Final Project

Analyzing Public Use Microdata Samples (PUMS) to Predict the Property Values in Oklahoma

Team Data Brewers

Ismael Beraza

Rohit Deshpande

Sujana Chamarty

Trevor Tippetts

University of Central Oklahoma

MSBA Advance Applied Analytics

Dr. Ho-Chang Chae

July 30, 2020

Table of Contents

Executive Summary 4

Introduction 5

Literature Review 5

Data Understanding 47

Data Collection 47

Data Description 47

Data Preparation 48

The following table consists of the imported variables in SAS® Enterprise Miner™: 49

Table 1: Data Dictionary (including roles, measurement levels and reason for rejection) 49

Exploratory Analysis 57

Summary Statistics 57

Interval Variables: 57

Distribution of the Interval Variables: 57

Data Visualization 65

Modeling and Evaluation 71

Data Partition 71

Transformation, Replacement, and Imputation 72

Modeling 73

High Performance Data Mining Nodes (HPDM) 74

Variable Clustering 77

Cluster Analysis 77

Decision Tree 84

Maximal Tree 84

Neural Network 88

Neural Network with Stepwise Regression 88

AutoNeural with Stepwise Regression 89

Control Point 94

Model Comparison: 94

Scoring 95

Recommendations 97

Conclusion 97

References 99

Executive Summary

For institutions that are inundated with data that tracks property values, they hope to improve their models which measure the dynamic economy of housing prices. House price valuation is one of the most important trading decisions that affect realtors and banks (Park & Bae, 2014). There can be an intentional and material misrepresentation of home value which we saw clearly during our last housing bubble that toppled lending institutions. Many commentators have also noted egregious unethical behavior on the part of lender, brokers and borrowers (Buttimer, 2011). When an individual or an investor applies for a mortgage, a bank initiates the process to appraise the home. An appraiser's value of a home has great influence in pushing prices up or down within their authority. Accountability in how a home is valued is subjective and remains at the discretion of the individual. Modeling, which is accessible and relatively easy to interpret, without the hindrance of hiding proprietary algorithms, is elusive for consumers and advantageous to banks and realtors, its asymmetric information is problematic. Machine learning is a technique of data analysis that automates analytical model building. It is a technology based on the idea that systems can learn from data, identify patterns and make decisions (SAS, 2020). In return, the models will be used to bring about a clear interpretation of the data and a consistency in property values. Machine learning models like decision trees, regressions, neural networks will be used to derive an optimal residential real estate value in a dynamic economy. This research will augment other property appraisals, as well as validate mortgage loans and lending decisions here in our very own neighborhoods.

Introduction

Description

The purpose of this project will be to explore U.S. Census Bureau PUMS (Public Use Microdata Samples) data to identify characteristics that are statistically related to variables which indicate property values in the state of Oklahoma. The two datasets, 'Oklahoma Population Individual Details' and 'Oklahoma Housing Records' will be scrubbed and merged using Python programming and the analysis will be facilitated by using SAS® Enterprise Miner. We would like to analyze and predict property values of residents within Oklahoma which would allow us to generate recommendations to banks, to realtors, and to consumers who want to purchase a home.

Literature Review

Review of Existing Literature

There are various sources which delve into statistical machine learning algorithms to predict and explain the factors affecting house prices, mortgage rates, income and other traits. However, there is not enough research on utilizing a statistical model to identify both individual and property descriptors related to the state of Oklahoma. Research completed by Liew (2013), examined the relationship between house prices and housing attributes. The results reinforced the data that homebuyers focus on the basic housing characteristics of a two, three, or four-room flats that relate to floor area, model type and age. Another study was conducted to determine which factors impact the market value of a housing unit and the findings of the research will principally help avoid overvaluation of housing units. The results can be used to help individuals estimate housing prices as a fraction of their wages if they relocate. After running the data through SAS® Enterprise Miner we found that the order of importance of the variables selected by the surrogate decision are fair market rent, insurance (monthly insurance cost), other costs (monthly insurance cost), structure

type (number of buildings in the building), ZINC2 (household income), and ROOMS (total number of rooms in the unit). The findings revealed that fair market rent, insurance and monthly insurance cost are the two most significant variables in examining the market value of a housing unit. It also verified that single houses and apartment complexes in the Midwest & South region have comparable market values. To those individuals looking to relocate in that region can be assured that homes values are reasonably assessed. Further findings revealed that the number of units in a building do not impact the current market value of a housing unit (Tanjil et al. 2016).

Forecasting property value is essential in helping buyers and investors make decisions about budget allocation, finding property funding strategies, and deriving suitable policies. Predicting housing prices can be tricky since market conditions are continuously changing. The main objective of this literature review is to dive into several research studies to explore the determinants of housing prices and to discover what techniques were used to predict home value. To provide a framework for understanding how home value is predicted and what factors are affecting it, a total of 23 peer-reviewed journals were chosen, which contains the following topics:

1. Independent variables such as amenities within the house, individual's economic status, constructional details, features of the property, immigrants, crime rates, resale values.
2. Predicting the property prices or resale price of the house.
3. Factors affecting the real estate from 1986 to 2019.
4. Real estate demand management.
5. The application of various statistical algorithms like decision trees, types of regression analysis, types of neural networks, and few optimization models.
6. Variables rejected due to a lack of relevance and failure of statistical significance.

The priority of our research is to identify the factors that affect housing prices. Much of the peer reviewed journals and articles focus on household characteristics to predict home value. The findings depict that location and size are the two attributes that highly impacts housing prices (Park 2014; Lu 2017).

After considering the variable selection process and machine learning algorithms to examine the market value of housing units, Tanjil et al. (2016) concluded that fair market rent and insurance are the two most important determinants in his study. On the other hand, Huang's (2019), outlook in that it is fundamental to leverage the inputs such as the tax amount, the land tax value dollar count, and tax value dollar count to consider the log errors as the best data preparation process.

The housing market is influenced by economic factors like interest rates, mortgage availability, and its supply and demand. The independent variables such as disposable income, interest rate, mortgage market liquidity, money supply, and housing stock supply affect the housing prices on a long run basis. For example, a lower level of housing construction reduces supply and increases housing prices (Shaaf, 2005-2006). Some of these macroeconomic factors overlap when looking at them to explain the housing market as Dr. Shaaf's article, *An Analysis of the Housing Market and the Oklahoma Experience*, tried to explain that by splitting the variables via demand and the supply side. In another example, Dr. Shaaf's view of how interest rates effect the number of houses which influences the prices or value has been seen time and again in our current housing economy. Higher interest rates lead to more expensive housing construction financing which results in lower supply of houses, and higher housing prices (Shaaf, 2005-2006). Furthermore, income and mortgage lending showed a positive correlation with the property value (White, 2015). Wu (2017), also stated that bank lending, housing supplies, and mortgage rate have a considerable impact on predicting home value. Enweim et al. (2010) found that "digital census" information can have a

positive effect on predicting home values. Digital census data are digital records, such as mobile phone connections, bank card transactions, geo-tagged Twitter, taxi usage, and so forth. This information is used to provide a benefit to predictive models, while also adding real-time data. Jones et al. (2015) insisted on other factors, which also affect property values, like crime rates, location, and whether the property is newly built. They are fundamental to predicting house prices by means of multilevel modeling and artificial neural networks.

Physical attributes of properties have also been studied to predict the housing prices. Montero et al. (2017) considered parametric and semi-parametric spatial hedonic model variants and found that inputs such as swimming pool, elevator, garage, and shopping areas have a strong association in predicting home values. Lu et al. (2017) suggested that high amenity areas such as pool area, garage, gym, and fence make a significant difference in housing prices among different neighborhoods.

Evaluation of Methodologies

Researchers utilized several machine learning techniques to ascertain the important variables in estimating home values. Plakandaras (2015) used distinct models known as the Random Walk (RW) Model, Bayesian Autoregressive, Bayesian Vector Autoregressive Model, and Ensemble Empirical Mode Decomposition (EEMD) - Support Vector Regression (SVR) to check the forecasting ability of different models. Based on the results, the novel method (EEMD-SVR) can be effectively used for predicting prices. Park (2014) used techniques such as Naïve Bayesian, c4.5, Ada Boost and Ripper. Park determined that location and size were the two most important factors and in all the tests Ripper was the model that predicted the home value accurately.

Many researchers indicated that non-linear models are more precise than conventional linear models (Huang, 2019; Mu 2014). According to Mu (2014), Support Vector Machine (SVM) and Least Squares Support Vector Machine (LSSVM) have higher prediction accuracy than Partial Least Squares (PLS). In trying to find diversity in the results, Huang (2019) also implemented linear and non-linear regression models, such as the decision tree, boosting, random forest, and SVM. The findings showed that non-linear models are superior to linear regression models for house value prediction. Fan et al. (2006) & Huang (2019) agree that as the branch of the decision tree grows, the nodes under this branch become purer and more informative about the independent variables. Bogin et al (2019) agrees with similar research on the subject, non-linear models have better predictive power. However, the predictive power comes at great computational cost, sometimes taking days longer than linear models. While some non-linear models, like random forest, can produce better results, one would need to ensure model performance is worth the added resource time to gain those results. In addition, Hong (2019) explains the advantages of Random Forest over hedonic models. Jones (2015) and Feng (2015) compared the Multilevel Modelling (MLM), Artificial Neural Networks (ANN) to evaluate the performance in terms of model fit, predictive accuracy and explanatory power and stated that MLM performance was much better than the ANN.

Shahhosseini, Hu and Pham (2019) optimized machine learning predictions of ensemble learners by finding the best weights for constructing ensembles for house price prediction. Based on the results, XGBoost and Random Forests are the best algorithms predicting the median price of the houses with the least MSE and MAPE, and highest R- squared values.

Another article, this time written by Worzala, Lenk and Silva titled, An Exploration of Neural Networks and its Application to Real Estate Valuation, had also wanted to find an applicable and

reliable method in appraising property by using multiple regression analysis or neural network models. Their results showed that multiple regression had not proven to be reliable. Multiple regression has often produced serious problems for real estate appraisal that primarily result from multicollinearity issues in the independent variables and from the inclusion of "outlier" properties in the sample (Worza, Lenk, & Silva, 1995). Their next choice was to see how neural network models performed under the same criteria. It also showed unfavorable results, for one thing the timing needed to perform a working model was strikingly different. While the computer run time for all of the multiple regression models could be performed in seconds, the running times for the neural networks varied from thirty seconds to forty-five hours (Worzala et al. , 1995)! Results like these are not reliable if you are trying to asses' various homes. Other problems described by this article in using neural network models are that they are not easy to use, results are inconsistent between neural network packages, results are inconsistent between runs of the same neural network software, and neural networks can have very long run times.

Gap in Existing Research

All the literature review contains household characteristics, economic factors and physical attributes of properties related to the housing prices but none of them utilized education-related determinants like school enrollment or individual's information such as the age of the children on housing prices in Oklahoma. Our study will support the decision-makers to reexamine the determinants that have control over the price fluctuation of the properties by using the factors listed above. Studies show that researchers are facing challenges with the text data or images (Aquino et al., 2019).

Although, several researchers used different machine learning techniques to predict home value, most of them obtained different models as the best one. Since these studies have mixed

opinions about the model that can accurately predict the housing prices, we would like to confirm specifically which model can effectively estimate the property value in Oklahoma.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
Huang, 2019	Survey data/U.S.	24,156	2016	The dependent variable is the home value. The independent variables presented in this paper are airconditioningtypeid, bathroomcnt, bedroomcont, buildingqualitytypeid, threequarterbathnbr, calculatedfinishedsquarefeet, fips, garagecarcnt, gargaretotalsqt, heatinggorsystemtypeid, latitude, longitude, lotsizesquarefeet, numberofstories, poolcnt, propertylandusettypeid, regionidcounty, regionidzip, roomcnt, unitcnt, yearbuilt, tax valuedollarcnt, structuretaxvaluedollarcnt, landtaxvaluedollarcnt, taxamaountm, and datediff.	This paper uses machine learning models to estimate home value. The methods implemented were linear regression and some non-linear models, such as decision tree, boosting, random forest, and SVM.	The objective of this research is to predict home value based on different house attributes. Linear and non-linear machine learning methods are used to estimate the log error of the home value. The analysis revealed that the conventional linear models are not predictive for complex home data sets, while tree-based non-linear models are most precise with the lowest MSE. In spite permitting the tree choose the important inputs, preprocessed big dataset is used. Also, Boosting was conducted to select important variables like taxamount, landtaxvaluedollarcnt,taxvaluedollarcnt, showing some coincidence with the tree-based method. The absolute error plot indicated that the estimation only works precisely when the original estimation (Zestimate) is accurate. When the original Zestimate is ambiguous, the estimation will fail to predict. The study highlights that variable selection is a fundamental procedure; removing inputs by intuition is too risky. Tree-based methods are highly convenient

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						when facing a model with a considerable number of independent variables in house value prediction.
Lu et al., 2017	Survey data/U.S.	2,920	1879-2010	The variables include house and neighborhood demographics like id, mssubclass, mszoning, lotfrontage, lotarea, street, alley, lotshape, landcontour, utilities, lotconfig, landslope, neighborhood, condition1, condition2, bldgtype, housestyle, overallqual, overallcond, yearbuilt, yearremodadd, roofstyle, roofmatl, exterior1st, exterior2nd, masvnrtype, masvnrarea, exterqual, extercond, foundation, bsmtqual, bsmtcond,	Multiple regression, hybrid regression algorithms, ridge regression, lasso regression, and gradient boosting regression	The research provides results drawn from data with 79 explanatory variables for part of residential home transactions in Ames, Iowa, and predicts the sales price of each covered home transaction. The researchers propose different regression models to predict individual house prices. The results indicate house price is determined by many factors such as location, size, house type, city, country, tax rules, economic cycle, population movement, interest rate, and those could affect demand and supply. After

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
				bsmtexposure, bsmtfintype1, bsmtfinsf1, bsmtfintype2, bsmtfinsf2,bsmtunfsf, totalbsmtsf. As well as, heating, heatingqc, centralair, electrical,1stflrsf, 2ndflrsf, lowqualfinsf, grlivarea, bsmtfullbath, bsmthalfbath, fullbath, halfbath, bedroomabvgr, kitchenabvgr, kitchenqual, totrmsabvgrd, functional, fireplaces, fireplacequ, garagetype, garageyrblt,garagefinish, garagecars, garagearea, garagequal, garagecond, paveddrive, wooddecksf, openporchsf, enclosedporch, 3ssnporch, screenporch, poolarea, poolqc, fence, miscfeature, miscval, and, mosold.		conducting various exploratory analyses, the study suggests that the high amenity areas experience greater price volatility and a significant difference in sale prices among different neighborhoods. Various algorithms have been used and optimized for the best results. The root means square error calculations suggest that the prediction of cheap house prices is essential. Hybrid regressions produce better results with reduced errors.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
Montero et al., 2017	Archival data/Madrid, Spain	10,512	2010	The response variable is Log_Price (Napierian logarithm of house price (euro/m2), individual covariates- Indicator variable for good condition, built-up area, physical attributes of properties such as house type, garage, swimming pool, elevator, etc., age, areal covariates, population, immigrants, and crime rate.	Parametric and Semi Parametric spatial hedonic models.	This article considers parametric and semi-parametric spatial hedonic model variants that account for spatial autocorrelation, spatial heterogeneity. The assessment of the prediction power of the competing models included in this research has been carried out by cross-validation. The housing prices are spatially autocorrelated and also exhibit spatial heterogeneity. For this reason, ‘‘hedonic house price models’’ have been replaced in the recent literature by the ‘‘spatial hedonic house price models,’’ and new models accounting for spatial autocorrelation and spatial heterogeneity has similarly emerged. These findings indicate that most of the quantitative covariates considered in the case study, exhibit a smooth nonlinear relationship with

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						house prices. With hedonic specifications that include more than one form covariates, even if we add one more such form in the model, it does not result in significant gains in the predictive power of the model. However, this research suggests that the inclusion of a spatial drift (spatial nonstationarity) in the model significantly improves its predictive power. The spatial drift is the important term of the model capturing nonlinearity and spatial heterogeneity.
Wu et al., 2017	Survey/Hong Kong	3,747	1996-2016	The dependent variable is housing price and the predictors are housing commencement, mortgage approved, mortgage rate, vacancy, household income, rent-price ratio, HIS.	Bayesian vector-autoregressive model	The study was done to estimate the short run factors that affect housing prices in Hong Kong. They identified the correlation of the seven variables with respect to the five shocks, namely housing supply, housing demand,

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						mortgage rate, market sentiment, and bank lending. The results show that bank lending and housing supply shocks were the key factors affecting the home value. Moreover, it was found that the low mortgage rate also has a significant impact on housing prices in the short run.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
Tanjil et al., 2016	Survey data/U.S.	36,675	2013	The dependent variable of this research is the market value of a housing unit. The predictors are low-income limit, structure type, household income, monthly insurance cost, number of rooms, fair market rent, housing cost as a fraction of income, monthly utility costs, MSA areas, the year the	This paper used several modeling techniques like decision tree with different number of branches and depth, neural network with distinct number of hidden units, and different network architecture	The objective of this study is to determine which factors impact the market value of a housing unit. The findings of the research will principally help avoid overvaluation of a housing unit from borrowers' and lenders' point of view. Also, these results will help individuals estimate

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
				unit was built, census region, adequacy/condition of a housing unit and number of bedrooms.	(multilayer perceptron, ordinary radial, normalized radial, generalized linear model). As well as, Polynomial Regression, PLS (NIPALS, SVD, Eigenvalue, and RLGW algorithm), Gradient Boosting (square error and Huber Mregression loss function), Memory Based Reasoning (MBR) with only numeric variables passed through PCA, MBR with both categorical and numeric variables passed through PCA were applied to estimate the	housing prices as a fraction of their wages if they have to relocate to different places. To reduce the number of input variables, LARS, LASSO, Adaptive LASSO, Variable Selection, Stepwise regression, Variable Clustering, PCA only with numeric variables, and PCA with all variables were tested. The important variables selected by the surrogate decision are fair market rent, insurance, other costs, structure type, household income, and the total number of rooms in the unit). It was also verified that single houses and apartment complexes in the Midwest & South region will have comparable market values. Nevertheless, the number of units in the building does not impact the current

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
					actual market value of a housing unit.	market value of a housing unit.
White, 2015	Survey data /U. S.	22,314	1983-2011	The dependent variable is house prices, while Gross mortgage lending, Income, Inflation, Interest rates, Migration, Money supply and, Housing supply are treated as independent variables.	Johansen cointegration approach was used to find out the long run and short run factors affecting the housing prices.	This study aims to estimate the temporary and permanent factors affecting the housing prices in the UK. The results depict that disposable income, interest rate, mortgage market liquidity, money supply, and housing stock supply affects the housing prices on a long run basis. The author agrees that the financial crisis made lenders more risk averse. The variable 'mortgage lending' was found to be significantly related in both the long-run and short-run. Also, the long-run impact of income and mortgage lending showed a positive correlation in almost all regions. The study revealed that the role

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						of the lending industry is significant in housing price changes.
Plakandaras, 2015	Survey data / U.S.	6,732	1890-2012	This study used real house prices as the dependent variable and the fiscal policy variable, real GDP per capita, unemployment, long term interest rate, short term interest rate, inflation rate, population, real construction cost, real stock price and, real oil price as the independent variables.	The techniques used are Random Walk (RW) model, Bayesian Autoregressive, Bayesian Vector Autoregressive model, and a novel ensemble Empirical Mode Decomposition (EEMD) - Support Vector Regression (SVR) forecasting methodology.	The research was done to predict the housing prices, to check the forecasting ability of different models, and to identify the most accurate model. The data was divided into two sets, in sample and out of sample observations. The EEMD-SVR model outperformed with the lowest MAPE percent (1.985) than other models. The author then argues that this novel method can be effectively used for predicting the housing prices and thereby predicting the business cycle of the economy.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
Park et al., 2014	Survey data /U.S.	5,359	2004-2007	The dependent variable is house price and the independent variables are basement type value, baths full, baths half, bedrooms, exterior type value, exterior features type value, cooling type value, fireplaces, total square, garage spaces, hot water type value, style type value, lot sqft, parking type, elementary school rate, middle school rate, high school rate, list month, list price ratio, fixed mortgage rates ratio, adjusted mortgage rates ratio, city, zip5, year built, days on market and, high or low.	The machine learning algorithms, such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost, were used for predicting housing prices.	The study was conducted to predict the housing prices by analyzing housing data in Virginia, US. Prediction models such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost were used to compare, evaluate the performance, and check how accurately each model can predict whether the closing price is greater or less than the listing price. 28 variables were selected in the final model using stepwise logistic regression. From the results, it was clear that RIPPER model predicted the accurate prices in all the tests, and location is one of the factors that significantly impact the housing price because of the regional differences. The study concluded by proving that machine learning algorithms can precisely evaluate and predict the

