Introduction and Problem Definition

At the outset, our team wished to create a user friendly interface that uses an advanced pricing model to output predicted housing prices using various attributes for a given property. Buying and selling a home is often one of the largest personal financial decisions for most people. Case et al (2001) show that home values have a positive correlation with consumer spending and savings. Thus, housing prices are of great interest to central banks, policy makers, governments in addition to potential and current homeowners (Schulz & Werwatz, 2004). Dong et al (2020) illustrate the importance of understanding how the housing market is central to macroeconomic planning.

Literature Survey: Pagourtzi et al (2003) and Graczyk et al (2009) review common practices of real estate valuation. Basic methods include reviewing comparable transactions, replacement value, investment income method, etc. Currently, homeowners and realtors frequently use the comparable transactions method. Per Pagourtzi and Graczyk, hedonic regression (regression technique which explains how various factors estimate the price of a good) and artificial neural networks (ANN) represent more advanced practices for house valuation. Kim et al (2020) demonstrate machine learning approaches perform better than the comparables method. However, advanced models can be too complex for most users limiting their adoption. Huang (2019) compared various machine learning techniques against Zillow's Zestimate and found that most models performed worse concluding that additional complexity may not be beneficial.

Rosen (1974) and Kauko (2002) address the lack of theoretical justification of hedonic regression pricing models which provides a framework for its use as a valuation model. Selim (2009) applies a similar method to estimate the effects of 46 property attributes on housing prices, which successfully identifies the significant determinants. However, their work does not use macroeconomic or geographic factors. Tse (2002) points out that the hedonic regression pricing models ignore correlation between location and property characteristics and suggests stochastic coefficients on variables (e.g., housing age, neighborhood quality, etc.) to measure how their effects vary across space.

Fan et al (2006) use techniques such as decision trees to identify the housing attributes that best explain and predict housing prices using a dataset of Singapore public houses. They conclude that their approach was superior to hedonic pricing; however, tree approaches usually identify one significant feature per node which could oversimplify the problem when there are many attributes. Park & Bae (2015) and Jha et al (2020) study additional important attributes such as mortgage rates, public school ratings, etc. They compare and contrast a variety of machine learning algorithms such as C4.5, RIPPER, Naive Bayesian, and AdaBoost to a set of Fairfax County (Park & Bae, 2015) and Volusia County (Jha et al, 2020) homes. They conclude that certain machine learning algorithms have predictive power than relatively simple hedonic pricing models. However, only limited work has been done in selecting the ideal ANN model. We plan on exploring some of these ANN algorithms in our work. Truong et al (2020) compared multiple machine learning approaches finding drawbacks for each and highlighting that time complexity blooms when maximizing for accuracy. Poursaeed et al (2018) propose a novel Convolutional Neural Networks (CNN) algorithm to analyze listing images to predict prices; however this method appears to be open to manipulation by realtors and is not something we plan to explore. Limsombunchai, Gan & Lee (2004) compared the predictive power of hedonic model and ANN model on a dataset of 200 homes in New Zealand finding a poor performance of the hedonic model with out-of-sample data compared to the ANN models; however, they note

that the blackbox nature of ANN could lead to difficulties in interpreting results. For our analysis, accuracy in out-of-sample data is more important since the goal is to help homeowners and prospective homeowners understand housing pricing. Gu, Zhu & Jiang (2011) created a housing price forecasting model by using a genetic algorithm and support vector machine (G-SVM) approach which optimized the training parameters of SVM and reduced the time required for parameter optimization. Mu et al (2014) show that SVM models perform better than partial least squares with nonlinear data. We note inconsistent conclusions about the use of hedonic pricing and ANN models as well as drivers of housing prices in our literature survey; at least some of these inconsistencies are due to algorithm choice and feature selection.

Proposed method: As described in our literature survey, the comparable sales method, one of the most common valuation methods suffers from questionable reliability due to the subjectiveness of variable selection. From our literature review, we conclude that a Machine Learning (ML) model based method represents a more objective approach to valuation. More accurate valuations lead to better capital allocation. We innovate by: 1) considering macroeconomic factors that our literature survey doesn't seem to focus on (e.g., stock market prices and interest rates) and 2) using a variety of ML techniques and choosing the "best" (as defined by our chosen criteria) for a given dataset to predict prices. The following sections include details across data gathering, cleaning, analysis and modeling, visualizations, and summary activities which are denoted as by respective team member names, where appropriate. At the time of this final submission, all sections needed for an initial proof-of-concept have been completed.

