https://posit.cloud/content/6265838

ADVANCED BUSINESS ANALYTICS CLASS

COMPARING PREDICTIVE MODELS FOR HOUSING PRICES: COMPREHENSIVE INCORPORATION OF ALL VARIABLES AND CATEGORICAL DUMMY VARIABLES VS. SIMPLIFIED NUMERIC AND LEVEL VARIABLES APPROACH

Clarissa Xaviera Nadia (29022003)
Reza Setiadi Shihran (29022004)
Safira Fabilia (29022011)
Farras Tamir (29022014)

INTRODUCTION

BACKGROUND ISSUE

The importance and significance of prediction analytics have become essential for businesses across industries. Predictive modeling has emerged as a powerful tool that holds the potential to transform the way businesses strategize. **This research report presents an in-depth analysis and prediction model using the Ames Housing Dataset from Kaggle**, understanding the factors that influence house prices and how predictive analytics can be used as a tool to forecast property values. **This research report focuses on housing price prediction by developing multiple predictive models, this study aims to uncover the most effective approach for forecasting house prices** and understanding the factors that influence property values in Ames, Iowa.

Prediction analytics plays a critical role in the world of business by providing valuable insights into the future. The ability to accurately predict house prices in the dynamic housing market can change the way real estate professionals, investors, and homeowners approach their decisions.

INTRODUCTION

BACKGROUND ISSUE

The Ames Housing Dataset serves as the foundation of this research, providing a comprehensive set of data to analyze and build predictive models. By leveraging this dataset, we aim to find valuable insights into property pricing and demonstrate the significance of prediction analytics in driving decision-making within the real estate industry. **By comparing several prediction models, this study seeks to present a comprehensive understanding and implications on the Ames housing market.** Ultimately, this research aims to contribute to the body of knowledge and researchers surrounding predictive modeling in real estate, as well as to the practitioners in the industry to make more informed and strategic decisions in the housing market.

RESEARCH QUESTION

"What is the most effective modeling approach for predicting housing prices: a comprehensive model utilizing all variables, including categorical variables transformed as dummy variables, or a simplified model featuring only numeric and level variables without incorporating categorical variables?"

Literature Review

LITERATURE REVIEW

Prediction models play a vital role in various disciplines, and their significance is particularly clear in the field of housing price forecasting. The study of housing prices holds substantial importance within the business context. As discussed earlier in the introduction of this study, predictive modeling and data analytics offer valuable insights for several parties like researchers and practitioners or actors in the industry such as real estate professionals, investors, and even homeowners.

Numerous studies have explored the use of different prediction models to comprehend the complex dynamics of the housing market:

Joshi et al. (2022) investigated house price prediction in Bangalore using a diverse set of methods, including linear regression, bagging classifier, K-nearest neighbor, XGB, decision tree, gradient boosting, and random forest. **Their findings revealed that the random forest algorithm proved to be the most appropriate and effective in handling the available data.** Similarly, **Madhuri et al. (2019)** explored various prediction models, including multiple linear, Ridge, LASSO, Elastic Net, gradient boosting, and Ada Boost Regression, and **identified gradient boosting as the most accurate algorithm for house price predictions.**

These studies exemplify the importance of understanding and studying different prediction models as they can significantly impact the accuracy and reliability of housing price forecasts.

LITERATURE REVIEW

Truong et al. (2020) explored into the prediction of housing prices using various advanced models, including random forest, extreme gradient boosting, light gradient boosting machine, hybrid regression, and stacked generalization. **Their results showcased the superior performance of Random Forest for training sets and Stacked Generalization Regression for test sets.**

Zhang et al. (2021) examined major factors influencing housing prices using Spearman correlation coefficient and established a multiple linear regression model for housing price prediction. **Their study demonstrated the effectiveness of the multiple linear regression model** in predicting and analyzing housing prices.

Kang et al. (2020) examine how well house price appreciation potentials can be predicted by combining multiple data sources. This study models house price appreciation through multiple linear regression, geographically weighted regression, and gradient boosting machine.

Manasa et al. (2020) explored house price prediction using several methods. In this study, modeling explorations apply some regression techniques such as linear regression, Lasso and Ridge regression, support vector regression, and boosting algorithms such as Extreme Gradient Boost Regression (XG Boost).

Prediction models enable researchers and stakeholders to capture underlying patterns and relationships within the data, identify crucial predictor variables, and explore significant factors affecting housing prices. These models provide valuable insights into the dynamic nature of the housing market.

LITERATURE REVIEW

The previous studies highlight the need to leverage multiple data sources and explore different prediction methods to model house price potentials effectively. Emphasizing housing price studies in business decision-making ensures a competitive edge and a better understanding of market trends. Through predictive analytics, understanding house prices allows stakeholders to identify potential investment opportunities, assess risk factors, and optimize strategies for profit maximization.