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						housing prices than using other methods.
Mu, 2014	Survey data / U.S.	452	2014	The dependent variable is the values of Boston suburb houses. The inputs included are per capita crime rate by town, proportion of nonretail business acres per town, and, index of accessibility to radial highways.	The methods used to forecast the home value are Vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS).	In this research, the values of Boston suburb houses are forecast by distinct Machine Learning methods. If the housing values can be accurately predicted, the government can make a reasonable urban planning, for instance, the government and developers can make decisions about whether developing the real estate on specific areas or not. According to the predicting results of home's value of Boston suburb, Vector machine (SVM) has a higher prediction accuracy than least squares support vector machine (LSSVM), and partial least squares (PLS). The paper reveals that

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						because of the presence of strong nonlinearity about the home's value of Boston suburb in the data, the predicted result of PLS algorithm is not convenient and efficiency is considerably low.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
Fan et al., 2006	Survey data /Singapore	4,192	1997-1998	The dependent variable of this study is the resale price. At the same time, the housing attributes are treated as inputs like rsale price, resale date, floor area, flat type, age of flat,	The methodology used in this paper was the decision tree algorithm.	This research examines the relationship between the resale prices of Singapore public housing of all types and housing attributes.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
				floor level (the floor on which a flat is located), the number of rooms within a flat, upgrading situation, location, and neighborhood. The sample encompasses almost all public housing variety, such as two-room, three-room, four room, five-room and six-room flats (executive apartments), and their sub-types.		The location and neighborhood attributes of the flats are measured by their categorized straight-line distances to expressways, mass rapid transit (MRT) stations, bus interchanges, popular primary schools, industrial estates, or private housing estates. The built tree shows that home buyers focus on the basic housing characteristics of two- and three-room flats or four-room flats such as floor area, model type, and flat age. Nonetheless, homebuyers of five-room flats pay more attention to floor level in addition to the basic housing characteristics. On the other hand, Homebuyers of executive apartments are less concerned about primary quantitative attributes and pay more attention to quality and service, such as amenities and a good environment. These imply that model type,

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						floor area, and flat age are more important variables in regressing and predicting the resale prices of HDB 4-room units. It is also noticeable that, as the branch of the decision tree grows, the nodes under this branch become purer and more informative about 4-room flats.
Quigley, 1999	Survey data/ U.S.	259	1986-1994	This study is predicting real estate prices. The independent variables include a local consumer price index, employment, and income disaggregated by industry category, the number of households and total population, vacancy rates for owner-occupied housing, commercial offices, and rental housing, unemployment rates, and, the volume of the mortgage.	Different regression models (logarithmic models, percentage models, linear models).	The research paper explains the relationship between the real estate prices and the economic conditions undertaken systematically across major U.S. housing markets. Different mathematical models explain the effects upon supplier and demand behavior of differing prices in the real estate market. The results show that the increase in household income is associated with an increase in the price of owner-occupied

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						housing. In addition to the construction permits, the results show an increase in real estate prices and economical construction activity in metropolitan areas with higher prices. The prices are low in regions with higher vacancy rates. The complete specification variable explains about 29 percent of the variation in log housing prices. The results of 5 regression models from a combination of the autoregressive structure with economic fundamentals show that changes in employment, income, in the number of households, and the number of construction permits are essential determinants of the course of housing prices.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
Lee, 1999	Survey data /U.S.	2,563	1989-1991	The variables used are home characteristics, for sale properties (property control variables) include lot size (square footage), living area (square footage), house type (dummy variables for semidetached and row house), garage (dummy), masonry (dummy), and stone (dummy), property-specific attributes, period of sale, neighborhood quality, macro locational amenities, existence and, programs characteristics of assisted housing in proximal areas.	OLS Regression	This research paper focuses on public housing's effects on racial segregation, poverty concentration, and its impact on real estate prices. The results suggest that in Philadelphia, the homeownership programs have a more beneficial impact on surrounding neighborhoods than any rental program. However, the negative impact of a rental program on property values is modest on adding control variables for neighborhood characteristics and property specific. Among the property specific control variables, except masonry (being the most common material used for houses), the rest of the variables produce positive coefficients and are statistically significant.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						Masonry variable is considered important on including the neighborhood quality variables. The results of the seasonal impacts indicate that negative impacts exist on sales in winter and spring. The locational amenity variable for distance to the central business district produces a negative correlation. The variables for park and river accessibility produce a statistically positive contribution to the property values. It is also evident from the results that high sales occurred in areas with large or high-rise public housing. In conclusion, the neighborhood impacts are important for producing positive outcomes for residents; these findings would lend support to recent proposals intended

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						to diversify the population of public housing residents.
Bogin et al, 2019	Uniform Appraisal Dataset/ US	420,370	2014-2016		Hedonic regression estimated using ordinary least squares, Ridge Regression, LASSO, Elastic Net, Random Forest, Boosting	The baseline for the study is the hedonic regression using ordinary least squares. This is a common method that is likely the most widely used throughout the appraisal industry. The authors found the ridge regression performed worse than the baseline model, while the LASSO regression model improved the model slightly. These models were rejected in part due to the increased complexity over the base line model. Two different types of random forest were tested. While both forest models improved the predication capability, it came at significant cost to model efficiency, taking 48 hours to run instead of a few

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						minutes with the baseline model. Additionally, the random forest models suffered from overfitting. The authors concluded, while random forest models show promising results, its complexity and computational efficiency may not be a good trade off, unless more power can be obtained to run a model.
Enwei et al, 2010	New York City Department of Finance, 311 Complaints, New York Police Department, NYC Taxi and Limousine Commission, American Community Survey, Longitudinal Employer/ NYC, NY	83,876	2010-2015	Data comes from NYC property sales maintained by the New York City Department of Finance. 311 complaints- data collected by local governments regarding noise, illegal parking, dead trees, animals, etc. Crime complaints as reported by the New York Police Department. Taxi trips- Trip record data was obtained from the NYC Taxi and Limousine Commission. American Community Survey (ACS)- Census data	Hedonic model, Dummy Variables, Cross-Validation, Linear Regression, LASSO, Neural Network, Random Forest, Gradient Tree Boosting	The authors use the hedonic model as a baseline for measurement. The value added by the regression, neural network, and random forest models are the ability better generalize the data from different regions (zip codes, census tracts, community districts). The authors were able to better generalize the added digital census data when using fresher (real-time) information using the non-linear regression models.

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
				containing population density, unemployment, median family income, education, etc. Longitudinal Employer-Household Dynamics (LEHD)- provides data about human mobility. Individuals traveling to work from home.		These models all provided better model performance over the hedonic model, which supported the authors hypothesis that additional "digital census" data can be beneficial when seeking to predict housing prices.
McClusky et al,	Lisburn District Council Area (Northern Ireland)	2,694	2002-2004	Time Adjusted Sales Price, Size, Garage, Age, Subclass, Class and Ward	Neural Networks, multiple regression, OLS Regression	The allure to use artificial neural networks are its advantages over multiple regression analysis. Neural networks do not need binary variables or to linearize variables through transformations. Over-fitting is less of an issue now with the current appraisal software in the market. Multiple regression analysis have a relatively easy ability to be explained which provide appraisals the power to justify their results. If artificial neural networks can be adjusted to

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						withstand the dynamics of real-world applications and can be used by the appraisal industry, over time it can be used more frequently than multiple regression analysis. This paper did not provide a definitive answer to one model over another but explained deficiencies in each model and brought further discourse to arguments over which model should be used.
Shaaf, 2005	Oklahoma City, State of Oklahoma, and U.S.	N/A	2005-2006	Income and wealth growth, interest rates and the interest rate policy of the Federal Reserve, inflationary expectations, the magnitude and the degree of speculation in the housing market, and population growth Regulatory burdens regarding construction, purchasing, and foreclosures	None	Competing views between a housing boom and its demise are discussed in this paper written by Mohammad. An optimistic outlook that envisions a robust demand and a market that will sustain higher home prices are contradicted by arguments discussed by Robert Shiller. Mr. Shiller, the contrarian, brings our attention to the fact that

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						inflated prices leads to a housing bubble. Oklahoma City's median-hope appreciation was 14% between 2002 and 2005, this number may not reflect the 22% median home-appreciation experienced by our nation but can still make an impact in our surroundings (Shaaf, 2005-2006).
Worzala et al, 1995	Fort Collins (Colorado) Board of Realtors Multiple Listing Services (MLS)	288	1993 - 1994	Southeast home, Ranch home, Number of Bathrooms, Lot size, Area of basement, Total area, Garage, Selling Price, Number of bedrooms, and Number of fireplaces	Multiple Regression, Neural Networks	A common method used to validate property and real estate appraisals, known as the sales comparison approach, has not provided reliable and verifiable data. Much of the results ascertained by the arbiters of property value estimations have been criticized by their inaccurate appraised subjective values. As a result, a more robust endeavor in appraisal analysis brought about an application for exploration

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						in multiple regression analysis and its comparison to Neural Network models. However, in the research promulgated by these authors, even these two approaches were found lacking. As explained by the authors, in the use of multiple regression for asset valuation: Multiple regression has often produced serious problems for real estate appraisal that primarily result from multicollinearity issues in the independent variables and from the inclusion of "outlier" properties in the sample (Worza, Lenk, & Silva, 1995).

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						In an effort to measure the performance between multiple regression analysis and neural networks, two criteria were used: the mean absolute error between the predicted and actual sales price of the test sample and the percentage of houses in the test sample whose absolute error was less than 5% of the actual sales price. A further look to visualize which independent variables were used to measure the performance for both models: number of bedrooms, number of bathrooms, total square footage, number of garages, number of fireplaces, number of stories, lot size and the real estate's age. The authors chose to break down their results to three types of cases in

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						measuring the performance of both models. The first case was using the entire data, as is. The second case was using the data that filtered the sale prices of houses between $105K and $288K.The motivation behind this second case was their need to reexamine another study which disseminated neural networks ability in surpassing multiple regression analysis in evaluating house prices. The third case narrowed down the results to a specific zip code area. The results of the first case showed that the neural network model outperformed multiple regression analysis but only marginally. The second case also proved to show that multiple regression analysis outperformed neural

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						network models in certain price ranges. In the final case, where the characteristics of the houses were more homogenous, the neural network model outperformed the multiple regression analysis results. In the end, neural network modeling did not outshine multiple regression analysis and the results were not as consistent as was hoped with using this type of analysis. In certain cases, the processing time for the neural networks varied from thirty second to forty-five hours. This type of reporting, in this type of market, is unacceptable and cannot be fully implemented without a more definitive outlook.
J. Hong et al., 2019	South Korea's Ministry of Land,	16601	2006–2017	Elapsed year (transaction year-construction year), Area, Floor level of a property, Heating system,	2 techniques have been used in this research appear and	The main objective of this paper is to forecast the housing price and comparing the two

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
	Infrastructure, and Transport.			Apartment brand, Number of units in the apartment complex, Number of buildings in the apartment complex, Parking lot, Floor area ratio, Building coverage ratio, The top floor of an apartment, The lowest floor of an apartment, Latitude, Longitude, Distance to national park, Distance to high school, Distance to redevelopment area, Distance to university, Distance to general hospital, Distance to museum, Distance to subway station, Transaction period, Gross domestic product (GDP), Annual growth rate in real GDP, Land price fluctuation rate in Seoul, Mortgage interest rate.	evaluated the best model/ approach to determine the housing prices. Conventional hedonic pricing model estimated by OLS regression and Random Forest which is an ensemble of Decision Tree's model.	statistical models which have been used to predict the house price. This paper explains the features of the Random Forest (RF) predictor in comparison to the conventional OLS-based predictor. This paper proves that the predictive performance of a machine learning-based predictor can be superior to that of the OLS-based approach. Researchers used apartment transaction data from 2006–2017 in Gangnam, one of the most developed areas in Korea. They collected data set covering 40% of all transactions in the selected area, and the samples were randomly divided into training sets consisting of 90% of all transactions and test sets consisting of the remaining 10% of transactions. They also used the averages of 10 experiments to compare

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						the performance measurements in order to eliminate the possibility that the results occurred by chance. The average percentage deviation between the predicted and actual market price was only around 5.5% for the machine learning predictor and almost 20% for the OLS-based predictor
Aquino et al., 2019	Survey data/Brazil	12223582	2015 to 2018.	id, floors, rooms, collected_on, property_id, operation, property_type, place_name, place_with_parent_names, country_name, state_name, geonames_id, currency, description, title, lat_lon, lon, lat, surface_covered_in_m2, surface_total_in_m2, expenses, proce, image_thumbnail	Random forest (RF) and recurrent neural networks (RNN)	The objective of this research is to predict property prices based on housing advertisements. This study proves that enriching the dataset and combining different ML architectures can be a better alternative for the prediction of housing prices in Brazil. This paper shows the importance of deep learning architectures since their high performances on massive amounts of data. The enriched random forest

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						works well with numeric variables. However, it cannot manage raw or text data or image. On the other hand, KISS represents all types of data through the embedding layers, but it did not address numeric features as RF does. It was concluded that combining the strengths of these two methods yielded an accurate result.
Adelino et al, 2015	ScienceDirect/ U.S	775 counties	2002-2007	Total employment, Unemployment rate, Percent college educated, Percent employed, Workforce as a percentage of population, Percent of homes owner-occupied, Average household income, Growth in income, Growth in house prices, Number of counties	Two stage least squares regressions	The study was done to identify the causal effect of housing prices on small scale business. The instrument used for house price growth is the Saiz measure of housing supply elasticity. The results depict that there is a causal effect of increase in the house prices with the creation of small-scale companies. The effect was found more in industries like manufacturing, that need startup capital and in

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						which housing collateral is more important. Collateral lending channel through real estate plays an important role in shaping employment dynamics. It was also found that there is no causal effect on employment at large firms
Shahhosseini, M.,Hu G.,Pham H.	U.S. Census Service regarding housing information in the Boston metropolitan area and Ames, IOWA	2930	2006-2010	YearBuilt,Neighborhood, Street, BldgType, MSSubClass, Foundation, LotArea, RoofStyle, Bedroom, FullBath, TotalBsmtSF, 1stFlrSF, TotRmsAbvGrd, GrLivArea, GarageCars, GarageArea, OverallQual, ExterQual, KitchenQual, BsmtQual, SalePrice	Multiple learners including LASSO regression, Random Forests, Deep Neural Networks, Extreme Gradient Boosting (XGBoost), and Support Vector Machines with three kernels (polynomial, RBF, and sigmoid) have	The objective of this paper is to optimize machine learning predictions of ensemble learners by finding the best weights for constructing ensembles for house price prediction. Two housing datasets for Boston and Ames have been chosen to demonstrate and validate the optimization model. . The predictions made by base learners are used as inputs of proposed optimization model to find the optimal weights. The results showed that the designed ensemble can outperform the benchmark

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
					been chosen as base learners for prediction	ensemble as well as all the individual base learners Based on the results for Boston housing dataset, XGBoost and Random Forests are the best algorithms predicting the median price of the houses with the least MSE and MAPE, and highest R-squared values. In other words, not only these two models predict with highest accuracy, they explain the variation in the target more than other chosen models. Moreover, prediction results of Ames housing dataset finds LASSO and Random Forests as the models with the least MSE and MAPE.
So, K., Orazem, P., Otto, D.	Public Use Microdata Samples (PUMS) of the 1990 United States Census.	8876	1990	commuting time, education level, gender, hourly wage, number of children, unearned income, housing price	Empirical modeling on joint decisions, Regression, Empirical estimates	The objective is to analyze the effects of housing prices, wages and commuting time on joint residential and job locations choices. This studyshows that an

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						intraregional empirical model of individual joint choice sof residential and job locations can yield plausible results. Nonmetropolitan residents trade off lower housing costs for lower wages in the local labor market. Those that opt to com mute to urban markets trade off higher wages for the disamenity of commuting time. All of these results are consistent with the underlying predictions of the Alonso–Mills–Muth model. That residential choices were influenced by differences in wages and housing prices is also consistent with previous interregional empirical studies based on the Roback model. Results suggest that improvements in transportation that lower commuting time will increase nonmetropolitan populations and will

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						increase the number of nonmetropolitan commuters to metropolitan markets. If, instead, policies encouraged eco nomic expansion in both markets which increased wages equally, population growth would be concentrated in metropolitan areas
Y., Jones, K.	the Land Registry of England and Wales	65,302	2001-2013	The actual sold price, The date of the sale, The address, The unit postcode, property type (detached houses, semi-detached houses, terraces or flats), duration (leasehold or freehold), and whether the property is newly-built at the time of the sale, Duration, Crime Rate, Location, House Price	Multilevel Modelling (MLM), Artificial Neural Networks (ANN)	The principal objective of this paper is to present Multilevel Modelling and Artificial Neural Networks approaches and compare their performance in terms of model fit, predictive accuracy and explanatory power. this paper explores a range of realistic scenarios as to how the effects of locations can be captured and compares the performance of all three approaches under each scenario. Neither ANN nor HPM is capable of including

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						neighborhood in the model due to the large number of categorical variables required for practical specification, while MLM is able to specify by simply defining them as macro-level units. All performance measures show that MLM is superior to ANN and HPM in each scenario, indicating that specification of neighborhood is helpful in house price predictions, even when the locations or neighborhood characteristics have been included in the model. The estimations of the coefficients of MLM are more robust and reliable than HPM, which reveals the effects of the measured variables at various spatial scales. The multilevel approach is capable of more fully exploiting the data on both properties and neighborhoods to achieve

Sources	Data Source/Country of origin	Sample Number	Timeline of Data	Variables	Techniques used	Objective and important findings
						better predictions, to understand what is driving the predictions.

Data Understanding

Data Collection

The data was published and pulled from the U.S. Census Bureau website and is the American Community Survey (ACS) Public Use Microdata Sample Files (PUMS) for 2017 in Oklahoma.

Data Description

The "Oklahoma Population Records" dataset is a .csv file composed of 74,753 records with 287 unique variables. These variables include a household identifier called "SERIALNO," and the other variables that we have classified into the following categories: (1) Citizenship, (2) Disabilities, (3) Education, (4) Employment, (5) Identification, (6) Income, (7) Insurance, (8) Language, (9) Military Service, (10) Nationality, (11) Relationships, (12) Transportation, and (13) Other. Records in this dataset are represented by the identifier "P" in the variable "RT," or Record Type. The "Oklahoma Housing Unit Records" dataset is also a .csv file, and it is composed of 39,081 records with 237 variables. The target variable is 'VALP' (property value). The median of property value in the sample is $120,000. It shares the household identifier "SERIALNO" with the "Oklahoma Population Records," and is composed of housing related variables such as number of bedrooms, house heating fuel, running water, and other living conditions, and the variables related to the earnings such as, household income, total earnings, wages, and, salaries.

Selection of Variables

We have selected 50 variables based on the literature review conducted for this project and analysis. Researchers have conducted analysis on physical attributes of housing and property values like lot size, insurance being paid, household income, electricity and gas bills being paid etc and how they are related in explaining the real estate prices (Huang, 2019; Lu et al., 2017; Tanjil et al., 2016; Fan et al., 2006; Lee, 1999, Hong, Jengei & Choi, Heeyoul "Henry & Kim,

Woo-sung. (2020)). We have also included our variables of interest from the individual details like education attainment of person, number of related children in household, language being spoken, family type and age of everyone in house. We have run two models to check if the education details of individual and number of related children in household are contributing in explaining the target-value property.