Data gathering (Raghuveer): At the outset of the project proposal, we had planned to use Zillow Group's API to extract live data on real estate property features and past transactions. However, in the recent past, Zillow has restricted information available via public API access to largely aggregated datasets. While they offer commercialized API access that has more detail, this too has highly restrictive terms, i.e., we cannot save down the data for later use. Without detailed property level information, we were unable to move forward with implementing ML modeling. Real estate property transactions and assessments are publicly available in most jurisdictions albeit not always online or via API. Some cities and counties offer this rich dataset via publicly accessible APIs. For this project, we narrowed our scope to focus on two locales, namely, City of Philadelphia and Fairfax County. The datasets were downloaded via Python using their public API. The datasets we downloaded were about ~2-3 million rows and across 70+ columns. We also downloaded datasets from Hartford, CT and San Francisco, CA amounting to an additional ~3 million rows. For reasons discussed below, we excluded these two datasets. Overall, we downloaded ~3GB of raw data. Next, we gathered economic and financial data: 1) a range of interest rates at monthly intervals, 2) inflation and unemployment data specific to the locales and nationally, and 3) stock market index returns in each month. The rationale for gathering this information is that higher interest rates, higher unemployment reduce demand for homes and negatively impact home prices.

Data cleaning and processing (Raghuveer): After data gathering, we cleaned the data for use in our models. We needed to address the following: 1) align dimensions across datasets, 2) rename dimensions consistently, 3) exclude erroneous data, and 4) exclude other outliers. Downloaded datasets were rich, offering detailed property-level information. For example: street address, number of bedrooms, square footage, construction quality, basement details, etc. In addition, they provided the last sale date and price. San Francisco's dataset is mostly feature

complete, however, it crucially lacks actual transaction data (when a property is sold and for how much) leading to its exclusion. This transaction information is critical because a real transaction is the clearest representative of actual market value that will be used to train our ML models. The datasets downloaded were across several tables containing different pieces of information. We joined across different tables, excluding erroneous data and outliers to generate the final clean dataset. One example of excluded information is: some counties report a deed transaction where the home is transferred from an individual to their own trust (not a real transaction) with a transaction value of $1 (or nominal amounts). The final cleaned datasets across the two locales has ~1 million data points with 20+ selected columns/features.

Next, we planned to join this data against the Google Maps API to download additional features for each property such as: distances to closest freeway, public transit, grocery stores, gas stations, parks and recreation areas, high schools, etc. However, after review by Daniel, the Google Maps API was found to be prohibitively expensive; we estimated the cost of downloading the required additional features to be >$8k, putting it out of scope for this project. Our hypothesis was that proximity to transportation, grocery stores, highly rated high schools, and parks will increase home values while proximity close to undesirable businesses would reduce values. As of this report, the complex data gathering and cleaning process has been completed. (As an audit to confirm accuracy of our dataset, we compared randomly selected public data against Zillow's information for the property.)

Analysis and Modeling (Entire Team): The objective of this activity is to explore the cleaned dataset and build ML models to predict real estate prices. From the literature survey, our team has shortlisted a number of analytical models that can be used to predict real estate prices. The following ML models will be explored: Random Forest, Regression Model, LASSO Variable Selection, ARIMA/Time Series Analysis, Support Vector Machine (SVM), and Gradient Boosting. We compared the efficacy of these models across the following metrics: 1) Root Mean Square error (RMSE), 2) Coefficient of determination (R^2), and 3) Mean Absolute Error (MAE).

Random Forest (Chelsea): A Random Forest (RF) regression model written in Python was used to explore the City of Philadelphia dataset. As discussed, economic datasets were joined. The following preparations were applied before model creation. Next, some additional cleaning and alignment steps were undertaken (See appendix 1 for cleaning steps.) 70% of the dataset was randomly selected for model training (See appendix 2 for a list of features).'N-estimator', 'max_features' and 'max_depth' were set to 500, 'sqrt', and 100, respectively. Default values were used for other parameters. Table 1 shows the performance of the RF model (test dataset). RF model resulted in 0.74 R^2 which was comparable or even better compared to previously reported values (0.69-0.9) from hedonic price models and neural network models (Limsombunchai, Gan, Lee 2004). The MAE and RMSE of the RF model was 54500.2 and 159480.1, respectively. RMSE in this study was significantly lower than the values (449111.5 to 1435810.8) reported by Limsombunchai, Gan & Lee (2004). This indicates our approach achieves respectable performance compared to contemporary approaches. The training time of RF was over 5 minutes due to the large dataset.

