrease in the error rate with the exception of the macro Ridge regression model, which was an interesting takeaway. We will now explore the output in more detail by model and dataset below:

**Ridge Regression Model** - This method is introduced to determine the effect of the penalty factor in this model.

1. **Macro Analysis** - The Ridge regression model has a very high R-Squared value and an MSE value of 27. This indicates that the model is "good". However, both the alpha and the MSE value increased from 0.5 to 1 and 27 to 31 respectively after variable selection. The alpha is the penalty value and since it increased by 100%, the magnitude of the remaining coefficients reduced. However, this change made no impact to the prediction accuracy and increased the magnitude of the error slightly as well.

2. **Micro Analysis** - With the micro factors, the error rate decreases marginally after the variable reduction process. Similarly, the MSE decreases marginally at a similar percent of the original value with all variables.

**Linear Regression Model** - Linear regression can be used to find a relationship between the dependent variable and the independent variables to find a best fit straight line. This performs better if the variables are not extremely correlated. If they are, it is usually better to use Ridge regression.

1. **Macro Analysis** - The Linear regression performs much better after the dropping of variables. In fact, the MSE dropped by over 100% from 234 to 104. In addition, the R-Squared value increases from 0.79 to 0.91 indicating that the model is better fitted. Compared to the Ridge and Random Forest, this regression model has the best improvement from the variable selection method via the Lasso regression model.
2. **Micro Analysis** - The RMSE value before dropping the variables is very large. However, after dropping the variables the RMSE drops to a number that is still very large but more reasonable considering the magnitude of sale price is hundreds of thousands. The R-Squared value was the biggest change from a negative value to a very high positive value. We assume that even though the RMSE is very large, the fitting is better and this is further supported by the decreased error rate. The RMSE can also be explained by variance in the prediction of sale prices which are in the hundreds of thousands. Hence, even if the predictions are close, the magnitude of variance is also large.

An important call out for the micro Linear regression is that the R-Squared value before dropping variables was negative which indicates that the model is predicting very poorly, specifically worse than if the mean of target was used for every prediction. This indicates that the variable selection performed was necessary for the micro Linear regression to improve in performance.

**Random Forest Model** - Random Forest modeling does not overfit due to its approach using bagging. In fact, after a certain number of trees, the performance just plateaus, which helps us eliminate the idea of overfitting with our output metrics. <sup>[3]</sup>

1. **Macro Analysis** - The Random Forest interestingly increases in number of trees when the variables were removed. Most research supports the fact that the larger the number of data points, the larger the number of trees will be needed to optimize Random Forest. However, no evidence suggests the same for the number of variables. For Random Forests to make accurate predictions, it is important for there to be uncorrelated variables. So, because the initial dataset had many variables that were correlated, it is possible that due to the bagging function of Random Forest, the model did not fully capture the patterns of the data. This resulted in less trees and a less ideal performance in the initial model. The model with only the top variables as determined by LASSO variable selection is more complex, but has an increased performance. The error rate with our thresholding of 5% and 10% is zero which indicates that the trees and splits are very good at assisting in predictions of the housing price index data.
2. **Micro Analysis** - The number of trees needed decreases significantly once variables are removed. In addition, after training the data, the error rate is approximately 15% prior to

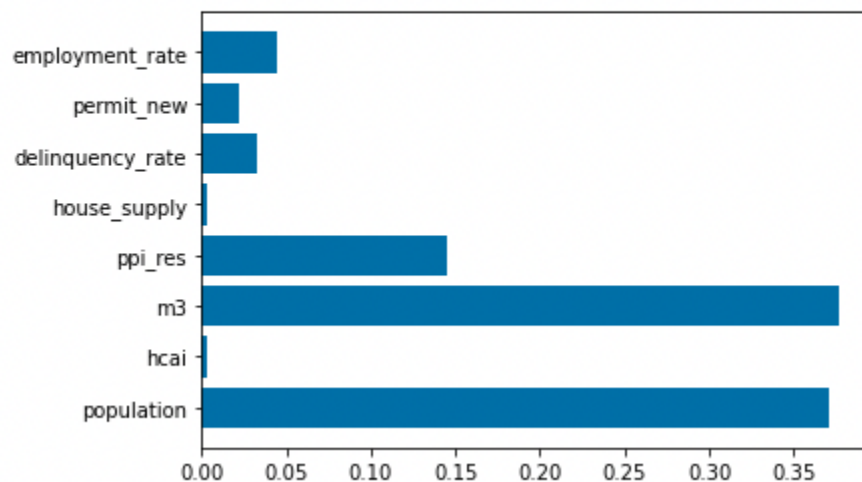
dropping variables and 14% after. This is a fairly good error rate considering variability in sale price data.