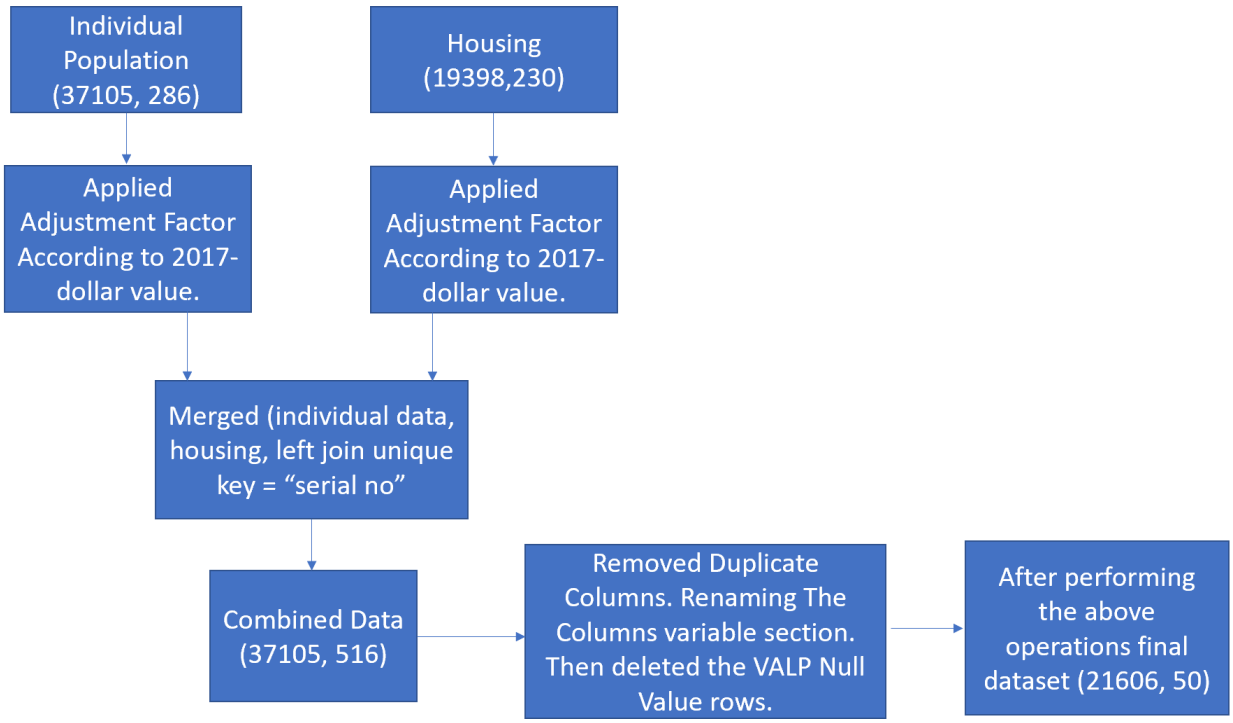
Data Preparation

Data Cleaning and Reshaping

The Oklahoma Population Records and Oklahoma Housing Unit Records datasets were cleaned separately and merged to obtain a meaningful dataset according to our objective. Out of 50 variables, 5 variables were refactored using Python and Excel to match the 2017-dollar value.

- **h_income:** We divided ADJINC by 1,000,000 to obtain the inflation adjustment factor and multiplied it to the PUMS variable value (HINCP).
- **ELEP:** We divided ADJHSG by 1,000,000 to obtain the inflation adjustment factor and multiplied it to the PUMS variable value (ELEP).
- **GASP:** We divided ADJHSG by 1,000,000 to obtain the inflation adjustment factor and multiplied it to the PUMS variable value (GASP).
- **INSP:** We divided ADJHSG by 1,000,000 to obtain the inflation adjustment factor and multiplied it to the PUMS variable value (INSP).
- **CONP:** We divided ADJHSG by 1,000,000 to obtain the inflation adjustment factor and multiplied it to the PUMS variable value (CONP).

2017 Dataset



The following table consists of the imported variables in SAS® Enterprise Miner™:

Table 1: Data Dictionary (including roles, measurement levels and reason for rejection)

Variable Name	Description	Measurement Level	Reason for Rejecting	Role
ACCESS	Access to the Internet 1 .Yes, by paying a cell phone company or Internet service provider 2.Yes, without paying a cell phone company or Internet service provider 3 .No access to the Internet at this house, apartment, or mobile home.	**Nominal**		**Input**
ACR	Lot size 1. House on less than one acre 2. House on one to less than ten	**Nominal**		**Input**

	acres 3. House on ten or more acres.			
ADJHSG	Adjustment factor for housing dollar amounts	**Interval**	An adjustment factor, used for calculations of other variables	**Rejected**
ADJINC	Adjustment factor for income and earnings dollar amounts	**Interval**	An adjustment factor, used for calculations of other variables	**Rejected**
AGEP	Age	**Interval**		**Input**
BATH	If the property has a bathtub or shower 1. Yes 2. No	**Binary**		**Input**
BDSP	Number of bedrooms	**Interval**		**Input**
BLD	Units in structure 2. One-family house detached 3. One-family house attached. Filtered and considered the above two levels.	**Nominal**		**Input**
BROADBND	Cellular data plan for a smartphone or other mobile device 0. No 1. Yes	**Interval**	Rejected as we have ACCES column	**Rejected**
CONP	Monthly condo fee	**Interval**		**Input**
COW	Class of worker 1 .Employee of a private for-profit company or business. 2. Employee of a private not-for-profit organization 3.Local government employee 4 .State government employee 5 .Federal government employee 6 .Self-employed in own not incorporated business, professional practice, or farm 7. Self-employed in own incorporated business,	**Nominal**		**Input**

	professional practice or farm8. Working without pay in family business or farm 9. Unemployed and last worked 5 years ago or earlier or never worked.			
ELEP	Monthly electricity cost	**Interval**		**Input**
ENG	Ability to speak English 1. Very well 2. Well 3. Not well 4. Not at all	**Nominal**	Rejected as we have LANX column	**Rejected**
FENGP	Ability to speak English 0. No 1.Yes	**Binary**	Rejected as we have LANX column	**Rejected**
FER	Gave birth to child within the past 12 months 0.No 1.Yes	**Binary**	Rejected as we have FPARC column	**Rejected**
FESRP	Employment status recode 0. No 1.Yes	**Binary**		**Input**
FKITP	Complete kitchen facilities 0. No 1.Yes	**Binary**		**Input**
FMARHYP	Year last married 0. No 1.Yes	**Binary**	Rejected as we have MAR column	**Rejected**
FMARP	Marital status 0. No 1.Yes	**Binary**	Rejected as we have MAR column	**Rejected**
FPARC	Family presence and age of related children 1.With related children under 5 years only 2. With related children 5 to 17 years only 3. With related children under 5 years and 5 to	**Nominal**		**Input**

	17 years 4. No related children.			
FSCHP	School enrollment 0. No 1.Yes	**Binary**		**Input**
FTAXP	Property taxes (yearly amount)	**Binary**		**Input**
GASP	Gas (monthly cost)	**Interval**		**Input**
h_income	House income	**Interval**		**Input**
HHL	Household language 1. English only 2. Spanish 3. Other Indo-European languages 4. Asian and Pacific Island languages 5. Other language	**Nominal**		**Input**
HHT	Household/family type 1. Married couple household 2. Other family household: Male householder, no spouse present 3. Other family household: Female householder, no spouse present 4. Nonfamily household: Male householder: Living alone 5. Nonfamily household: Male householder: Not living alone.	**Nominal**		**Input**
INSP	Fire/hazard/flood insurance (yearly amount).	**Interval**		**Input**
LANX	Language other than English spoken at home 1. Yes, speaks another language 2. No, speaks only English.	**Binary**		**Input**

MAR	Marital status 1 .Married 2 .Widowed 3. Divorced 4.Separated 5. Never married or under 15 years old	**Nominal**		**Input**
MV	When moved into the house or apartment 1.12 months or less 2. 13 to 23 months 3. 2 to 4 years 4.5 to 9 years 5. 10 to 19 years 6. 20 to 29 years 7.30 years or more.	**Nominal**		**Input**
NPF	Number of persons in family	**Interval**		**Input**
NRC	Number of related children in household	**Interval**		**Input**
PERNP	Total person's earnings	**Interval**	Rejected as we have h_income column	**Rejected**
PINCP	Total person's income	**Interval**	Rejected as we have h_income column	**Rejected**
PUMA	Public use microdata area code	**Nominal**	Rejected as we are only considering data from Oklahoma	**Rejected**
R18	Individuals under 18 years old in household. 0 .No person under 18 in household 1 .1 or more persons under 18 in household.	**Binary**		**Input**
R65	Individuals of 65 years and over in household. 0. No person 65 and over 1. 1 person 65 and over 2 .2 or more persons 65 and over.	**Nominal**		**Input**

RAC1P	Recoded detailed race code. 1 .White alone 2. Black or African American alone 3. American Indian alone 4. Alaska Native alone 5. American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races 6 .Asian alone 7. Native Hawaiian and Other Pacific Islander alone 8. Some Other Race alone 9. Two or More Races.	**Nominal**		**Input**
RMSP	Number of rooms	**Interval**		**Input**
RT	Record Type H .Housing Record or Group Quarters Unit P .Person Record	**Nominal**	Rejected due to left inner join since we have person's table on the left.	**Rejected**
RWAT	If the property has hot and cold running water 1. Yes 2. No 9. Case is from Puerto Rico, RWAT not applicable.	**Binary**		**Input**
SCH	School enrollment. 1 .No, has not attended in the last 3 months 2.Yes, public school or public college 3.Yes, private school or college or home school	**Nominal**		**Input**
SCHG	Grade level attending. 1 .Nursery school/preschool 2 .Kindergarten 3. Grade 1 4. Grade 2 5. Grade 3 6. Grade 4 7. Grade 5 8. Grade 6 9.Grade 7 10. Grade 8 11 .Grade 9 12. Grade 10 13. Grade 11	**Nominal**	Rejected as we have SCHL column	**Rejected**

	14. Grade 12 15. College undergraduate years (freshman to senior) 16. Graduate or professional school beyond a bachelor's degree.			
SCHL	Educational attainment 1. No schooling completed 2. Nursery school, preschool 3. Kindergarten 4. Grade 1 5. Grade 2 6. Grade 3 7. Grade 4 8. Grade 5 9. Grade 6 10 .Grade 7 11 .Grade 8 12 .Grade 9 13 .Grade 10 14 .Grade 11 15. 12th grade - no diploma 16. Regular high school diploma 17. GED or alternative credential 18. Some college, but less than 1 year 19. 1 or more years of college credit, no degree 20. Associate's degree 21. Bachelor's degree 22. Master's degree 23. Professional degree beyond a bachelor's degree 24. Doctorate degree.	**Nominal**		**Input**
SERIALNO	Person serial number of housing unit	**Nominal**	Rejected since SERIALNO does not impact the outcome of our dependent variable.	**Rejected**
SEX	Gender 1. Male 2. Female	**Binary**		**Input**

SRNT	Specified rental unit 0. A single-family home on 10 or more acres. 1. A single-family home on less than 10 acres or any other type of building, including mobile homes, with no regard to acreage.	**Binary**		**Input**
TYPE	Type of housing unit 1. Housing unit 2.Institutional group quarters 3. Noninstitutional group quarters	**Nominal**		**Input**
VALP	Property value	**Interval**		**Target**
YEAR	Year	**Nominal**	Rejected since YEAR does not impact the outcome of our dependent variable.	**Rejected**

The redundant and adjusting factor variables have been rejected and the final analysis had been conducted on the input and target variable which resulted in 50.

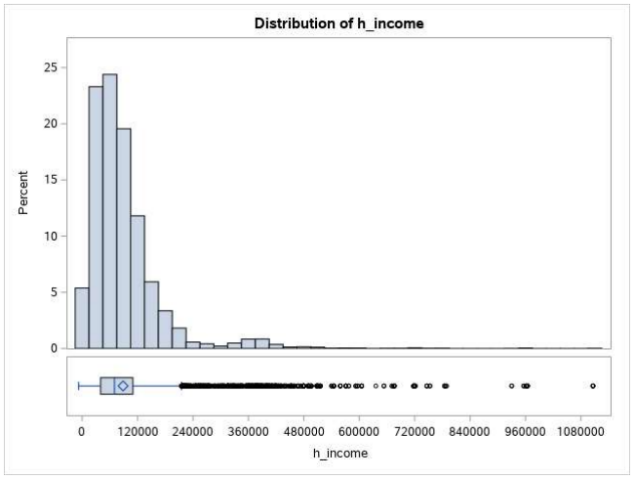
VARIABLE SUMMARY		
Role	**Measurement Level**	**Frequency Count**
INPUT	BINARY	10
INPUT	INTERVAL	10
INPUT	NOMINAL	14
REJECTED	BINARY	4
REJECTED	INTERVAL	5
REJECTED	NOMINAL	6
TARGET	INTERVAL	1

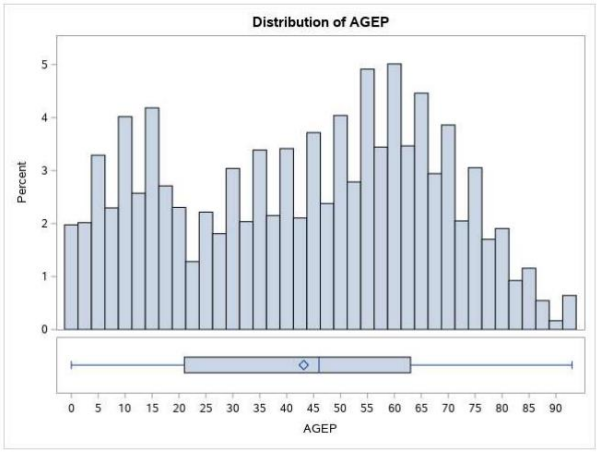
Exploratory Analysis
Summary Statistics

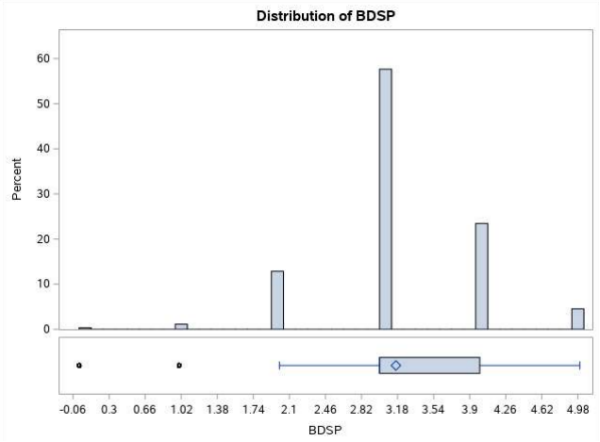
Interval Variables:

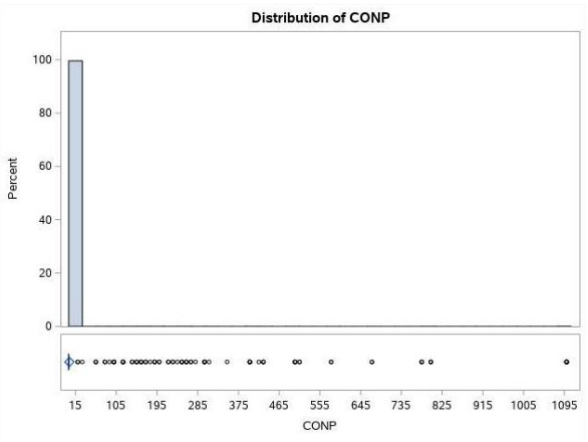
Variable	Mean	Std Dev	Median	N	Std Error	Variance	Range	Skewness	Kurtosis
h_income	88597.66	82087.69	70000.00	21606	558.4583979	6738388149	1113700.00	3.5121797	21.2136949
AGEP	43.1733778	24.0435312	46.0000000	21606	0.1635728	578.0913921	93.0000000	-0.1286449	-1.1027576
BDSP	3.1634731	0.7714716	3.0000000	21606	0.0052485	0.5951684	5.0000000	-0.0290852	1.1469904
CONP	1.6162177	32.3651935	0	21606	0.2201867	1047.51	1100.00	26.0919437	768.5827937
ELEP	163.1556049	89.7799719	150.0000000	21606	0.6107905	8060.44	559.0000000	1.4437611	3.1521417
GASP	55.0158752	57.4732030	50.0000000	21606	0.3910013	3303.17	399.0000000	2.5503143	10.8064691
NPF	3.4107548	1.4775297	3.0000000	18931	0.0107386	2.1830940	7.0000000	1.0019919	0.6966303
NRC	0.9853281	1.2982881	0	21606	0.0088325	1.6855519	7.0000000	1.2687159	1.1590809
RMSP	6.5940942	2.0279259	6.0000000	21606	0.0137964	4.1124835	15.0000000	1.3654674	3.8072823
VALP	178143.04	206014.12	135000.00	21606	1401.55	42441819069	1873000.00	5.6095852	41.4998264
INSP	1373.41	1162.49	1200.00	21606	7.9086500	1351384.95	7200.00	1.6451638	5.3829129

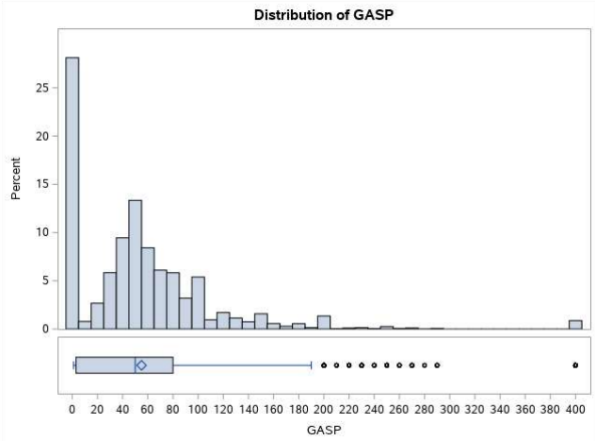
Distribution of the Interval Variables:

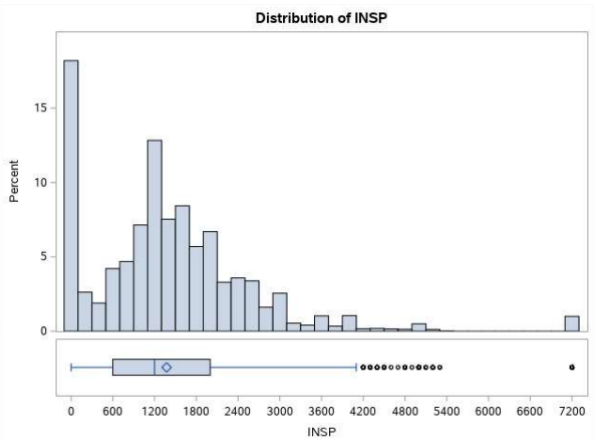


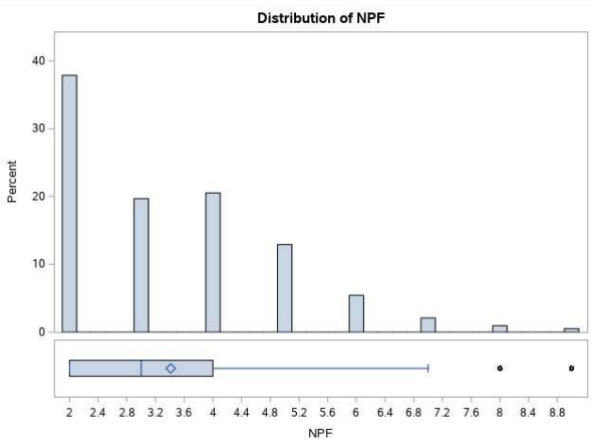


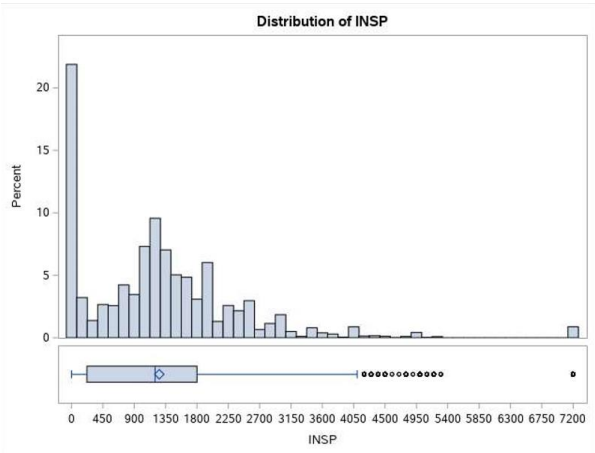


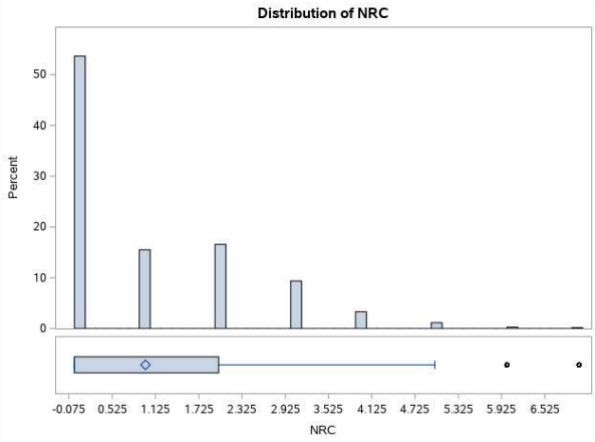


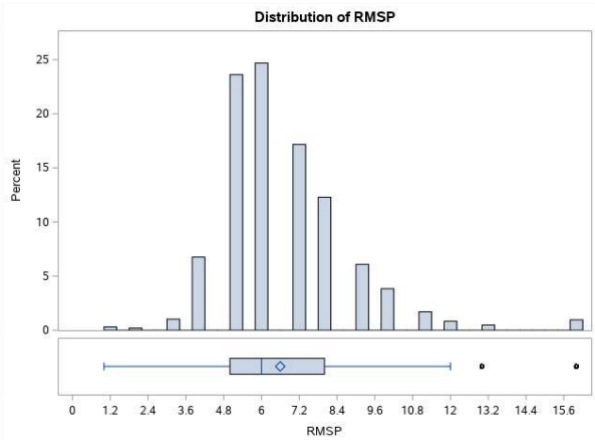

Distribution of GASP
Percent
0 5 10 15 20 25
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400
GASP

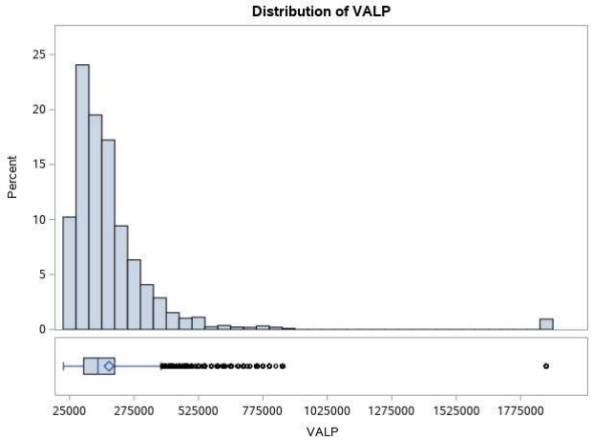

Distribution of INSP
Percent
0 5 10 15
0 600 1200 1800 2400 3000 3600 4200 4800 5400 6000 6600 7200
INSP


Distribution of NPF
Percent
0 10 20 30 40
2 2.4 2.8 3.2 3.6 4 4.4 4.8 5.2 5.6 6 6.4 6.8 7.2 7.6 8 8.4 8.8
NPF


Distribution of INSP
Percent
0 5 10 15 20
0 450 900 1350 1800 2250 2700 3150 3600 4050 4500 4950 5400 5850 6300 6750 7200
INSP


Distribution of NRC
Percent
0 10 20 30 40 50
-0.075 0.525 1.125 1.725 2.325 2.925 3.525 4.125 4.725 5.325 5.925 6.525
NRC


Distribution of RMSP
Percent
0 5 10 15 20 25
0 1.2 2.4 3.6 4.8 6 7.2 8.4 9.6 10.8 12 13.2 14.4 15.6
RMSP



Handling the Outliers:

As seen from the above summary statistics, distribution plots, and Box and Whisker plots we see the skewness is high for H_INCOME, CONP and VALP variables. This can be handled by applying log transfer over the values.

Categorical Variables:

Chi-Square and Frequency Distribution:

Chi-Square Goodness of Fit Test/ Chi-Square Test of Homogeneity/ chi-square test for equal proportions: This explains how to conduct a chi-square goodness of fit test. The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

ACCESS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	18453	85.41	18453	85.41
2	490	2.27	18943	87.67
3	2663	12.33	21606	100.00

ACR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	13956	64.59	13956	64.59
2	4981	23.05	18937	87.65
3	2669	12.35	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	26692.3724
DF	2
Pr > ChiSq	<.0001

Chi-Square Test for Equal Proportions	
Chi-Square	9871.9031
DF	2
Pr > ChiSq	<.0001

BATH	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21592	99.94	21592	99.94
2	14	0.06	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	21550.0363
DF	1
Pr > ChiSq	<.0001

BDSP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	76	0.35	76	0.35
1	246	1.14	322	1.49
2	2778	12.86	3100	14.35
3	12456	57.65	15556	72.00
4	5070	23.47	20626	95.46
5	980	4.54	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	31046.2666
DF	5
Pr > ChiSq	<.0001

BLD	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	21376	98.94	21376	98.94
3	230	1.06	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	20695.7936
DF	1
Pr > ChiSq	<.0001

COW	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7237	59.16	7237	59.16
2	781	6.38	8018	65.54
3	923	7.55	8941	73.09
4	984	8.04	9925	81.13
5	585	4.78	10510	85.92
6	1027	8.40	11537	94.31
7	572	4.68	12109	98.99
8	64	0.52	12173	99.51
9	60	0.49	12233	100.00
Frequency Missing = 9373				

Chi-Square Test for Equal Proportions	
Chi-Square	29361.4855
DF	8
Pr > ChiSq	<.0001

FESRP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	20244	93.70	20244	93.70
1	1362	6.30	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	16501.4313
DF	1
Pr > ChiSq	<.0001

FKITP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21379	98.95	21379	98.95
1	227	1.05	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	20707.5398
DF	1
Pr > ChiSq	<.0001

FSCHP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	20331	94.10	20331	94.10
1	1275	5.90	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	16806.9581
DF	1
Pr > ChiSq	<.0001

FPARC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1318	6.96	1318	6.96
2	6420	33.91	7738	40.87
3	2276	12.02	10014	52.90
4	8917	47.10	18931	100.00
Frequency Missing = 2675				

Chi-Square Test for Equal Proportions	
Chi-Square	8039.9131
DF	3
Pr > ChiSq	<.0001

FTAXP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	17161	79.43	17161	79.43
1	4445	20.57	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	7483.8774
DF	1
Pr > ChiSq	<.0001

HHL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	19350	89.56	19350	89.56
2	1416	6.55	20766	96.11
3	210	0.97	20976	97.08
4	395	1.83	21371	98.91
5	235	1.09	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	65564.9030
DF	4
Pr > ChiSq	<.0001

LANX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1491	7.26	1491	7.26
2	19041	92.74	20532	100.00
Frequency Missing = 1074				

Chi-Square Test for Equal Proportions	
Chi-Square	15001.0959
DF	1
Pr > ChiSq	<.0001

HHT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15557	72.00	15557	72.00
2	1075	4.98	16632	76.98
3	2299	10.64	18931	87.62
4	887	4.11	19818	91.72
5	340	1.57	20158	93.30
6	1206	5.58	21364	98.88
7	242	1.12	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	59674.0383
DF	6
Pr > ChiSq	<.0001

MAR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	11040	51.10	11040	51.10
2	1299	6.01	12339	57.11
3	1928	8.92	14267	66.03
4	207	0.96	14474	66.99
5	7132	33.01	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	19631.2623
DF	4
Pr > ChiSq	<.0001

MV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1420	6.57	1420	6.57
2	1119	5.18	2539	11.75
3	3524	16.31	6063	28.06
4	4057	18.78	10120	46.84
5	5794	26.82	15914	73.66
6	2780	12.87	18694	86.52
7	2912	13.48	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	4936.3846
DF	6
Pr > ChiSq	<.0001

R18	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11466	53.07	11466	53.07
1	10140	46.93	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	81.3791
DF	1
Pr > ChiSq	<.0001

R65	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	15135	70.05	15135	70.05
1	3458	16.00	18593	86.05
2	3013	13.95	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	13121.0422
DF	2
Pr > ChiSq	<.0001

RAC1P	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	16487	76.31	16487	76.31
2	741	3.43	17228	79.74
3	1880	8.70	19108	88.44
4	2	0.01	19110	88.45
5	65	0.30	19175	88.75
6	388	1.80	19563	90.54
7	5	0.02	19568	90.57
8	351	1.62	19919	92.19
9	1687	7.81	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	94623.6382
DF	8
Pr > ChiSq	<.0001

RWAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21571	99.84	21571	99.84
2	35	0.16	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	21466.2268
DF	1
Pr > ChiSq	<.0001

SCH	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	16245	77.48	16245	77.48
2	4159	19.84	20404	97.31
3	563	2.69	20967	100.00
Frequency Missing = 639				

Chi-Square Test for Equal Proportions	
Chi-Square	19312.6215
DF	2
Pr > ChiSq	<.0001

SCHL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	527	2,51	527	2,51
2	324	1,55	851	4,06
3	280	1,34	1131	5,39
4	235	1,12	1366	6,51
5	289	1,38	1655	7,89
6	314	1,50	1969	9,39
7	293	1,40	2262	10,79
8	299	1,43	2561	12,21
9	369	1,76	2930	13,97
10	328	1,56	3258	15,54
11	528	2,52	3786	18,06
12	458	2,18	4244	20,24
13	559	2,67	4803	22,91
14	636	3,03	5439	25,94
15	302	1,44	5741	27,38
16	4614	22,01	10355	49,39
17	698	3,33	11053	52,72
18	1289	6,15	12342	58,86
19	2698	12,87	15040	71,73
20	1355	6,46	16395	78,19
21	2982	14,22	19377	92,42
22	1149	5,48	20526	97,90
23	282	1,34	20808	99,24
24	159	0,76	20967	100,00
Frequency Missing = 639				

Chi-Square Test for Equal Proportions	
Chi-Square	30868,1135
DF	23
Pr > ChiSq	<.0001

SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	10556	48.86	10556	48.86
2	11050	51.14	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	11.2948
DF	1
Pr > ChiSq	0.0008

SRNT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21606	100.00	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	0.0000
DF	0
Pr > ChiSq	.

TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21606	100.00	21606	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	0.0000
DF	0
Pr > ChiSq	.

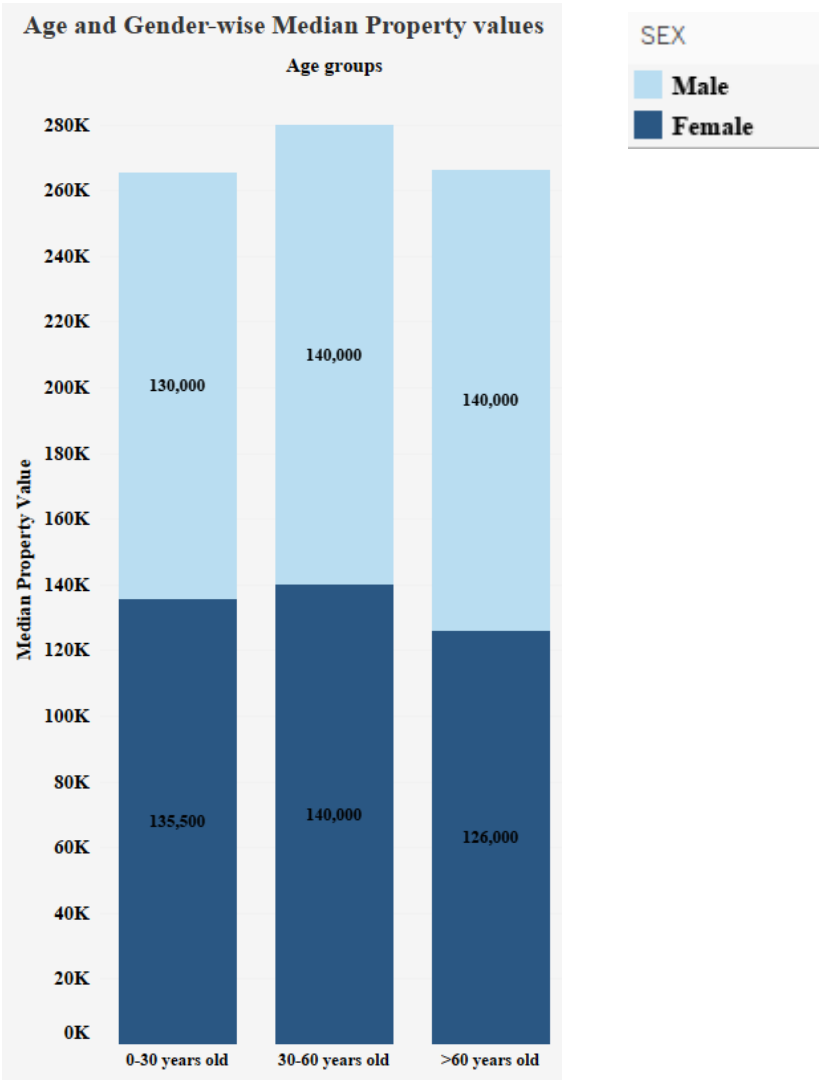
Sample Size = 21606

Pearson Correlation Coefficients

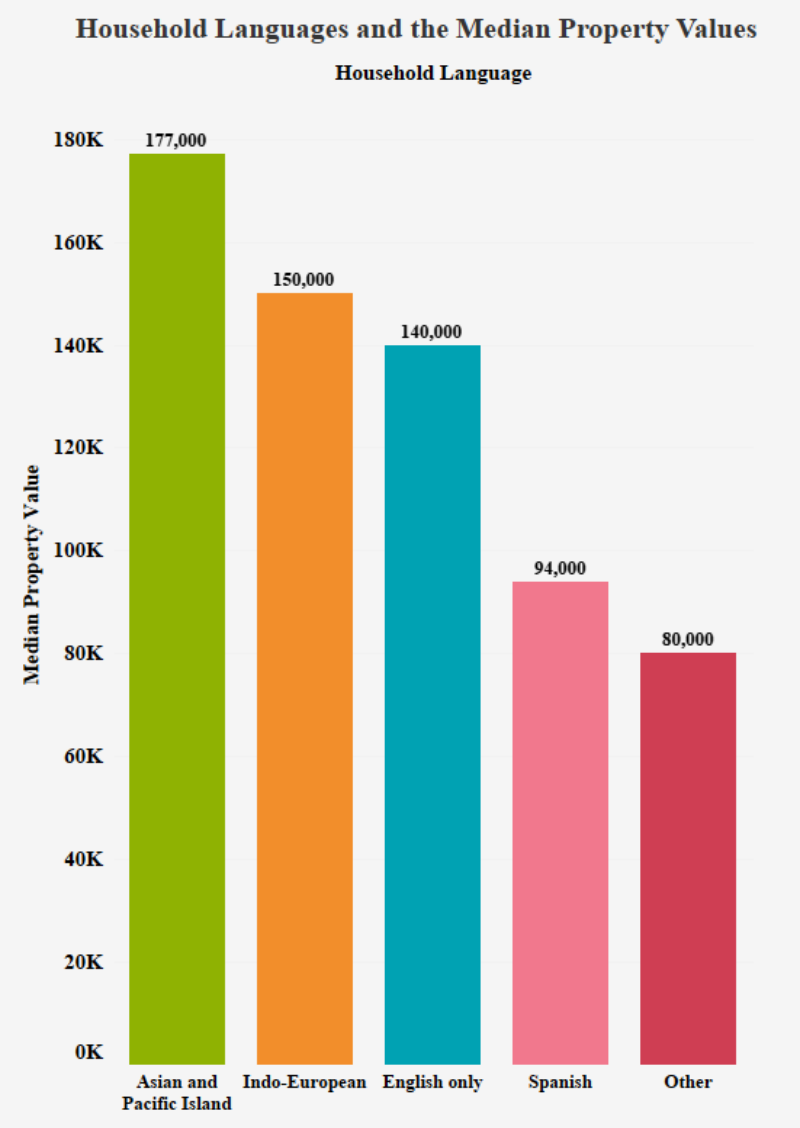
Pearson Correlation Coefficients Number of Observations										
	h_income	AGEP	CONP	ELEP	GASP	INSP	VALP	NPF	NRC	RMSP
h_income	1.00000 21606	-0.09837 21606	0.02495 21606	0.21807 21606	0.06314 21606	0.38800 21606	0.39868 21606	0.08658 18931	0.07180 21606	0.31272 21606
AGEP	-0.09837 21606	1.00000 21606	0.02184 21606	-0.12090 21606	-0.03226 21606	-0.01134 21606	-0.00337 21606	-0.54243 18931	-0.62142 21606	-0.03559 21606
CONP	0.02495 21606	0.02184 21606	1.00000 21606	0.00224 21606	-0.00111 21606	-0.01075 21606	-0.00288 21606	-0.01235 18931	-0.01622 21606	0.02461 21606
ELEP	0.21807 21606	-0.12090 21606	0.00224 21606	1.00000 21606	-0.03151 21606	0.20511 21606	0.24769 21606	0.16944 18931	0.16440 21606	0.22423 21606
GASP	0.06314 21606	-0.03226 21606	-0.00111 21606	-0.03151 21606	1.00000 21606	0.06087 21606	0.07747 21606	0.07093 18931	0.06182 21606	0.12377 21606
INSP	0.38800 21606	-0.01134 21606	-0.01075 21606	0.20511 21606	0.06087 21606	1.00000 21606	0.53348 21606	-0.00765 18931	0.00841 21606	0.34266 21606
VALP	0.39868 21606	-0.00337 21606	-0.00288 21606	0.24769 21606	0.07747 21606	0.53348 21606	1.00000 21606	0.00947 18931	0.02850 21606	0.34625 21606
NPF	0.08658 18931	-0.54243 18931	-0.01235 18931	0.16944 18931	0.07093 18931	-0.00765 18931	0.00947 18931	1.00000 18931	0.85299 18931	0.08877 18931
NRC	0.07180 21606	-0.62142 21606	-0.01622 21606	0.16440 21606	0.06182 21606	0.00841 21606	0.02850 21606	0.85299 18931	1.00000 21606	0.08696 21606
RMSP	0.31272 21606	-0.03559 21606	0.02461 21606	0.22423 21606	0.12377 21606	0.34266 21606	0.34625 21606	0.08877 18931	0.08696 21606	1.00000 21606

Data Visualization

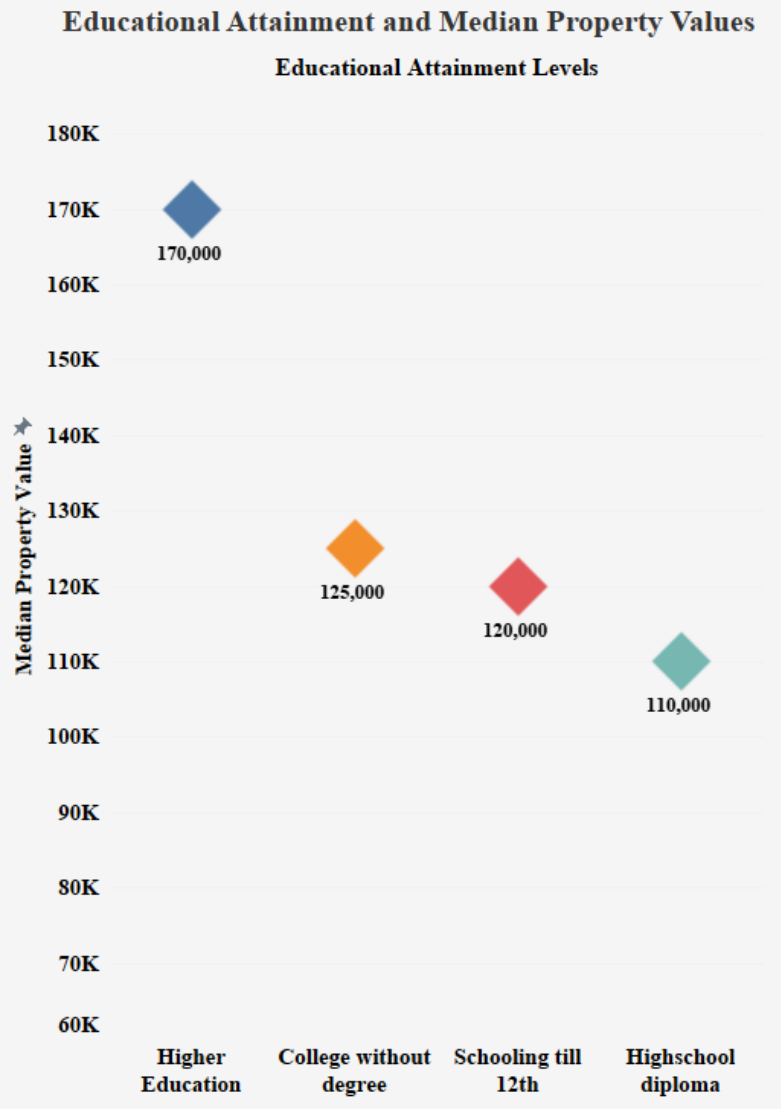
Property value are affected by many factors such as age, sex, household language, education, and marital status which are being compared to property value.