Random forest model was also applied to predict housing prices for all the properties assumed they were listed in Dec 2020. These predicted values were used in Tableau to generate interactive visualizations.

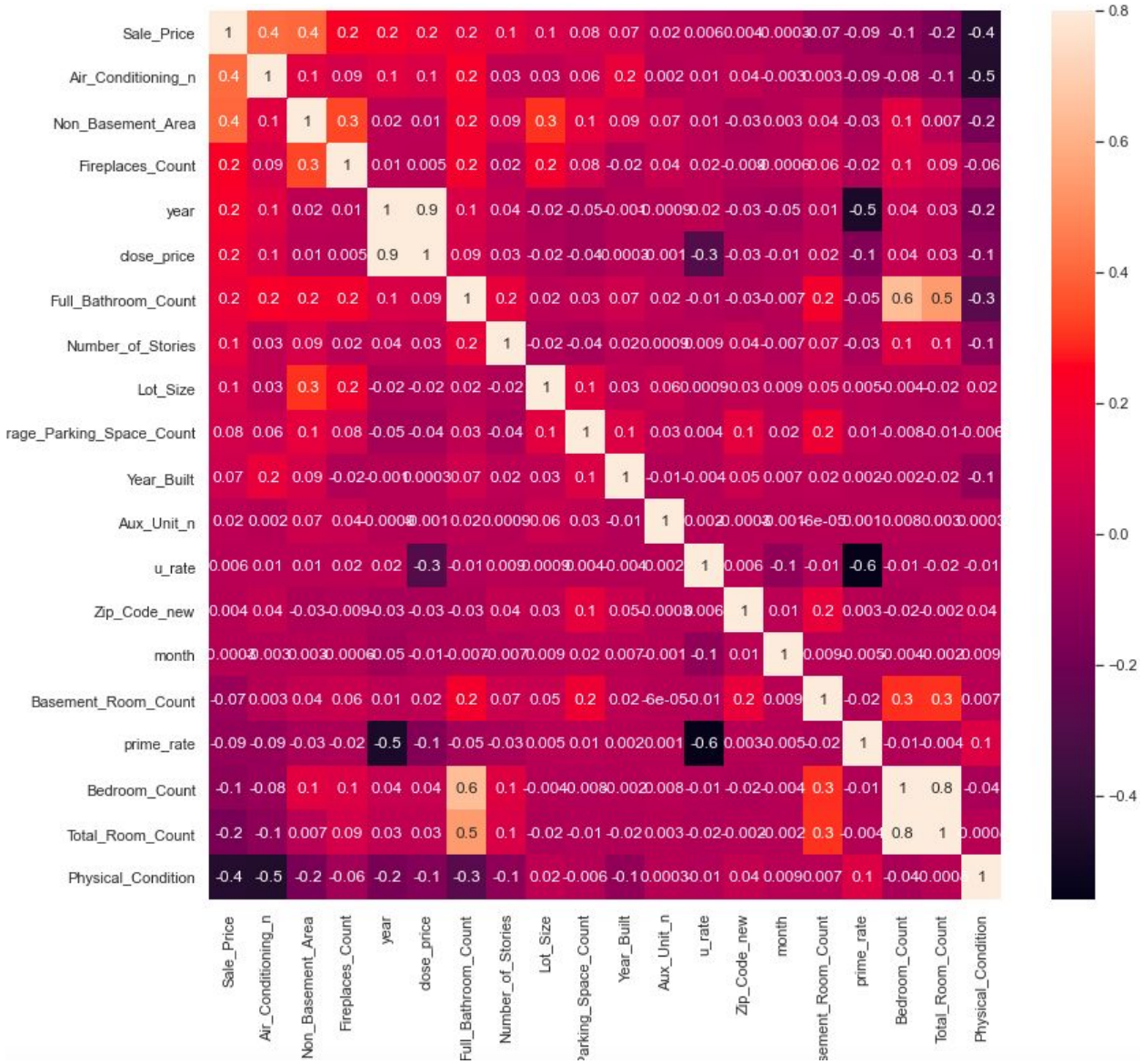
Regression Analysis (Malika)**:** The Philadelphia housing price dataset was applied to build a Multi-Linear Regression Model using Python. Regression analysis is a statistical method that analyzes the relationship between two or more variables. Explanatory analysis was completed to understand the strongest relationship between independent (e.g., Building Style, Bedroom Count, etc.) and dependent (i.e., Sale Price) variables leading to 14 variables that have the strongest impact on property's sale price. Completed regression model led to 0.199 R^2, 125406.1 MAE, and 584101.7 RMSE values. Given low values for R^2, this model is not a good fit for the real estate pricing model and therefore was not selected for predictions.