	Previous Studies	Prediction Models/Predictive Modeling Methods
Forecasting House Price	Joshi et al. (2022)	Random forest, linear regression, gradient boosting, XGB, K-nearest neighbour, bagging classifier, and decision tree
	Kang et al. (2021)	Multiple linear regression, geographically weighted regression, and gradient boosting machine
	Madhuri et al. (2019)	Multiple linear regression, Ridge, LASSO, Elastic Net, gradient boosting, and Ada Boost Regression
	Manasa et al. (2020)	Linear regression, Lasso and Ridge regression, support vector regression, and extreme gradient boost regression
	Truong et al. (2020)	Random forest, extreme gradient boosting, light gradient boosting machine, hybrid regression, and stacked generalization
	Zhang (2021)	Multiple linear regression

01 Joshi, I., Mudgil, P., & Bisht, A. (2022). House Price Forecasting by Implementing Machine Learning Algorithms: A Comparative Study. *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3* (pp. 63-71).

02 Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy, 111*, 104919.

03 Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. *In 2019 International conference on smart structures and systems (ICSSS)* (pp. 1-5).

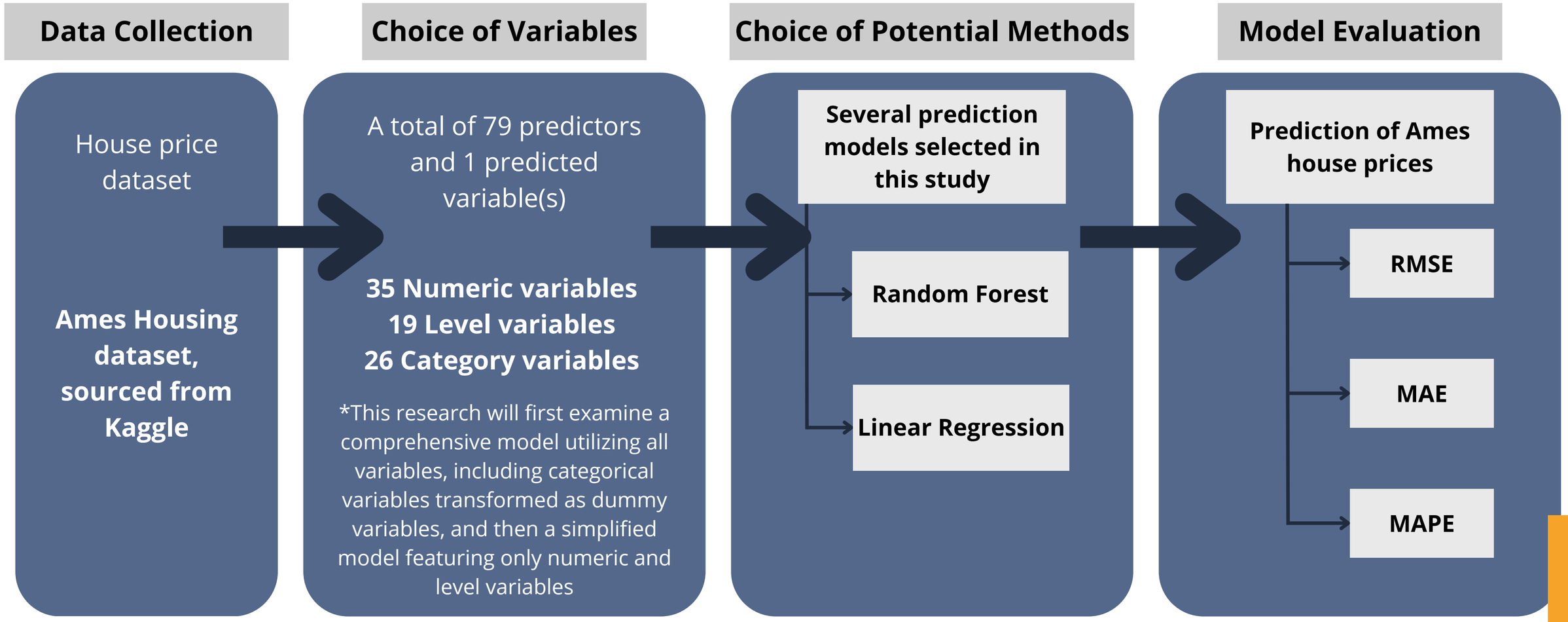
04 Manasa, J., Gupta, R., & Narahari, N. S. (2020). Machine Learning based Predicting House Prices using Regression Techniques. *In 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)* (pp. 624-630).