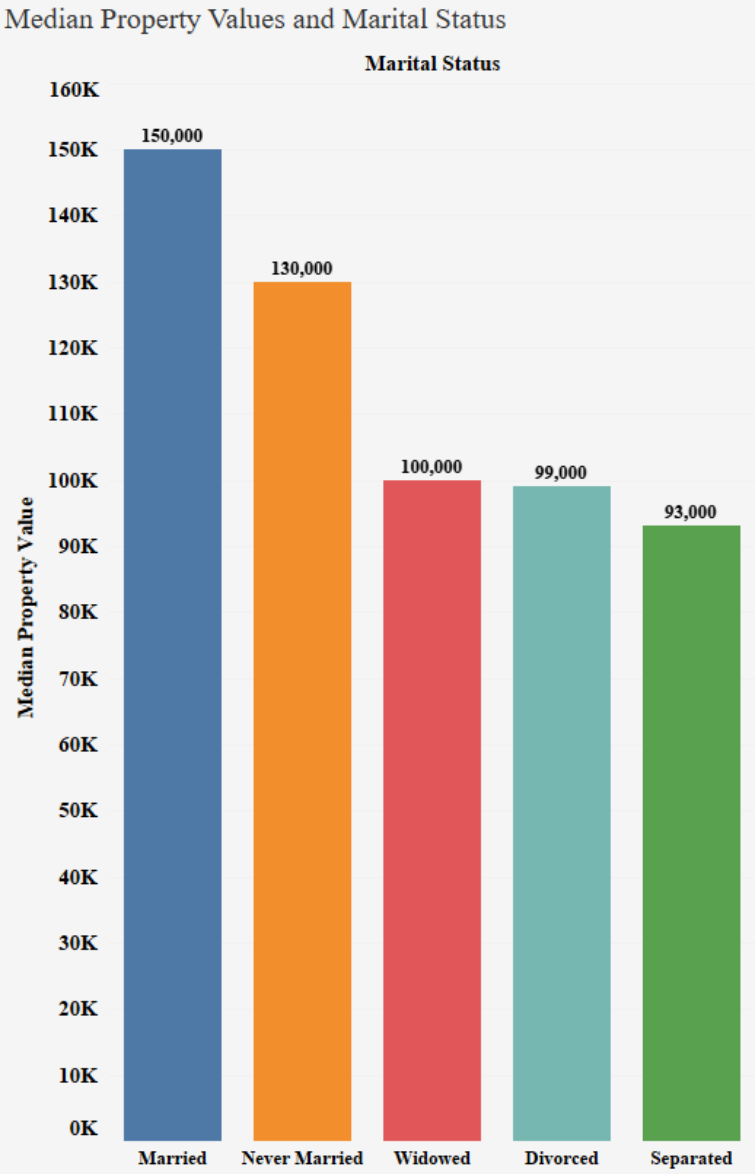
There is not a significant difference in property values when comparing sex. Age did have an impact on property value. Individuals who are 30 to 60 years old have property values that are nearly $20,000 more valuable than other age groups.



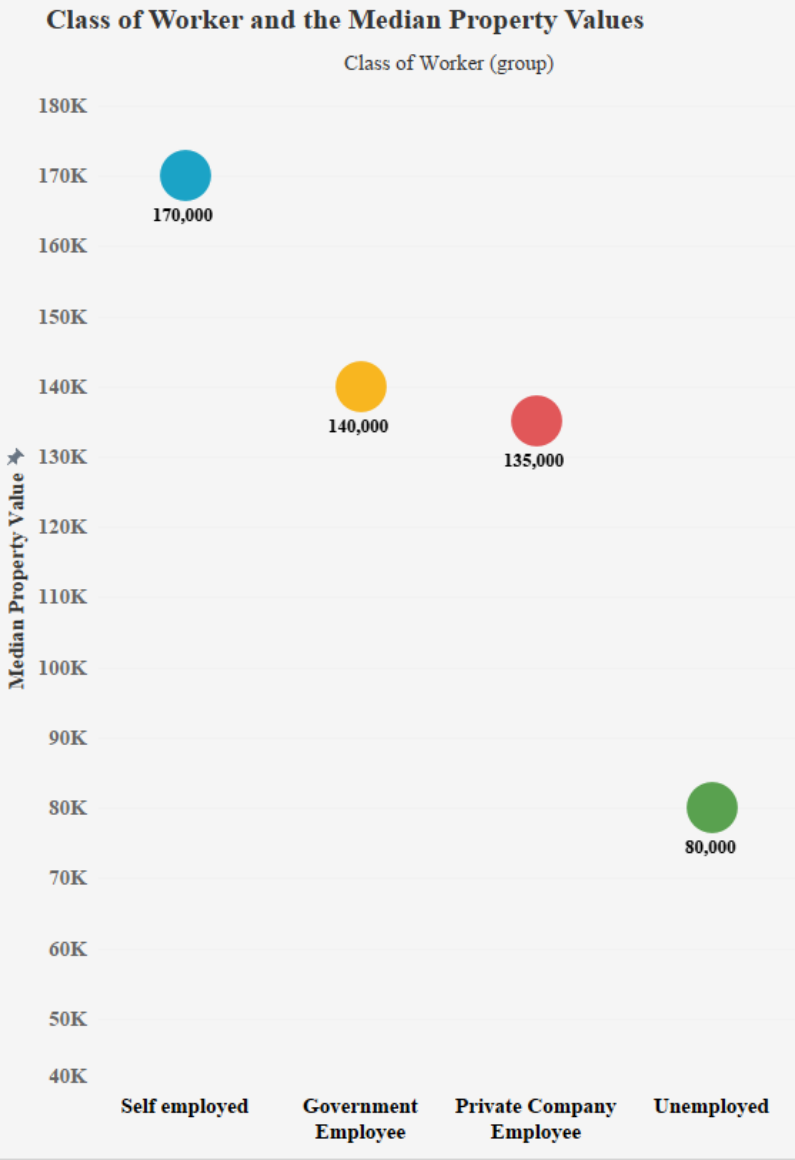
When comparing property value and household language, households that speak an Asian and Pacific Island language have on average property values worth 15% more than the next highest group.



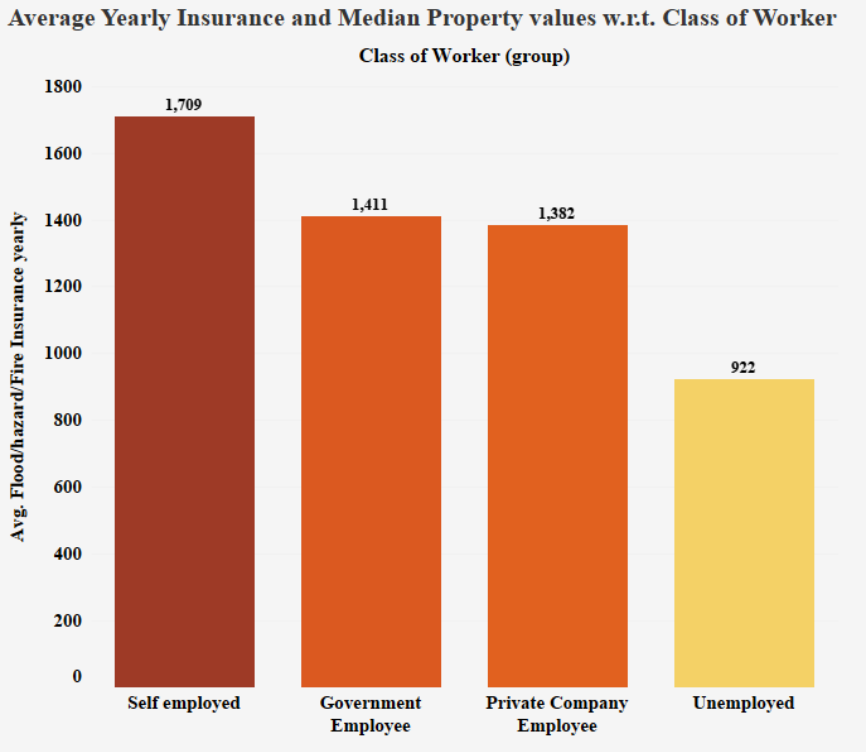
Educational attainment has a major impact on property values. This also translates to household income. Both variables are highly correlated. This graph clearly displays the value that higher education brings to individuals and families.

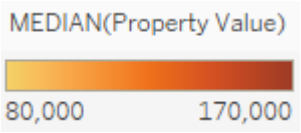


Married couples have, on average, property values that are $20,000 more valuable than those who never marry, and more than $50,000 more valuable than individuals who have divorced or separated.



Individuals who are self-employed have the highest property values.





Modeling and Evaluation

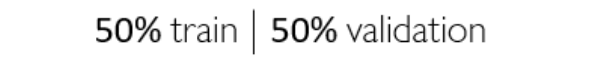
Data Partition

In predictive modeling, the standard strategy for honest assessment of model performance is data splitting (Christie et al., 2017). Our dataset was split into 50% training, and 50% validation.

Table 5: Variable Summary

VARIABLE SUMMARY		
Role	**Measurement Level**	**Frequency Count**
INPUT	BINARY	10
INPUT	INTERVAL	10
INPUT	NOMINAL	14
REJECTED	BINARY	4
REJECTED	INTERVAL	5
REJECTED	NOMINAL	6
TARGET	INTERVAL	1

Table 6: Partition summary



Type	Data Set	Number of Observations
DATA	EMWS1.Ids6_DATA	21606
TRAIN	EMWS1.Part_TRAIN	10803
VALIDATE	EMWS1.Part_VALIDATE	10803

Transformation, Replacement, and Imputation

We transformed the inputs that are highly skewed, highly kurtotic and had high standard deviation. Table 7 shows the list of variables that were transformed using a logarithmic transformation to reduce statistical deviations.

Table 7: List of Transformed Variables

Variables	Standard Deviation	Skewness	Kurtosis
CONP	40.0358345	22.3975841	550.0737267
ELEP	90.21963551	1.39363121	2.913676947
GASP	57.10760942	2.566472149	11.03591403
INSP	1176.48097	1.68841097	5.521816642
RMSP	2.035660098	1.409654787	3.913353052
VALP	218952.6301	5.518355914	38.41318206
h_income	82543.05604	3.502114851	20.79987757
LOG_CONP	0.407588109	14.65331217	216.8830334
LOG_ELEP	0.644948814	-2.18405531	11.9304033
LOG_GASP	1.342148333	-0.541069135	-1.126161099
LOG_INSP	2.796905457	-1.519360957	0.59149497
LOG_RMSP	0.254239504	0.02785427	2.314343818
LOG_VALP	0.860062571	-0.573922635	3.129335629
LOG_h_income	0.812639529	-0.737094017	4.641698526

The replacement node is used to reassign non missing values to any of the levels before performing imputation (Christie et al., 2017). Table 8 illustrates the inputs that have been grouped according to its similar categories and frequency distribution.

Table 8: Replacement

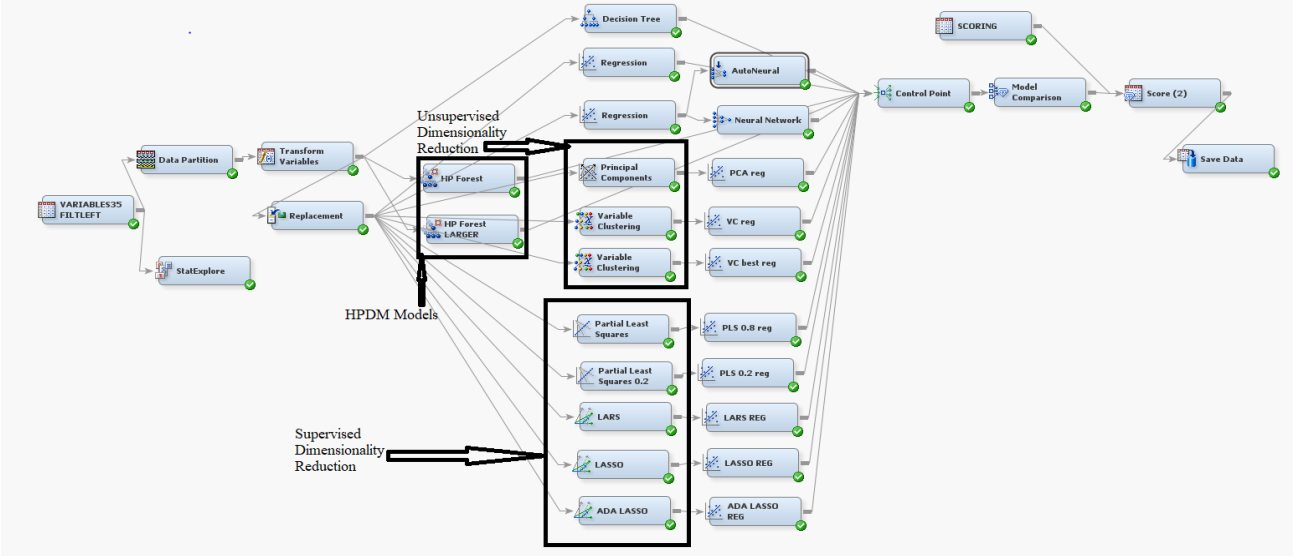
Replacement Counts

Obs	Variable	Role	Label	Train	Validation
1	ACCESS	INPUT		0	0
2	BLD	INPUT		113	117
3	COW	INPUT		7185	7184
4	SCH	INPUT		562	640
5	SCHL	INPUT		10384	10371

The impute node is used to replace the missing values in the dataset (Christie et al., 2017). We did not use the impute node as the ACS data had a higher quality in terms of completeness of responses as it is achieved by telephone and personal interviewing of households that do not respond by mail. Additionally, according to the data dictionary, a null response in most categorical variables is not a lack of response but rather it meant that the variable did not apply to that individual or was an exception. Imputing missing values for these class level variables would incorrectly change the interpretation of all observations and could drastically change the outcomes of the analyses.

Modeling

Several predictive models were built using SAS® Enterprise Miner™ to predict the property value in Oklahoma. The following are the models that we used:



High Performance Data Mining Nodes (HPDM)

The definition given by the textbook to explain this model was:

> Enterprise Miner HPDM nodes use SAS High-Performance Analytics procedures. SAS High-Performance Analytics procedures are written using threaded kernel extensions. They are designed to run in a distributed environment. HPDM are multi-threaded, to take advantage of multiple cores in single-machine mode. HPDM use algorithm choices that consider data movement and replication.

Random Forest

Random Forest is an ensemble model designed to use predictions from the classification of regression trees. The model excels in its ability to overcome instability seen in a single classification or regression tree. It randomly samples different variables amongst the data and it compares the results with data excluded from the original sampling. The terms used to refer to this type of technique is a bagged sample and an out-of-bag sample. Within our results we

provided the average squared error for the out-of-bag sample which did better than all the other models.

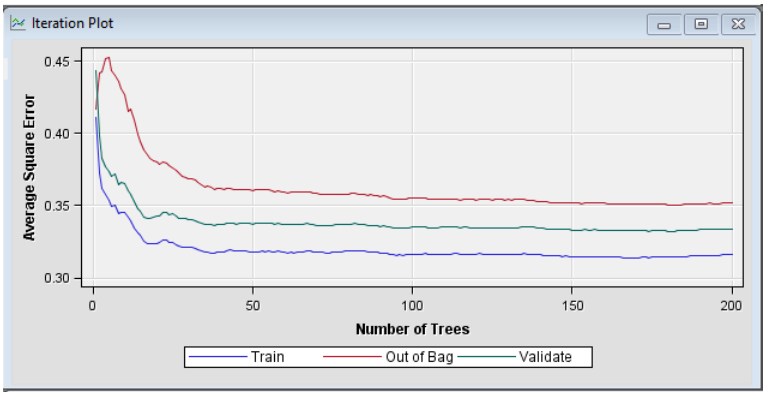
A further motivation to use the HPDM Random Forest node was its capacity to provide better diversity and predictive accuracy. These two attributes were important for our team to achieve our objective that we sought out. Assessments that deal with property valuation in the world among us, our team believed it is important to provide results that provides consistency.

Property Selection in Random Forest

The model had three options which it exemplified in its capability to perform. Those three options are the number of trees, the number of inputs and the sampling strategy. We were able to adjust each option under the properties selection by changing the maximum number of trees property, the number of vars, and the proportion of observations in each sample and the number of observations in each sample. We chose to adjust the default value of the proportion of observations to 0.6 and 0.8 with the maximum number of trees to 100 and 200, respectively. The best results were obtained with the Random Forest Model with 0.8 proportion and 200 trees.