Support Vector Machine (SVM) and Gradient Boosting (Daniel): A Support Vector Regression (SVR) model and a Gradient Boosting model were written in Python to test performance on the Philadelphia and Fairfax datasets. First, economic and stock price data was joined to the dataset. The datasets were explored by plotting features against the sale price to determine and observable trends and/or outliers. Significant price and square footage outliers were removed from the dataset. Through the data exploration, it was found that square footage had one of the highest correlations with the sale price and economic factors such as unemployment and prime rate had very low or negative correlation with sale price. Visualizing the physical condition rating through a box plot showed that smaller rating numbers were associated with higher prices. Lastly, unnecessary features (e.g., parcel ID) and features with a lot of missing data were removed, and the remaining features were converted to all be numeric. Given the tendency of home prices to increase over time, the first 85% of the data in terms of date sold was used as the training set. For the Gradient Boosting model, a gridsearch was performed to determine the best parameter values for n_estimators and max_depth. The result was a value of 100 for n_estimators and 9 for the max_depth; default values were used for the remaining parameters. Table 1 above shows the performance results; the Gradient Boosting model achieved an R^2 of ~.65 on the test set and an RMSE of 169,781. This performance did not surpass the Random Forest model, but shows that Gradient Boosting provides a benefit relative to a multi-linear regression approach. For the SVR model, the feature data was normalized and PCA used to narrow the number of components to 15 based on the explained variance. Different Kernel and C values were compared to determine optimal parameters. Although each of the kernel methods performed similarly, surprisingly the linear method performed slightly better than RBF and Polynomial. The final model was trained using a linear kernel and C set to 1000. This model achieved an R^2 of .57 and an RMSE of 192,915. Overall, both models showed improvement over multi-linear regression with Gradient Boosting outperforming SVR, but the results did not improve upon the performance of the Random Forest model. Although we built models for both Philadelphia and Fairfax County, the absence of proper GIS data in the Fairfax dataset led to its eventual exclusion because we were unable to visualize it.

Table 1. Evaluation metrics for prediction models on the City of Philadelphia dataset

Model	R^2	MAE	RMSE
Multi-Regression	0.19	125406.1	584101.7
Random Forest	**0.74**	**54500.2**	**159480.1**
Gradient Boosting	0.65	69507.0	169781.0
Support Vector Machine	0.57	94301.0	192915.0

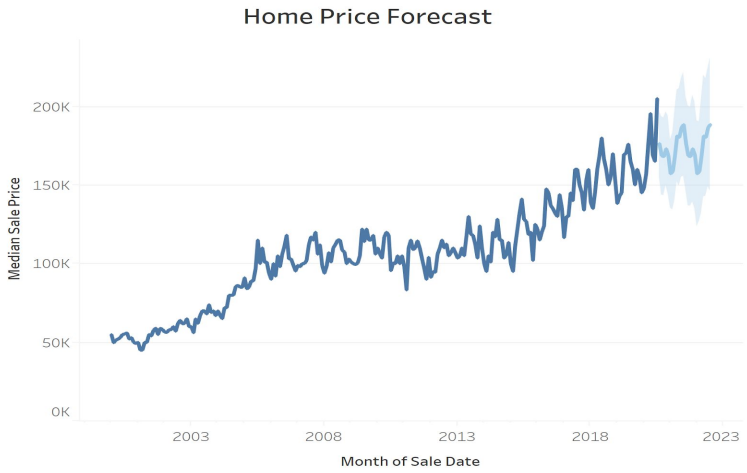
Matrix 1: Correlation matrix heatmap of various features of City of Philadelphia