## Analysis of Top Features of Macro and Micro Data

### Top Ranked Macro Economic Features

Rank	Ridge ( <i>coef. value</i> )	Linear ( <i>coef. value</i> )	Forest ( <i>feature importance</i> )
1	Population (19.27)	M3 (26.81)	M3 (0.38)
2	PPI Res (15.14)	Population (26.22)	Population (0.37)
3	Hcai (12.35)	Permit_New (14.96)	PPI Res (0.15)
4	Permit New (12.48)	Hcai (12.78)	Employment Rate (0.04)
5	House Supply (12.24)	House Supply (11.49)	Delinquency Rate (0.03)
6	M3 (11.35)	Employment Rate (-9.06)	Permit_New (0.02)
7	Delinquency Rate (-7.12)	Delinquency Rate (-5.63)	Hcai (0.003)
8	Employment Rate (-4.9)	PPI Res (-2.33)	House Supply (0.003)

**Figure 10:** Top Ranked Macro Economic Features by coefficient values of regression equation after variable selection.



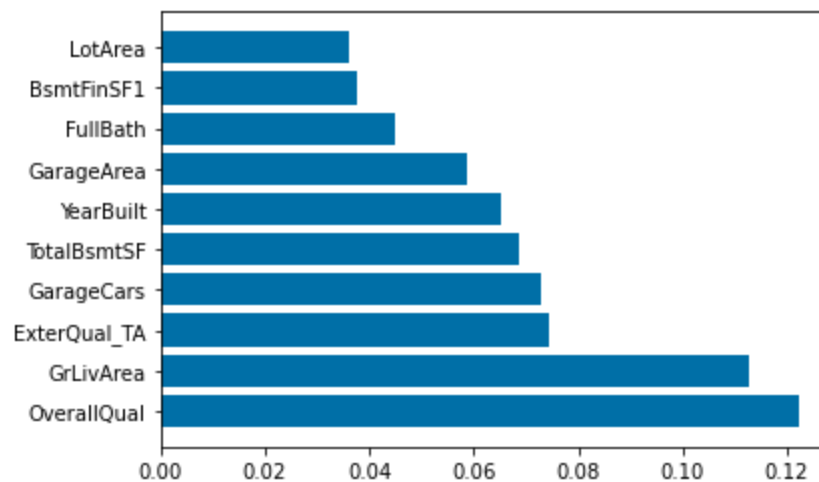
**Figure 11:** Top Ranked Macro Economic Features via Random Forest Regression<sup>[6]</sup>

The overall top three features for macro were the following: Population, M3 Money Supply, and PPI\_Res (Producer Price Index involved with Residential Construction). Population was ranked very high for all three models, ranking second for both the Linear and the forest model, while first for the Ridge models. On the flip side, M3 Money Supply was the top feature for Linear and Random Forest. Looking further into Linear, Population and M3 Money Supply were approximately tied as the feature coefficients are 26.81 and 26.22, respectively. This indicates that population is almost just as important as M3 Money Supply. However, M3 Money Supply was ranked slightly lower at #6 for the Ridge model. It is important to note that even if the rank was #6 for population, the coefficients for features 2-6 are between 11 and 15. Hence, ranking as #6 still holds a high weight. The third most important feature consistently was the PPI\_Res which was ranked 2 for Ridge and 3 for Random Forest. On the flip side, it was the least important feature for Linear, which was quite surprising to see. Lastly, when looking at the Random Forest feature importance values, one can note that M3 Money Supply, Population, and PPI\_Res are drastically more important than all the remaining variables.

### Top Ranked Microeconomic Features

Rank	Ridge <i>(coef. value)</i>	Linear <i>(coef. value)</i>	Forest <i>(feature importance)</i>
1	Above Ground Living Area (29196)	Roof Material: Clay/Tile (-584630)	Overall Quality (0.12)
2	Roof Material: Clay/Tile (-18895)	Proximity to various conditions: near positive off-site feature (-197887)	Above Ground Living Area (0.1128)
3	Overall Quality (11104)	Roof Material: Metal (134570)	External Material Quality: Typical (0.0745)
4	Proximity to various conditions: near positive off-site feature (-9223)	Garage Quality: Excellent (132954)	Garage Cars (0.07)
5	Basement Finished Square Feet (7780)	Garage Condition: Excellent (-120463)	Total Basement Square Footage (0.069)
6	Kitchen Quality: Excellent (7574)	Roof Material: Wood Shingles (96042)	Year Built (0.07)
7	Neighborhood: NorthRidge (7266)	Roof Material: Tar & Gravel (93038)	Garage Area (0.06)
8	Basement Quality: Excellent (6181)	Roof Material: Composite Shingle (87284)	Full Bath (0.04)
9	Neighborhood: NorthRidge Heights (6029)	Roof Material: Wood Shakes (86970)	Basement Finished Square Feet (0.04)
10	Year Built (5320)	Roof Material: Roll (86722)	Lot Area (0.04)

**Figure 12:** Top Ranked Microeconomic Features



**Figure 13:** Top Ranked Microeconomic Features via Random Forest Regression<sup>[6]</sup>

The overall top features for the micro dataset are: Above Ground Area, Roof Material: Clay/Tile and Overall Quality. Considering the rank of these variables across the three different models, these have the highest coefficient weight or feature importance. Roof Material in the original dataset is a categorical variable. Due to one-hot encoding, its potential values became their own variables that we used in the training and test sets. The values were types of materials like: shingles, membrane, metal or gravel and tar. The roof material clay/tile appeared as a top variable for Linear and Ridge regressions. Due to the negative coefficients, there is an inverse relationship between this material roof and the price of the home. Above Ground Area is a numerical variable of the total above ground square footage of the house. The Overall Quality variable is a measure of the overall material and finish of the house. This is a scaled value from one (very poor) to ten (very excellent). This appears as a top variable for Ridge and Random Forest.