Random Forest Results:

Iteration Plot



Fit Statistics

Statistics Label	Train	Validation
Average Squared Error	0.315852	0.333656
Divisor for ASE	10803	10803
Maximum Absolute Error	4.659808	4.68956
Sum of Frequencies	10803	10803
Root Average Squared Error	0.562007	0.57763
Sum of Squared Errors	3412.148	3604.485

Loss Reduction Variable Importance

Variable	Number of Rules	MSE	OOB MSE	Valid MSE	Absolute Error	OOB Absolute Error	Valid Absolute Error
LOG_INSP	4349	0.145390	0.13986	0.13757	0.075700	0.071882	0.073946
LOG_h_income	3488	0.059268	0.05180	0.05316	0.033102	0.029422	0.030664
BDSP	1986	0.029061	0.02719	0.03149	0.017377	0.016182	0.019919
LOG_RMSP	2786	0.030043	0.02610	0.02810	0.018725	0.016210	0.017535
ACR	1656	0.015388	0.01305	0.01441	0.009681	0.008264	0.008836
LOG_ELEP	1940	0.010380	0.00750	0.00587	0.005259	0.003495	0.002838
HHT	709	0.007903	0.00606	0.00616	0.004837	0.003809	0.003863
ACCESS	610	0.005445	0.00480	0.00511	0.003357	0.002995	0.003167
SCHL	465	0.008165	0.00468	0.00519	0.004303	0.002417	0.003070
MV	1574	0.004138	0.00165	0.00179	0.003178	0.001422	0.001493
NPF	1720	0.002984	0.00117	0.00084	0.002200	0.000941	0.000750
RAC1P	834	0.002552	0.00105	0.00124	0.001932	0.001006	0.001062
LOG_GASP	1633	0.003092	0.00086	0.00038	0.002004	0.000645	0.000436
HHL	699	0.001529	0.00083	0.00105	0.001336	0.000863	0.001035
FPARC	907	0.001940	0.00078	0.00056	0.001388	0.000541	0.000410
NRC	1593	0.001750	0.00077	0.00070	0.001508	0.000698	0.000718
FTAXP	595	0.001411	0.00056	0.00070	0.000670	0.000098796	0.000169
MAR	254	0.001086	0.00055	0.00068	0.000679	0.000348	0.000400
R65	606	0.001021	0.00020	0.00017	0.000768	0.000171	0.000137
LOG_CONP	97	0.000239	0.00019	0.00003	0.000143	0.000087006	0.000024703
BLD	98	0.000225	0.00014	0.00010	0.000149	0.000093735	0.000053078
R18	494	0.000521	0.00009	0.00005	0.000494	0.000182	0.000142
FKITP	65	0.000088	0.00004	-0.00001	0.000073381	0.000028732	0.000003087
LANX	239	0.000360	0.00004	0.00007	0.000305	0.000073405	0.000094387
SCH	233	0.000410	0.00002	0.00000	0.000264	-0.000027565	-0.000029486
RWAT	18	0.000092	0.00002	0.00001	0.000031035	0.000004431	0.000001936
FSCHP	185	0.000167	0.00001	-0.00003	0.000149	0.000019356	0.000005946
SRNT	0	0.000000	0.00000	0.00000	0	0	0
BATH	0	0.000000	0.00000	0.00000	0	0	0
TYPE	0	0.000000	0.00000	0.00000	0	0	0
COW	211	0.000731	-0.00001	0.00014	0.000437	-0.000005423	0.000050656
FESRP	230	0.000276	-0.00009	-0.00005	0.000193	-0.000049246	-0.000018313
SEX	153	0.000131	-0.00010	-0.00016	0.000109	-0.000113	-0.000116
AGEP	516	0.001063	-0.00028	-0.00033	0.000658	-0.000150	-0.000156

Variable Clustering

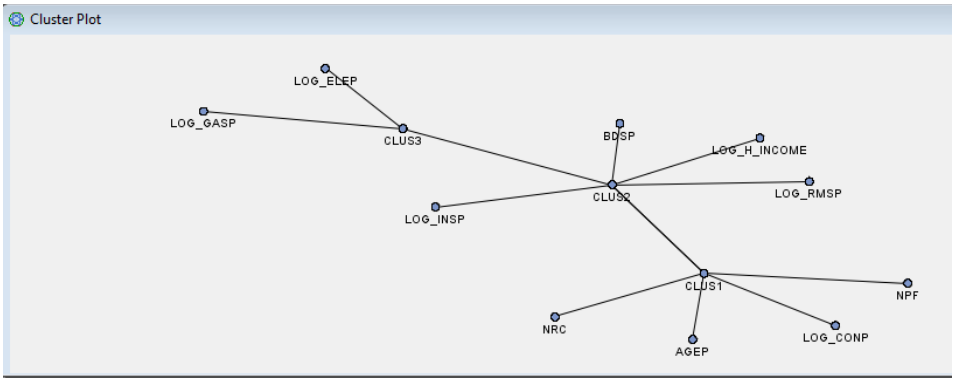
The Variable Clustering tool is useful for data reduction, such as choosing the best variables or cluster components for analysis. Variable clustering removes collinearity, decreases variable redundancy, and helps reveal the underlying structure of the input variables in a data set (SAS Institute Inc., 2017).

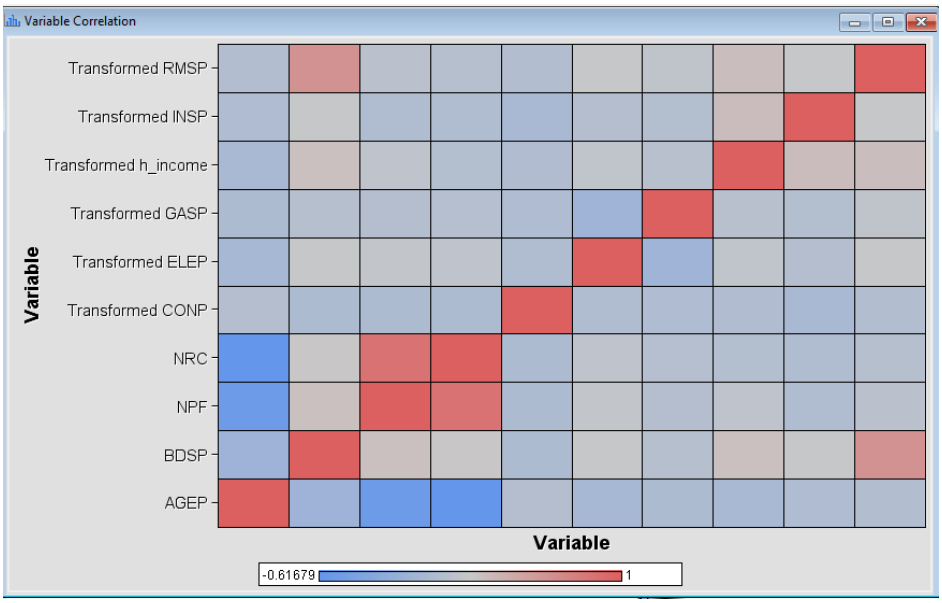
The individual and census data sets used in the project included a high number of variables. To quickly and efficiently review these variables, variable clustering was performed to find the "best variables". In addition, variable clustering was useful since many variables in the dataset had high correlation.

Cluster Analysis

Cluster analysis is a data reduction method that attempts to group training dataset cases based on similarities in input variables (Christie et al., 2017). We used the cluster node to segment our dataset and 3 clusters were selected. The important variables based on the cluster analysis output are as follows:

Cluster analysis Results:





3 Clusters		R-squared with			
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	AGEP	0.6356	0.0061	0.3666	
	LOG_CONP	0.0029	0.0003	0.9975	Transformed CONP
	NPF	0.8338	0.0315	0.1716	
	NRC	0.8812	0.0126	0.1203	
Cluster 2	BDSP	0.6405	0.0538	0.3799	
	LOG_INSP	0.2451	0.0001	0.7549	Transformed INSP
	LOG_RMSP	0.6547	0.0020	0.3460	Transformed RMSP
	LOG_h_income	0.3797	0.0055	0.6237	Transformed h_income
Cluster 3	LOG_ELEP	0.5685	0.0369	0.4480	Transformed ELEP
	LOG_GASP	0.5685	0.0093	0.4355	Transformed GASP

No cluster meets the criterion for splitting.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	2.588962	0.2589	0.2589	1.812318	0.0027	
2	4.354384	0.4354	0.3335	1.133917	0.0029	0.9974
3	5.410555	0.5411	0.4800	0.998457	0.0029	0.9975

Variable Clustering Regression Results:

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	LOG_RMSP	1	1	1304.75	<.0001
2	ACR	2	2	194.57	<.0001
3	REP_SCHL	3	3	85.82	<.0001
4	HHT	2	4	96.78	<.0001
5	MV	6	5	30.29	<.0001
6	REP_ACCESS	2	6	58.55	<.0001
7	RAC1P	7	7	11.67	<.0001
8	REP_SCH	1	8	38.16	<.0001
9	REP_COW	3	9	13.17	<.0001
10	R65	2	10	14.10	<.0001
11	FPARC	3	11	6.09	0.0004
12	HHL	4	12	4.55	0.0011
13	LOG_GASP	1	13	7.68	0.0056

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	37	1608.928859	43.484564	80.28	<.0001
Error	8876	4807.708486	0.541653		
Corrected Total	8913	6416.637345			

Model Fit Statistics

R-Square	0.2507	Adj R-Sq	0.2476
AIC	-5427.5260	BIC	-5425.1915
SBC	-5157.9016	C(p)	36.9397

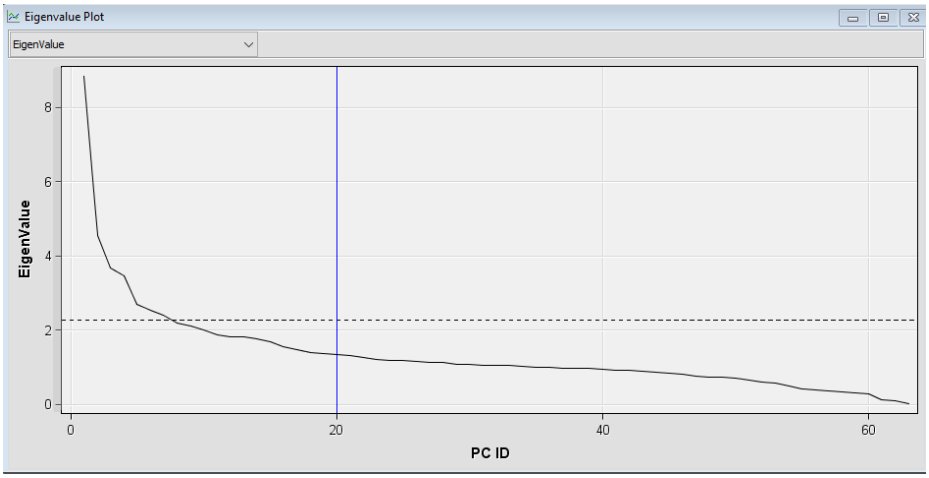
Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
ACR	2	232.0781	214.23	<.0001
FPARC	3	9.4330	5.81	0.0006
HHL	4	10.1380	4.68	0.0009
HHT	2	63.5123	58.63	<.0001
LOG_GASP	1	4.1594	7.68	0.0056
LOG_RMSP	1	498.2619	919.89	<.0001
MV	6	87.7365	27.00	<.0001
R65	2	11.9091	10.99	<.0001
RAC1P	7	20.7544	5.47	<.0001
REP_ACCESS	2	54.2429	50.07	<.0001
REP_COW	3	22.0809	13.59	<.0001
REP_SCH	1	26.7687	49.42	<.0001
REP_SCHL	3	107.9610	66.44	<.0001

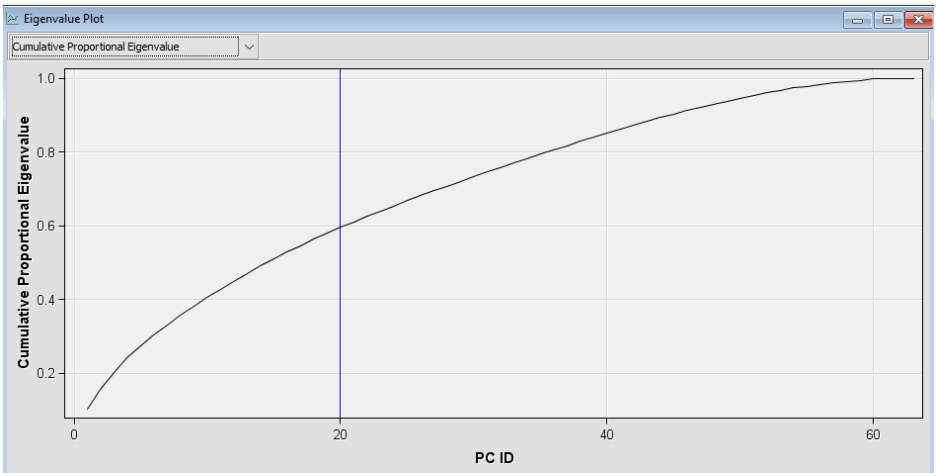
Statistics Label	Train	Validation
Akaike's Information Criterion	-5667.01	NaN
Average Squared Error	0.587656	0.565801
Average Error Function	0.587656	0.565801
Degrees of Freedom for Error	10765	NaN
Model Degrees of Freedom	38	NaN
Total Degrees of Freedom	10803	NaN
Divisor for ASE	10803	10803
Error Function	6348.452	6112.349
Final Prediction Error	0.591805	NaN
Maximum Absolute Error	5.602125	5.437957
Mean Square Error	0.589731	0.565801
Sum of Frequencies	10803	10803
Number of Estimate Weights	38	NaN
Root Average Sum of Squares	0.766587	0.752198
Root Final Prediction Error	0.769289	NaN
Root Mean Squared Error	0.767939	0.752198
Schwarz's Bayesian Criterion	-5390.09	NaN
Sum of Squared Errors	6348.452	6112.349
Sum of Case Weights Times Freq	10803	10803

Principal Component Analysis

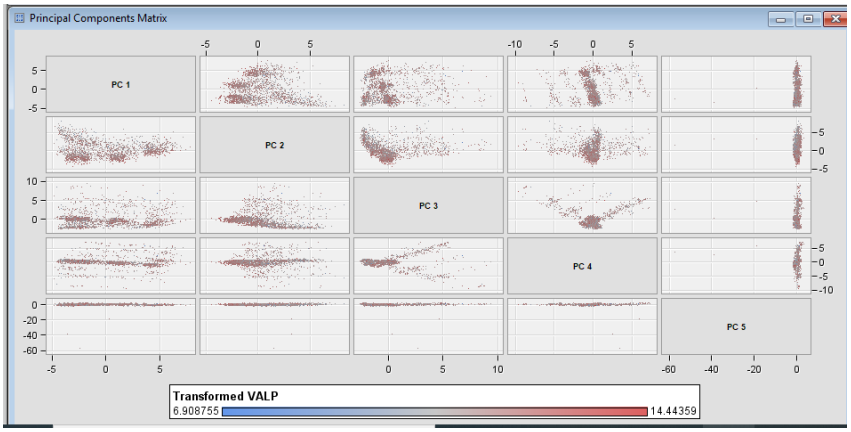
As discussed earlier, many variables in the project had high collinearity. A feature of principal component analysis is that when run, it constructs output variables that are uncorrelated (SAS Institute Inc., 2017). The results of the PCA node are described below.



The eigenvalues are plotted on the y-axis and the number of the principal components on the x-axis. According to the scree plot, we used ten principal components because it was the most drastic change in slope which occurred for PC ID 10.



The vertical blue line at principal component 20 means that 20 principal components were selected to pass on as input variables to successor nodes.



The scatter plots for the first five principal components are shown by default

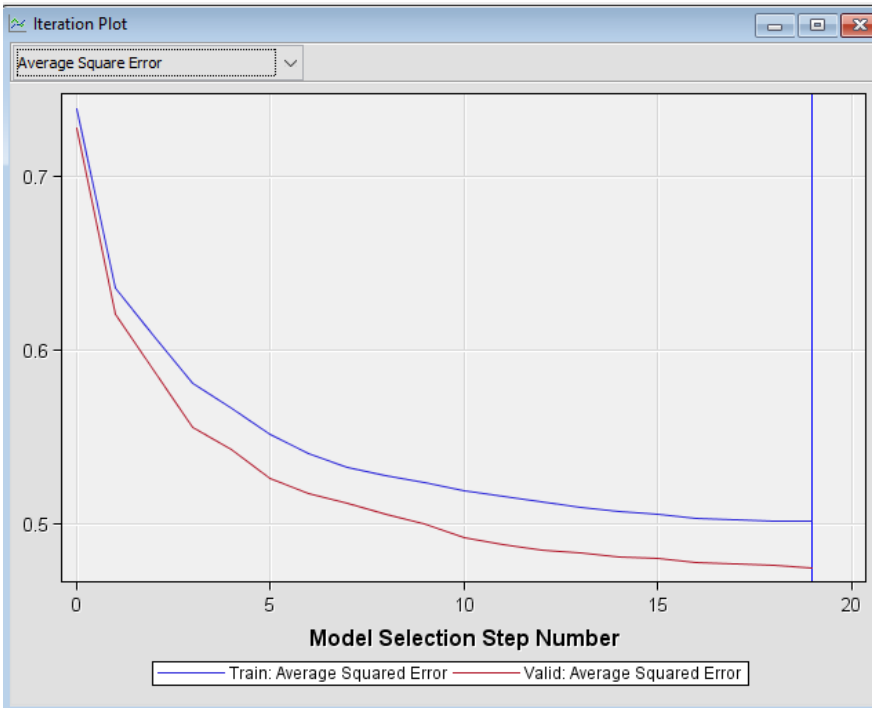
```
*-------------------------------------------------------------*
Summary of Exported Principal Components
*-------------------------------------------------------------*

Total number of input variables: 34
Maximum number cutoff of principal components: 20
Cumulative proportional eigenvalue cutoff: 0.99
Proportional eigenvalue increment cutoff: 0.001
Number of the selected principal components: 20
Total variation explained by the selected principal components: 0.5950812214
```

```
Variable Summary

             Measurement    Frequency
Role            Level         Count

INPUT         BINARY            10
INPUT         INTERVAL          10
INPUT         NOMINAL           14
REJECTED      NOMINAL            5
```

The number of input variables is 34 but we can see that there are only 20 input variables. Fourteen of them are nominal which are automatically coded using dummy variables. We also see that ten variables are binary and are also dummy-coded.

Passing the Principal Components to Successor Regression Node:

Statistics Label	Train	Validation
Akaike's Information Criterion	-7418.04	NaN
Average Squared Error	0.501392	0.475137
Average Error Function	0.501392	0.475137
Degrees of Freedom for Error	10783	NaN
Model Degrees of Freedom	20	NaN
Total Degrees of Freedom	10803	NaN
Divisor for ASE	10803	10803
Error Function	5416.533	5132.903
Final Prediction Error	0.503252	NaN
Maximum Absolute Error	5.671551	5.525705
Mean Square Error	0.502322	0.475137
Sum of Frequencies	10803	10803
Number of Estimate Weights	20	NaN
Root Average Sum of Squares	0.70809	0.689302
Root Final Prediction Error	0.709402	NaN
Root Mean Squared Error	0.708746	0.689302
Schwarz's Bayesian Criterion	-7272.29	NaN
Sum of Squared Errors	5416.533	5132.903
Sum of Case Weights Times Freq	10803	10803



Model selection was considered at the 19th step.

Below are the results of the 19th step:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	2573.788375	135.462546	269.67	<.0001
Error	10783	5416.533395	0.502322		
Corrected Total	10802	7990.321769			

Model Fit Statistics

R-Square	0.3221	Adj R-Sq	0.3209
AIC	-7418.0440	BIC	-7415.9668
SBC	-7272.2924	C(p)	19.1866

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.7478	0.00682	1722.82	<.0001
PC_1	1	0.0166	0.00229	7.26	<.0001
PC_10	1	0.1166	0.00481	24.24	<.0001
PC_11	1	0.1216	0.00500	24.31	<.0001
PC_12	1	0.0783	0.00504	15.53	<.0001
PC_13	1	-0.0897	0.00506	-17.71	<.0001
PC_14	1	-0.0417	0.00513	-8.12	<.0001
PC_15	1	0.0681	0.00524	13.00	<.0001
PC_16	1	0.0476	0.00547	8.69	<.0001
PC_17	1	0.0264	0.00562	4.69	<.0001
PC_18	1	-0.0255	0.00577	-4.42	<.0001
PC_19	1	0.0561	0.00581	9.65	<.0001
PC_2	1	-0.1509	0.00319	-47.24	<.0001
PC_20	1	0.0167	0.00588	2.84	0.0046
PC_4	1	-0.0359	0.00367	-9.80	<.0001
PC_5	1	0.0335	0.00415	8.08	<.0001
PC_6	1	0.0752	0.00429	17.54	<.0001
PC_7	1	0.0287	0.00440	6.53	<.0001
PC_8	1	0.0286	0.00460	6.22	<.0001
PC_9	1	0.0476	0.00470	10.13	<.0001

Decision Tree

One of the first models we implemented was a decision tree. It is easy to understand and uses non-linear relationships between input and target variables. The split search algorithm it executes are based on its logworth value and we can run through its branches to see it develop a classification system. Within our own data discovery, we found important variables that we continually found in other models. The following decision tree models were run by modifying the selection criteria:

Maximal Tree

The maximal tree was built using the interactive decision tree. The train node of the tree was accessed through the root node of the tree. The first splitting variable was INSP – annual insurance cost. After pruning, we obtained a subtree assessment plot with an average squared error criteria and obtained 46 leaves.