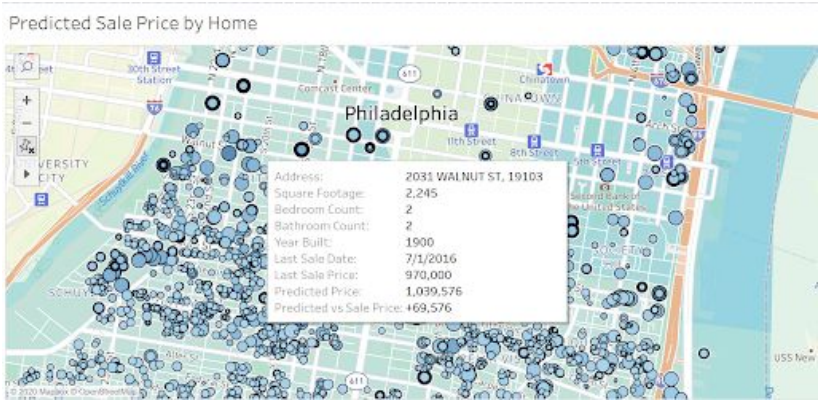
Data Visualization: The ultimate objective of this activity is to provide an interactive visualization tool that can illustrate model inputs (user filled: zip code, home features, etc.) and outputs (estimated cost, confidence intervals, etc.).

Visualization using Tableau (Aubry and Mitch): Using Tableau, we were able to create a dashboard with various visualizations to show the median house price sold and for given parameters (such as square footage, bedrooms, bathrooms, etc.) and dynamically show the predicted values by month and by individual home if it were listed today. The “Home Price Forecast” visual shows the historical median sale price by month starting in 2000 and forecasts the price for the future two years by month using SARIMA modeling available in Tableau. This gives us an understanding of the seasonal and overall trend impacting the Philadelphia housing market in the near future. Because we cannot predict how some of the other factors (macroeconomic, etc.) will change in the future, it is not possible to predict values too far into the future. For example, the below visual shows the historical median price for homes with 1 to

4 bedrooms, 1 to 4 bathrooms, and square footage between 501 and 2500 sqft. The data used is through June 2020 at which point the model predicts the price going forward along with a lower and upper bound estimate. This gives the end user the ability to change the parameters to a size of home that is of interest to them and understand the trend in the Philadelphia market to figure out when a good time would be to buy within the next two years.



Another way to view the data is by the “Predicted Sale Price by Home” visual which plots all of the homes fitting the specified parameters described above and gives us an understanding of the square footage size and predicted values of the homes in relation to one another by looking at the size of the circles and the color respectively. Upon zooming in and hovering over individual homes, users can see several attributes of the home along with the last sale price, date, and most importantly, the predicted price of the home if it were listed and sold today.



Summary and Conclusion (Entire Team): The final result of our project is an interactive and user friendly Tableau dashboard which utilizes Random Forest algorithms to predict housing prices in the Philadelphia’s market. Using this dashboard, a user can select desired attributes of zip code, desired square footage range, bedroom count, bathroom count, air conditioning, and physical condition to predict property’s price. Furthermore, using the “Home Price Forecast” view of our dashboard, one sees that Philadelphia’s housing market was strong over the last 20 years with some dips due to the recession. It is predicted to be stable in the upcoming next couple of years based on our assumptions about macroeconomic factors remaining fairly stable. Over the course of this project, all team members have contributed a similar amount of effort to this project.