## Summary Performance

From the analysis above, we believe that the Random Forest Model performed the best for the Macro factors. Of the other models, we would rate Ridge as the next best model as it had a very low error rate after variable reduction of 1% (when the threshold is within 10%). Linear regression performed fairly well too but the error rate and MSE was higher indicating that the precision and accuracy were lower as the variance from the target values were larger.

In addition, we believe that the Linear regression model performs the best for the Micro factors. This is mainly due to the error rate being the lowest by quite a margin at 3% (when the threshold is within 10%) while the other models had error rates between 14% and 16%. However, there is some concern in regards to overfitting and a poor training model as the R-Squared value was negative prior to the test model and further training splits would be recommended to determine if the model can be utilized for other housing data. It was promising to see that the model iterations with various different training and test splits we were able to implement led to the improved accuracy and precision seen. Outside of this model, we could utilize the Random Forest Regressor model for the Micro factors as a second option because the

MSE is marginally higher than Lasso and Ridge but the error rate is the lowest of the three at approximately 14%.

Finally, addressing the initial question for this paper, the top macroeconomic features that we believe impact the housing price index are population, M3 Money Supply, and PPI\_Res. The most impactful variables for Sale Price are the Above Ground Area, Clay/Tile roofing material and Overall Quality ranking as discussed in detail previously.

## **Further Considerations**

During the variable selection process, we note that the variables that we drop improves the Linear regression drastically but only marginally makes an impact on the Ridge and Random Forest processes. This is an interesting takeaway and something we would want to look into further. In addition, with the Micro Factor data, there are over 80 features that go into micro analysis of house prices. To reduce overfitting, a future project would focus on maybe 15-20 features max. As studies have shown, as you add more and more features, the R-Squared can continuously go up, hence giving a false indication of performance and fitting. In addition, it is important that our takeaways on best models for future Micro and Macro Factor analysis will differ from city to city or even from country to country as what is considered important will change based on infrastructure and norms.

Principal Component Analysis is utilized in this project for more exploratory analysis and data cleaning in our research project. However, it could be used to spot more outliers and even as a variable selection method in the future, specifically for micro factors as the most important or prominent features may differ from area to area and there may be very clear outliers depending on the area looked at.

Through our research, one of the key downfalls for the Random Forest model <sup>[2]</sup> is that the predicted values are not able to be outside the training set values for the targeted feature or variable. Hence, if the training set is not holistic of the maximum and minimum values of the data, there will be issues with predictions (extrapolation) of those said points. Additionally, the Random Forest model takes a lot of computational power for large datasets so if this research was broadened to larger areas, it could be a downside based on resource availability.

## **Breakdown of Effort**

Both team members played a role in the execution or revision of all major analysis components, the construction of the report, and modeling techniques. However, below is where one team member had slightly more of a focus:

Namrata Buxani - Methodology, Standardization, Data Exploration, PCA Analysis, Linear Regression, Gradient Boosting, Statistical Performance Analysis/Validation, Evaluation and Results

Chonel Chase - Methodology, Standardization, Lasso Regression, Ridge Regression, Random Forest Model, Statistical Performance Analysis/Validation, Evaluation and Results

## Appendix

1. Wright, S. (2022). *Stephen Wright*. <https://stephenallwright.com/good-mse-value/>
2. Mwit, D. (2022). *MLOps Blog*. MLOps Blog. <https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why>
3. Yiu, T. (2022). *Understanding Random Forest*. Toward Data Science. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
4. Moffitt, C. (2017). *Guide to Encoding Categorical Values in Python*. Practical Business Python. <https://pbpython.com/categorical-encoding.html>
5. Koehrsen, W. (2018). *Hyperparameter Tuning the Random Forest in Python*. Practical Business Python. <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>
6. Płoński, P. (2020). *Random Forest Feature Importance Computed in 3 Ways with Python*. mljar. <https://mljar.com/blog/feature-importance-in-random-forest/>
7. Savva, C. (2006). *Factors Affecting House Prices in Cyprus: 1988-2008*. Research Gate. [https://www.researchgate.net/publication/228632026\\_Factors\\_Affecting\\_House\\_Prices\\_in\\_Cyprus\\_1988-2008](https://www.researchgate.net/publication/228632026_Factors_Affecting_House_Prices_in_Cyprus_1988-2008)
8. K. (2022). *House Prices - Advanced Regression Techniques*. Kaggle. <https://www.google.com/url?q=https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data&sa=D&source=docs&ust=1670306598882353&usq=AOvVaw0KBCrfD6j3vy5k-CBkoTXo>