Figure 11: Leaf Statistics – Selection Criteria as largest and average square error

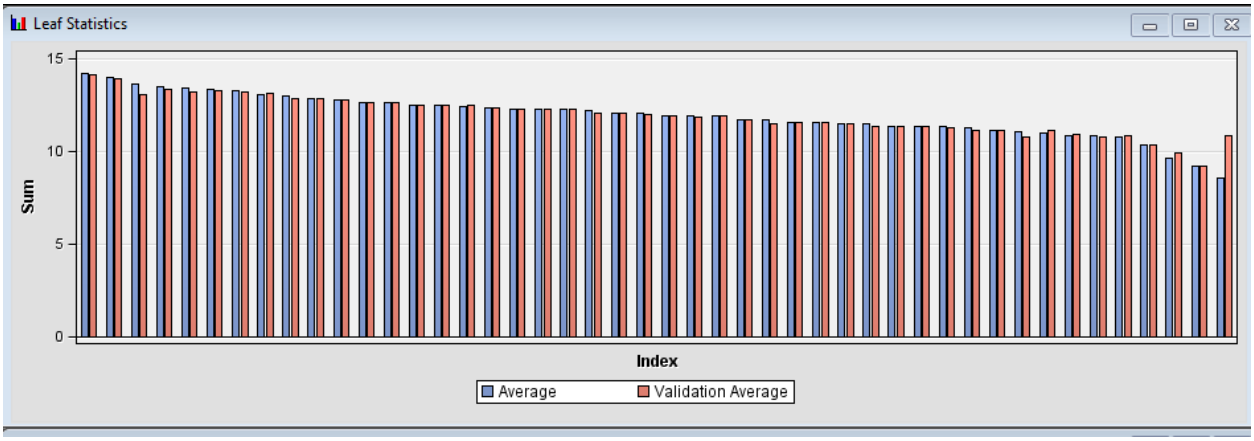
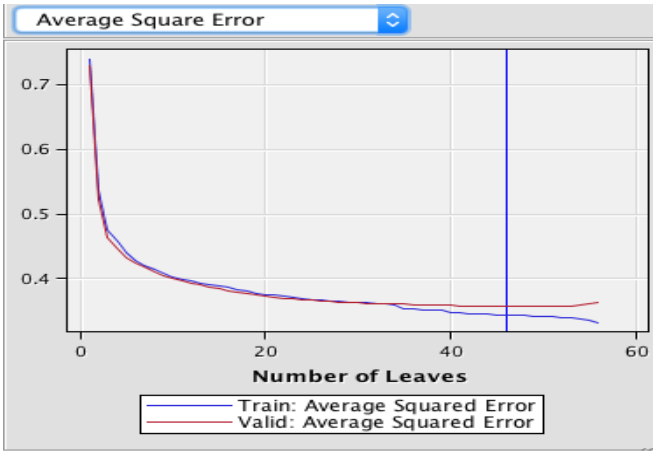


Figure 12: Subtree assessment plot – Average square error

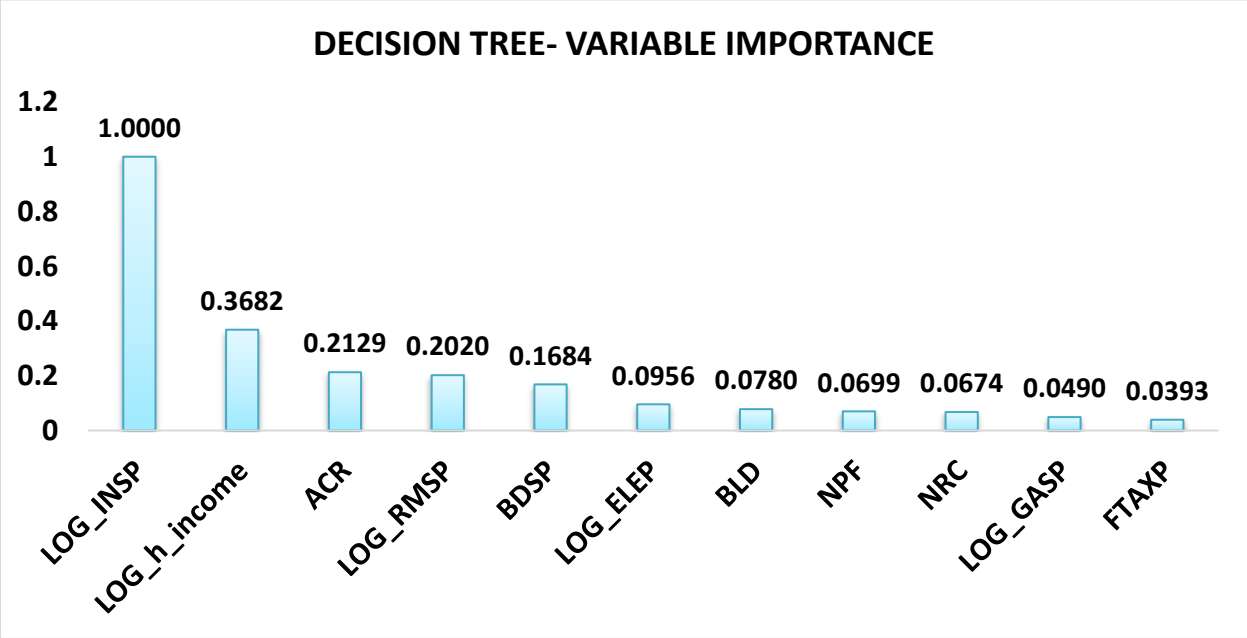


Subtree assessment plot – Average square error

Fit Statistics:

Statistics Label	Train	Validation
Sum of Frequencies	10803	10803
Maximum Absolute Error	4.663567	4.663567
Sum of Squared Errors	3717.529	3859.905
Average Squared Error	0.34412	0.357299
Root Average Squared Error	0.586617	0.597745
Divisor for ASE	10803	10803
Total Degrees of Freedom	10803	NaN

Variable importance – Decision Tree



The most important variable in this model is “LOG_INSP” (insurance), followed by “ “LOG_h_income” (house income) and ACR (Lot Size).

Regression Models – Stepwise Selection Method

Regressions assume a specific association structure between inputs and target (Christie et al., 2017). Different variable selection methods like stepwise, forward, and backward selection were used in regression models to find the optimal set of inputs. Out of three methods, stepwise regression was the best model for our dataset with an average of squared error of 0.465412.

Fit statistics - Stepwise Selection

Statistics Label	Train	Validation
Akaike's Information Criterion	-7969.09	NaN
Average Squared Error	0.474609	0.465412
Average Error Function	0.474609	0.465412
Degrees of Freedom for Error	10762	NaN
Model Degrees of Freedom	41	NaN
Total Degrees of Freedom	10803	NaN
Divisor for ASE	10803	10803
Error Function	5127.198	5027.849
Final Prediction Error	0.478225	NaN
Maximum Absolute Error	5.514543	5.372435
Mean Square Error	0.476417	0.465412
Sum of Frequencies	10803	10803
Number of Estimate Weights	41	NaN
Root Average Sum of Squares	0.688918	0.682211
Root Final Prediction Error	0.691538	NaN
Root Mean Squared Error	0.69023	0.682211
Schwarz's Bayesian Criterion	-7670.3	NaN
Sum of Squared Errors	5127.198	5027.849
Sum of Case Weights Times Freq	10803	10803

Summary of Stepwise Selection

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	LOG_INSP	1	1	3146.20	<.0001
2	LOG_RMSP	1	2	988.99	<.0001
3	LOG_h_income	1	3	631.47	<.0001
4	ACR	2	4	234.96	<.0001
5	BDSP	1	5	92.51	<.0001
6	REP_SCHL	3	6	29.11	<.0001
7	MV	6	7	13.81	<.0001
8	R65	2	8	32.59	<.0001
9	NPF	1	9	39.18	<.0001
10	FTAXP	1	10	31.61	<.0001
11	LOG_ELEP	1	11	30.19	<.0001
12	RAC1P	7	12	6.32	<.0001
13	REP_COW	3	13	9.18	<.0001
14	REP_SCH	1	14	19.16	<.0001
15	AGEP	1	15	19.56	<.0001
16	LOG_GASP	1	16	17.56	<.0001
17	NRC	1	17	16.28	<.0001
18	HHT	2	18	11.72	<.0001
19	SEX	1	19	4.68	0.0306
20	LANX	1	20	4.44	0.0351
21	REP_ACCESS	2	21	3.14	0.0432

Parameters likelihood table

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	7.2463	0.1496	48.43	<.0001
ACR	1	1	-0.2209	0.0103	-21.37	<.0001
ACR	2	1	0.00342	0.0114	0.30	0.7645
AGEP		1	0.00254	0.000626	4.06	<.0001
BDSP		1	0.1367	0.0118	11.61	<.0001
FTAXP	0	1	0.0369	0.00844	4.38	<.0001
HHT	1	1	0.0839	0.0225	3.72	0.0002
HHT	2	1	-0.0283	0.0348	-0.81	0.4174
HHT	3	0	0	.	.	.
HHT	4	0	0	.	.	.
HHT	5	0	0	.	.	.
HHT	6	0	0	.	.	.
LANX	1	1	0.0314	0.0151	2.07	0.0380
LOG_ELEP		1	0.0691	0.0110	6.27	<.0001
LOG_GASP		1	0.0211	0.00537	3.93	<.0001
LOG_INSP		1	0.1167	0.00266	43.92	<.0001
LOG_RMSP		1	0.4192	0.0351	11.95	<.0001
LOG_h_income		1	0.2174	0.0102	21.28	<.0001
MV	1	1	0.0680	0.0239	2.85	0.0044
MV	2	1	0.1227	0.0259	4.74	<.0001
MV	3	1	0.0737	0.0164	4.48	<.0001
MV	4	1	0.0482	0.0155	3.11	0.0019
MV	5	1	-0.0330	0.0133	-2.48	0.0132
MV	6	1	-0.1154	0.0179	-6.46	<.0001
NPF		1	-0.0718	0.00963	-7.46	<.0001
NRC		1	0.0532	0.0111	4.80	<.0001
R65	0	1	-0.0451	0.0136	-3.32	0.0009
R65	1	1	-0.0172	0.0147	-1.17	0.2401
RAC1P	1	1	-0.0753	0.0819	-0.92	0.3580
RAC1P	2	1	-0.0430	0.0881	-0.49	0.6253
RAC1P	3	1	-0.1243	0.0838	-1.48	0.1378
RAC1P	5	1	-0.0616	0.1307	-0.47	0.6374
RAC1P	6	1	0.1004	0.0947	1.06	0.2891
RAC1P	7	1	0.5794	0.5557	1.04	0.2972
RAC1P	8	1	-0.3076	0.0947	-3.25	0.0012
REP_ACCESS	1	1	0.0376	0.0175	2.15	0.0313
REP_ACCESS	2	1	-0.0208	0.0301	-0.69	0.4893
REP_COW	1	1	-0.0379	0.0394	-0.96	0.3356
REP_COW	2	1	0.0925	0.0430	2.15	0.0315
REP_COW	3	1	0.0640	0.0841	0.76	0.4470
REP_SCH	1	1	-0.0771	0.0138	-5.60	<.0001
REP_SCHL	1	1	0.1371	0.0481	2.85	0.0044
REP_SCHL	2	1	-0.0854	0.0217	-3.93	<.0001
REP_SCHL	3	1	-0.0758	0.0203	-3.74	0.0002
SEX	1	1	-0.0148	0.00680	-2.17	0.0302

Neural Network

Neural network is a natural extension of a regression model. Its prediction formula has a flexible addition that permits the trained neural network to model virtually any association between input and target variable (Christie et al., 2017). Neural networks and auto neural networks were attached to each of these nodes in an attempt to improve the predictive power of these models. After conducting various regressions, variable selection methods, partial least squares, consolidation and selection trees using the stepwise method, the variables included in the model were then passed through to a neural network in order to discover any additional association between the inputs and the target.

Neural Network with Stepwise Regression

The validation average squared error is 0.415689 and its optimal count was obtained at the 50th iteration.

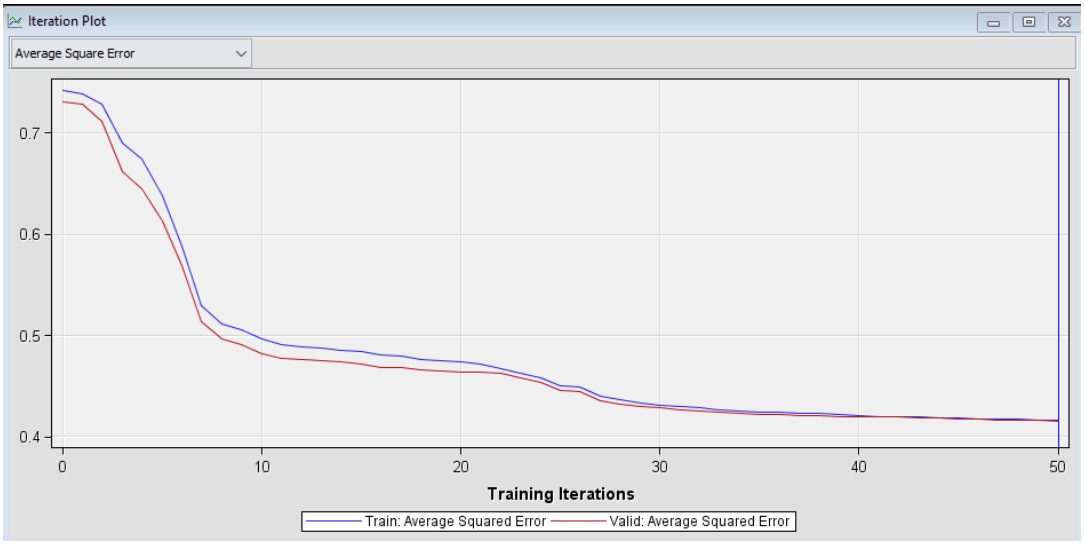


Table 31: Fit statistics – Neural Network Stepwise Regression

Statistics Label	Train	Validation
Average Squared Error	0.416456	0.415689
Maximum Absolute Error	4.880008	4.880008
Divisor for ASE	10803	10803
Sum of Frequencies	10803	10803
Root Average Squared Error	0.645334	0.64474
Sum of Squared Errors	4498.977	4490.69
Sum of Case Weights Times Freq	10803	10803
Final Prediction Error	0.427313	NaN
Mean Squared Error	0.421885	0.415689
Root Final Prediction Error	0.653692	NaN
Root Mean Squared Error	0.649526	0.64474
Average Error Function	0.416456	0.415689
Error Function	4498.977	4490.69

AutoNeural with Stepwise Regression

The auto neural tool presents an automatic way to explore alternative network architectures and hidden unit counts (Christie et al., 2017). The auto neural network is similar to the neural network except that the existing network weights were not reinitialized. In this case, the average square error is decreasing with a unit increase in our number of hidden inputs.

Table 32: Fit statistics - AutoNeural with Stepwise Regression

Statistics Label	Train	Validation
Average Squared Error	0.420518	0.417712
Maximum Absolute Error	4.880008	4.880008
Divisor for ASE	10803	10803
Sum of Frequencies	10803	10803
Root Average Squared Error	0.648474	0.646306
Sum of Squared Errors	4542.858	4512.54
Sum of Case Weights Times Freq	10803	10803
Final Prediction Error	0.435172	NaN
Mean Squared Error	0.427845	0.417712
Root Final Prediction Error	0.659676	NaN
Root Mean Squared Error	0.654099	0.646306
Average Error Function	0.420518	0.417712
Error Function	4542.858	4512.54

Iteration plot with 1 hidden unit

Supervised Interval variable selection

Partial Least Square Regression

As in principal component analysis (PCA), the partial least squares (PLS) algorithm extracts components as a linear combination of the original input variables. It constructs components explaining as much of the variation as possible of both the target and input variables.

Table 37: Fit statistics - Partial Least Square Regression

Statistics Label	Train	Validation
Average Squared Error	0.40374727	0.39469367
Divisor for ASE	10803	10803
Maximum Absolute Error	5.44082715	5.33086053
Sum of Frequencies	10803	10803
Root Average Squared Error	0.63541111	0.6282465
Sum of Squared Errors	4361.68179	4263.87567

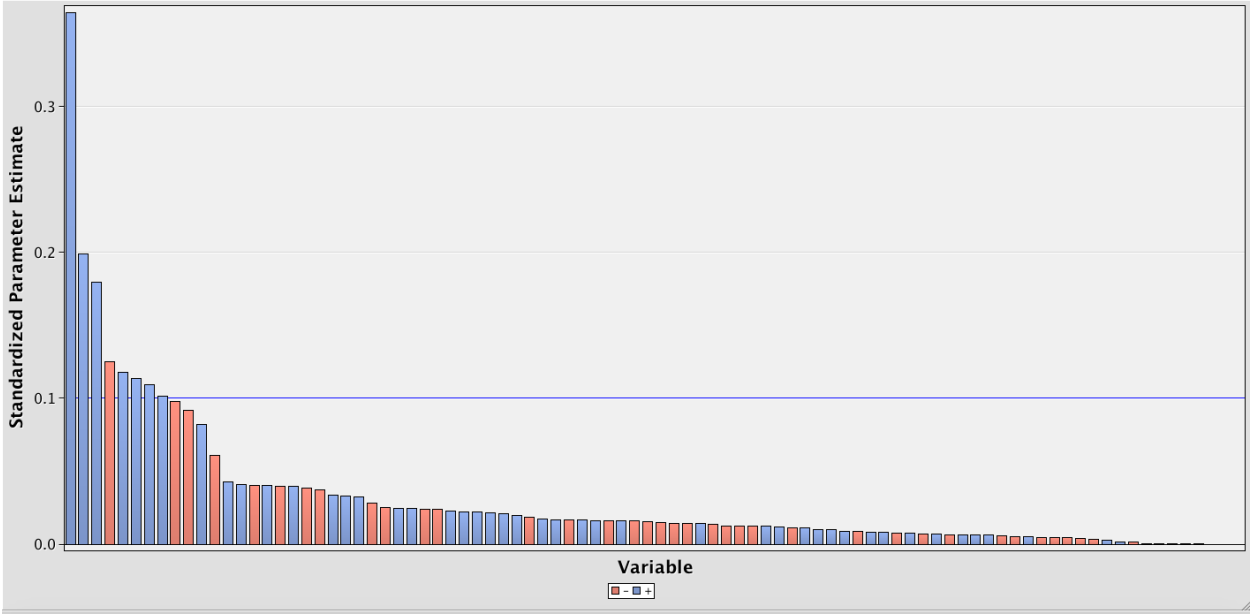


Figure 30: Iteration plot – Absolute Standardized Parmeter Estimates

In this plot, the absolute standardized parameter estimates for all variables are shown. The horizontal blue line shows the current value of the cutoff.

LAR (Least Angle Regression)

The least angle regression method, like forward selection, starts with no effects in the model and adds effects. The parameter estimates at any step are "shrunk" when compared to the corresponding least squares estimates. When the algorithm describes a step being taken, this is simply a new term added to the model. When the algorithm describes the length of the step, this is referring to the size of the parameter coefficient.

Table 38: Fit statistics – Least Angle Regression

Statistics Label	Train	Validation
Akaike's Information Criterion	-8115.8337	NaN
Average Squared Error	0.4668207	0.46014763
Average Error Function	0.4668207	0.46014763
Divisor for ASE	10803	10803
Error Function	5043.06406	4970.97484
Final Prediction Error	0.47177302	NaN
Maximum Absolute Error	5.46329979	5.33514699
Mean Square Error	0.46929686	0.46014763
Sum of Frequencies	10803	10803
Number of Estimate Weights	57	NaN
Root Average Sum of Squares	0.68324279	0.67834182
Root Final Prediction Error	0.68685735	NaN
Root Mean Squared Error	0.68505245	0.67834182

Least Absolute Shrinkage and Selection Operator (LASSO)

The lasso method adds and deletes parameters based on a version of ordinary least squares where the sum of the absolute regression coefficients is constrained. If the parameter *t* is small, some of the regression coefficients are zero. The parameter t is increased in discrete steps, that is, the

nonzero coefficients selected at a step, correspond to the selected parameters of the potential input variables.