References

1. Case et al (2001). Case, K. E., J. M. Quigley, and R. J. Shiller. (2001). “Comparing Wealth Effects: The Stock Market Versus The Housing Market,” Working Paper W8606, National Bureau of Economic Research.
2. Dong et al (2020). Dong, S., Wang, Y., Gu, Y., Shao, S., Liu, H., Wu, S., & Li, M. (2020). “Predicting the turning points of housing prices by combining the financial model with genetic algorithm.” PloS one, 15(4), e0232478.
3. Graczyk et al (2009). Graczyk M., Lasota T., Trawiński B. (2009). “Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA.” In: Nguyen N.T., Kowalczyk R., Chen SM. (eds) Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. ICCCI 2009. Lecture Notes in Computer Science, vol 5796. Springer, Berlin, Heidelberg.
4. Gu, Zhu & Jiang (2011). Gu, J., Zhu, M., & Jiang, L. (2011). “Housing price forecasting based on genetic algorithm and support vector machines.” Expert Systems with Application, 38(4), 3383–3386.
5. Huang (2019). Huang, Yitong. (2019). “Predicting Home Value in California, United States via Machine Learning Modeling.” Statistics, Optimization & Information Computing. 7. 10.19139/soic.v7i1.435.
6. Jha et al (2020). Jha, S. B., Pandey, V., Jha, R. K., & Babiceanu, R. F. (2020). “Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study.” arXiv preprint, Aug 2020.
7. Kauko (2002). Kauko, T. (2002). “Modelling the locational determinants of house prices. Neural network and value tree approaches.” Utrecht: Universiteit Utrecht (2002).
8. Kim et al (2020). Kim, Y., Choi, S., & Yi, M. Y. (2020). “Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices.” Sustainability, 12(14), 5679.
9. Limsombunchai, Gan & Lee (2004). Limsombunchai, V. , Gan, C., Lee, M. (2004). “House price prediction: hedonic price model vs. artificial neural network.” American Journal of Applied Sciences, 1(3) 193–201.
10. Mu et al (2014). Mu, J., Wu, F., & Zhang, A. (2014). “Housing value forecasting based on machine learning methods.” In Abstract and Applied Analysis (Vol. 2014). Hindawi. Article ID 648047, 2014.
11. Pagourtzi et al (2003). Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T. and French, N. (2003), "Real estate appraisal: a review of valuation methods." Journal of Property Investment & Finance, Vol. 21 No. 4, pp. 383-401.
12. Park & Bae (2015). Park, B., & Bae, J. K. (2015). “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data.” Expert Systems with Applications, 42, 2928–2934.

13. Poursaeed et al (2018). Poursaeed, O., Matera, T. & Belongie, S. (2018) “Vision-based real estate price estimation.” Machine Vision and Applications 29, 667–676 (2018).
14. Rosen (1974). Rosen, S. (1974). “Hedonic prices and implicit markets: product differentiation in pure competition.” Journal of Political Economy, 82(1), 34-55.
15. Schulz & Werwatz (2004). Schulz, R., Werwatz, A. “A State Space Model for Berlin House Prices: Estimation and Economic Interpretation.” The Journal of Real Estate Finance and Economics 28, 37–57 (2004).
16. Selim (2009). Selim, H. (2009). “Determinants of house prices in Turkey: Hedonic regression versus artificial neural network.” Expert Systems with Applications, 36(2), 2843–2852.
17. Truong et al (2020). Q. Truong, M. Nguyen, H. Dang, B. Mei, “Housing Price Prediction via Improved Machine Learning Techniques,” Procedia Comput. Sci. 174 (2020) 433–442.
18. Tse (2002). Tse, R.Y.C. (2002). “Estimating neighbourhood effects in house prices: towards a new hedonic model approach.” Urban Studies, 39 (7), 1165–1180.

Appendix

1. Random Forest (RF) Regression Model: Additional cleaning steps:
    a. A new variable called ‘Age at Sale’ was created, representing the age of a property at sale adjusted by remodeling situation.
    b. Categorical variables were converted to dummy variables. For example, ‘Tax district’ feature has too many categories, so any category with less than 500 values was removed to reduce complexity.
    c. Records were dropped if:
        i. ‘Sale Price’ was $\leq$ \$100 or $\geq$ \$10 million
        ii. ‘Age at Sale’ is negative (county erroneously reporting remodel as original built year due to poor records)
        iii. Record has ‘NA’ values.
        iv. Bedroom number > 7
        v. ‘Building category’ is multi-family
    d. In total, 26% rows were removed from the original dataset.
2. Random Forest (RF) Regression Model features:
    a. Property features:
        i. 'Zip code', 'Sale Year', ‘Age at Sale’, 'Bedroom Count', 'Full Bathroom Count', 'Total Room Count', 'Lot Size', 'Number of Stories', 'Fireplaces Count', 'Basement Garage Parking Space Count', 'Basement Room Count', 'Non Basement Area', 'Physical Condition', 'Building Category’*, ‘'Air Conditioning’, 'Aux Unit’, 'Elevation Type’.
    b. Macroeconomic features:
        i. ‘Unemployment rate’, ‘CPI’, ‘Federal Funds Rate' and 'Prime Rate'.

    *Single family was the only building category in the dataset.