Table 39: Fit statistics – Least Absolute Shrinkage and Selection Operator (LASSO)

Statistics Label	Train	Validation
Akaike's Information Criterion	-8115.8337	NaN
Average Squared Error	0.4668207	0.46014763
Average Error Function	0.4668207	0.46014763
Divisor for ASE	10803	10803
Error Function	5043.06406	4970.97484
Final Prediction Error	0.47177302	NaN
Maximum Absolute Error	5.46329979	5.33514699
Mean Square Error	0.46929686	0.46014763
Sum of Frequencies	10803	10803
Number of Estimate Weights	57	NaN
Root Average Sum of Squares	0.68324279	0.67834182
Root Final Prediction Error	0.68685735	NaN
Root Mean Squared Error	0.68505245	0.67834182
Schwarz's Bayesian Criterion	-7700.4417	NaN
Sum of Squared Errors	5043.06406	4970.97484
Sum of Case Weights Times Freq	10803	10803

Adaptive LASSO

This is a modification of the LASSO algorithm. The basics are the same, but weights are applied to the parameters in the LASSO constraint.

Table 40: Fit statistics – Adaptive LASSO

Statistics Label	Train	Validation
Akaike's Information Criterion	-8126.6972	NaN
Average Squared Error	0.46695626	0.46031898
Average Error Function	0.46695626	0.46031898
Divisor for ASE	10803	10803
Error Function	5044.52849	4972.82599

Control Point

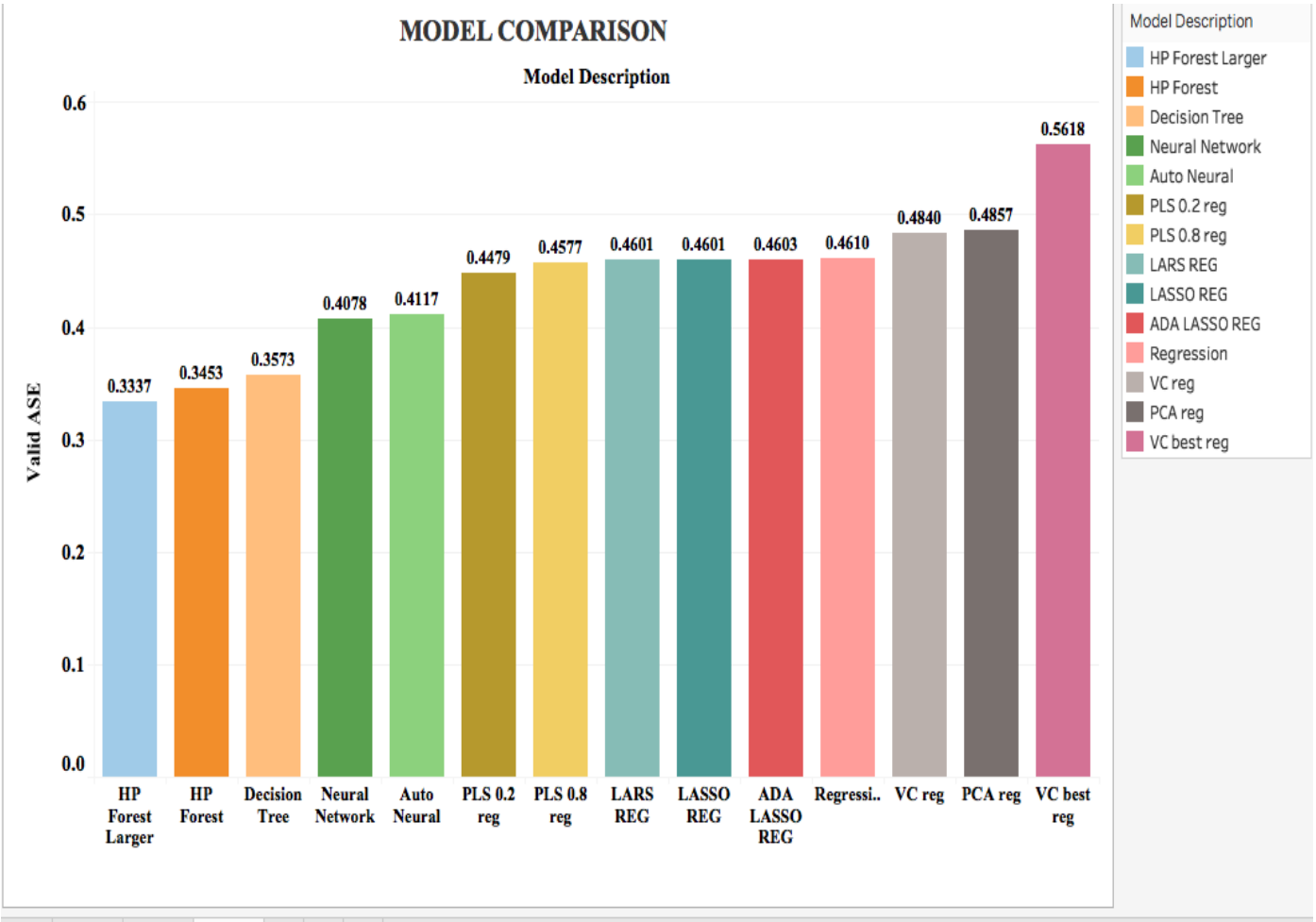
The control point tool was used to establish a control point to reduce the number of connections that were made in the process flow diagram.

Model Comparison:

After building several models we knew it was impertinent to compare their performance. The Model Comparison tool collected the information from the attached model nodes, and it enabled us to easily compare the average squared error.

Model Description	Target	Train ASE	Valid ASE	Train RMSE	Valid RMSE
HP Forest LARGER	LOG_VALP	0.3159	0.3337	0.5620	3604.4851
HP Forest	LOG_VALP	0.3369	0.3453	0.5804	3730.4605
Decision Tree	LOG_VALP	0.3441	0.3573	0.5866	3859.9048
Neural Network	LOG_VALP	0.4121	0.4078	0.6419	4405.2955
AutoNeural	LOG_VALP	0.4152	0.4117	0.6444	4447.1461
PLS 0.2 reg	LOG_VALP	0.4595	0.4479	0.6778	4838.9773
PLS 0.8 reg	LOG_VALP	0.4721	0.4577	0.6871	4944.5072
LARS REG	LOG_VALP	0.4668	0.4601	0.6832	4970.9748
LASSO REG	LOG_VALP	0.4668	0.4601	0.6832	4970.9748
ADA LASSO REG	LOG_VALP	0.4670	0.4603	0.6833	4972.8260
Regression	LOG_VALP	0.4676	0.4610	0.6838	4980.5967
VC reg	LOG_VALP	0.5041	0.4840	0.7100	5229.0853
PCA reg	LOG_VALP	0.5090	0.4857	0.7135	5246.8531

VC best reg	**LOG_VALP**	**0.5844**	**0.5618**	**0.7645**	**6069.5045**



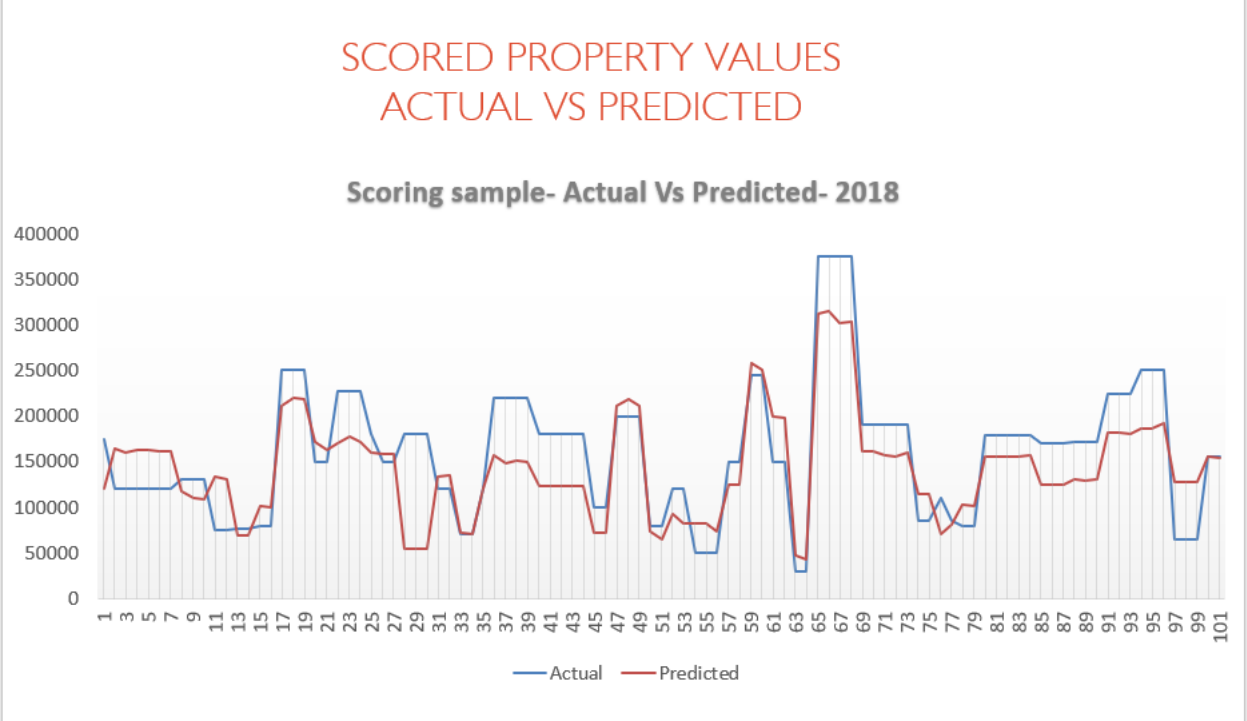
Scoring

The score data generates predicted values for a dataset that does not contain the target variable, "VALP" (property value). The score tool generates and manages scoring formulas in the form of a single SAS DATA step. This can be used in most SAS environments even without the presence of SAS® Enterprise Miner™.

In order to generate predicted target values outside of SAS® Enterprise Miner™, scoring node module was used and imported into SAS Studio. This predictive model was run against a

scoring dataset with 25,013 observations from 2018 US Oklahoma Census Bureau data. We deleted the "VALP" variable from the data and later compared the output "VALP" prediction column (by converting to inverse exponential as our "VALP" variable is in logarithmic target values) with the original "VALP" data. It yielded a similar pattern of predictions with deviations at the outliers which have high original "VALP" values as revealed in the chart below:

Figure 32: Scored property values with first 100 values



Interesting Findings

1. It was discovered that the HPDM Random Forest has the best average squared error (0.3337).
2. Yearly Insurance cost, Household Income, and persons per household explain the property value and are important variables.
3. No previous research speaks about the Oklahoma household language and its contribution to the property value.

4. Individual attributes MV (moved into the house) and ACR (size of the lot) play a major role in determining the property value.
5. Number of related children are more for never married category of the marital status variable.

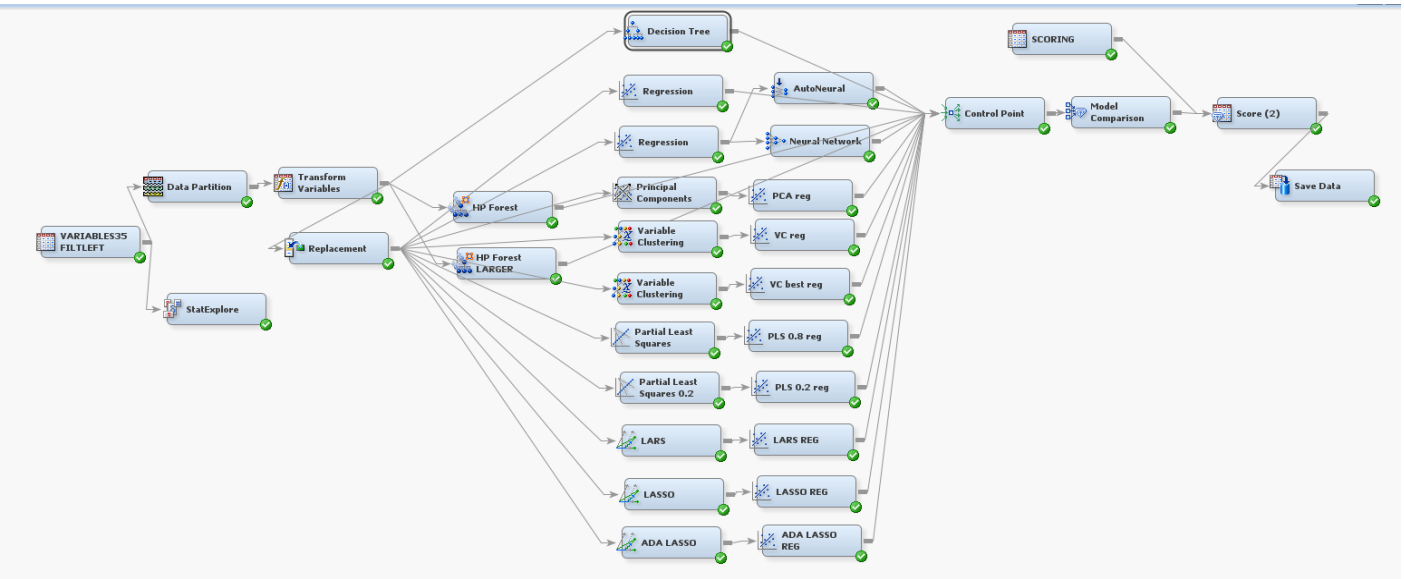
Recommendations

Based on the research findings, the following recommendations are proposed:

- Financial institutions can consider social variables like household language, number of related children and marital status to improve home valuation predictions.
- Most home insurance policies cover the replacement cost of the home; therefore, insurance companies can use significant factors like lot size, units in structure, and number of rooms to estimate the home's replacement cost.

Conclusion

We have explored the application of diverse machine learning techniques to estimate property value in Oklahoma. HPDM Random Forest model offers the best accuracy with the least average squared error. Factors that influence the property values: Insurance, house income, number of rooms, number related children in household and lot size. This model can be useful to assist financial institutions to get a more accurate property appraisal, as well as enable consistent mortgage loans and lending decisions by improving the existing models.


Decision Tree
SCORING
Regression
AutoNeural
Regression
Neural Network
Control Point
Model Comparison
Score (2)
Data Partition
Transform Variables
HP Forest
Principal Components
PCA reg
Save Data
VARIABLES35 FILTLEFT
Replacement
HP Forest LARGER
Variable Clustering
VC reg
Variable Clustering
VC best reg
StatExplore
Partial Least Squares
PLS 0.8 reg
Partial Least Squares 0.2
PLS 0.2 reg
LARS
LARS REG
LASSO
LASSO REG
ADA LASSO
ADA LASSO REG

References

Adelino, Manuel & Schoar, Antoinette & Severino, Felipe. (2013). House Prices, Collateral and Self-Employment. Journal of Financial Economics. 117. 10.1016/j.jfineco.2015.03.005.

Afonso, Bruno & Melo, Luckeciano & Dihanster, Willian & Sousa, Samuel & Berton, L.. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble.

Bogin, A.N., Shui, J. Appraisal Accuracy and Automated Valuation Models in Rural Areas. *J Real Estate Finan Econ* **60,** 40–52 (2020). https://doi.org/10.1007/s11146-019-09712-0

Christie, P., Georges, J., Thompson, J., & Wells, C. (2017). *Applied Analytics Using SAS Enterprise Miner.* Cary, North Carolina, United States: SAS Institute Inc.

Fan, G., Ong, S., & Koh, H. (2006). Determinants of house price: A decision tree approach. *Urban studies*, *43*(12): 2301-2315.

Feng, Y., & Jones, K. (2015). Comparing multilevel modelling and artificial neural networks in house price prediction. *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, 108-114.

Hong, Jengei & Choi, Heeyoul "Henry & Kim, Woo-sung. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. International Journal of Strategic Property Management. 24. 1-13. 10.3846/ijspm.2020.11544.

Huang, Y. (2019). Predicting home value in California, United States via machine learning modeling. *Statistics, optimization & information computing, 7*(1): 66-74.

Lee, C., Culhane, D. P., & Wachter, S. M. (1999). The differential impacts of federally assisted housing programs on nearby property values: A Philadelphia case study. https://repository.upenn.edu/spp_papers/64

Liew, C., Haron, N.A. (2013). Factors Influencing the Rise of House Price in Klang Valley. International Journal of Research in Engineering and Technology. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.4551&rep=rep1&type=pdf

Lu, Sifei., Li, Zengxiang., Qin, Zheng., Yang, Xulei., & Goh, Rick. (2017). A hybrid regression technique for house prices prediction. 319-323. 10.1109/IEEM.2017.8289904. https://www.researchgate.net/publication/323135322_A_hybrid_regression_technique_for_house_prices_prediction

Mills, Arlen C,M.A.I., S.R.A. (2019). Residential perspective on data collection and property description in the valuation process. *The Appraisal Journal, 87*(1), 42-53. Retrieved from http://vortex3.uco.edu/login?url=https://search-proquest-com.vortex3.uco.edu/docview/2235639830?accountid=14516

Montero, J.M., Mínguez, R., & Fernández-Avilés, G. (2018). Housing price prediction: parametric versus semi-parametric spatial hedonic models. *Journal of geographical systems*, *20*(1): 27–55. https://doi-org.vortex3.uco.edu/10.1007/s10109-017-0257-y.

Mu, J., Wu, F., & Zhang, A. (2014). Housing value forecasting based on machine learning methods. *Abstract and applied analysis*, *20*(14):1-7.

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*(6):2928–2934. https://doiorg.vortex3.uco.edu/10.1016/j.eswa.2014

Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic modeling*, *45*: 259–267.https://doiorg.vortex3.uco.edu/10.16

Quick Facts Oklahoma. (2018, July 01). United States Census Bureau:

https://www.census.gov/quickfacts/fact/table/OK/IPE120218#IPE120218

SAS Institute (2020). https://www.sas.com/en_id/insights/analytics/machine-learning.html

Quigley, J.M. (1999). Real estate prices and economic cycles. *International real estate review, 2*(1):1-20. https://escholarship.org/uc/item/58c6v2kx

SAS Institute Inc. 2017. Advanced Predictive Modeling Using SAS® Enterprise Miner: Course Notes. Cary, NC: SAS Institute Inc.

Shaaf, M. B. (2005-2006). An Analysis of the Housing Market and the Oklahoma Experience. *Southwest Business and Economics Journal,* 63-77. Retrieved July 07, 2020.

Shahhosseini, M., Hu, G., & Pham, H. (2019). Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction.

So, K.S., Orazem, P.F., & Otto, D.L. (2001). The Effects of Housing Prices, Wages, and Commuting Time on Joint Residential and Job Location Choices. *Transportation & the Spatial Structure eJournal*.

Tanjil, M., & Chakraborty, G. (2016). Predicting current market value of a housing unit across the four census regions of the United States using SAS® Enterprise Miner™.

United States Census Bureau. (2018). census.gov: https://www.census.gov/programs surveys/acs/data/pums.html

White, M. (2015). Cyclical and structural change in the UK housing market. *Journal of European real estate research, 8*(1):85-103. doi:http://dx.doi.org.vortex3.uco.edu/10.1108/JERER-02-2014-0011

Wu, T., Cheng, M., & Wong, K. (2017). Bayesian analysis of Hong Kong's housing price dynamics. *Pacific economic review*, *22*(3): 312–331. https://doi-org.vortex3.uco/10.12232

Worzala, E., Lenk, M., & Silva, A. (1995). An Exploration of Neural Networks and Its Application to Real Estate Valuation. *The Journal of Real Estate Research, 10*(2), 185-201. Retrieved June 28, 2020, from www.jstor.org/stable/24881837