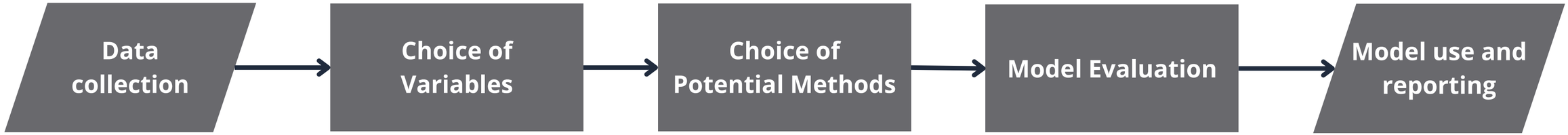
05 Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science, 174*, 433-442.

06 Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming, 2021*, 1-9.

Research Framework

RESEARCH FRAMEWORK





DATA COLLECTION

In this research, the data collection process centers around the selection of the **Ames Housing dataset, sourced from a Kaggle competition** and curated by Dean De Cock for data science education.

The decision to utilize this specific dataset originates from its exceptional attributes, setting it apart compared to other alternatives. **The Ames dataset is more comprehensive, encompassing a rich array of variables,** making it an ideal choice for conducting an in-depth analysis of factors influencing housing prices. Furthermore, **this dataset has been extensively used and tested by numerous data analysts,** supporting its reliability and suitability for predictive modeling.

By utilizing this well-established and robust dataset, this research aims to gain valuable insights into the dynamics of the housing market in Ames, Iowa, as well as contribute to the advancement of predictive modeling techniques in the context of the real estate industry.

Data Description

VARIABLES

Variables	Description	Type of data
SalePrice	The property's sale price in dollars. This is the target variable that trying to predict	numeric
MSSubClass	The building class	numeric
MSZoning	The general zoning classification	category
LotFrontage	Linear feet of street connected to property	numeric
LotArea	Lot size in square feet	numeric
Street	**Type of road access**	**category**
Alley	**Type of alley access**	**category**
LotShape	General shape of property	category
LandContour	Flatness of the property	category
Utilities	Type of Utilities available	category
LotConfig	Lot configuration	category
LandSlope	Slope of property	level
Neighborhood	Physical locations within Ames city limits	category
Condition1	Proximity to main road or railroad	category
Condition2	Proximity to main road or railroad (if a second is present)	category
BldgType	Type of dwelling	category
HouseStyle	Style of dwelling	category
OverallQual	Overall material and finish quality	level
OverallCond	Overall condition rating	level
YearBuilt	Original construction date	numeric

Variables	Description	Type of data
YearRemodAdd	Remodel date	numeric
RoofStyle	Type of roof	category
RoofMatl	Roof material	category
Exterior1st	Exterior covering on house	category
Exterior2nd	Exterior covering on house (if more than one material)	category
MasVnrType	Masonry veneer type	category
MasVnrArea	Masonry veneer area in square feet	numeric
ExterQual	Exterior material quality	level
ExterCond	Present condition of the material on the exterior	level
Foundation	Type of foundation	category
BsmtQual	Height of the basement	level
BsmtCond	General condition of the basement	level
BsmtExposure	Walkout or garden level basement walls	level
BsmtFinType1	Quality of basement finished area	level
BsmtFinSF1	Type 1 finished square feet	numeric
BsmtFinType2	Quality of second finished area (if present)	level
BsmtFinSF2	Type 2 finished square feet	numeric
BsmtUnfSF	Unfinished square feet of basement area	numeric
TotalBsmtSF	Total square feet of basement area	numeric
Heating	Type of heating	category

VARIABLES (Continue)

Variables	Description	Type of data
HeatingQC	Heating quality and condition	level
CentralAir	**Central air conditioning**	**category**
Electrical	Electrical system	category
1stFlrSF	First Floor square feet	numeric
2ndFlrSF	Second floor square feet	numeric
LowQualFinSF	Low quality finished square feet (all floors)	level
GrLivArea	Above grade (ground) living area square feet	numeric
BsmtFullBath	Basement full bathrooms	numeric
BsmtHalfBath	Basement half bathrooms	numeric
FullBath	Full bathrooms above grade	numeric
HalfBath	Half baths above grade	numeric
Bedroom	Number of bedrooms above basement level	numeric
Kitchen	Number of kitchens	numeric
KitchenQual	Kitchen quality	level
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)	numeric
Functional	Home functionality rating	level
Fireplaces	Number of fireplaces	numeric
FireplacesQu	Fireplace quality	level
GarageType	Garage location	category
GarageYrBlt	Year garage was built	numeric

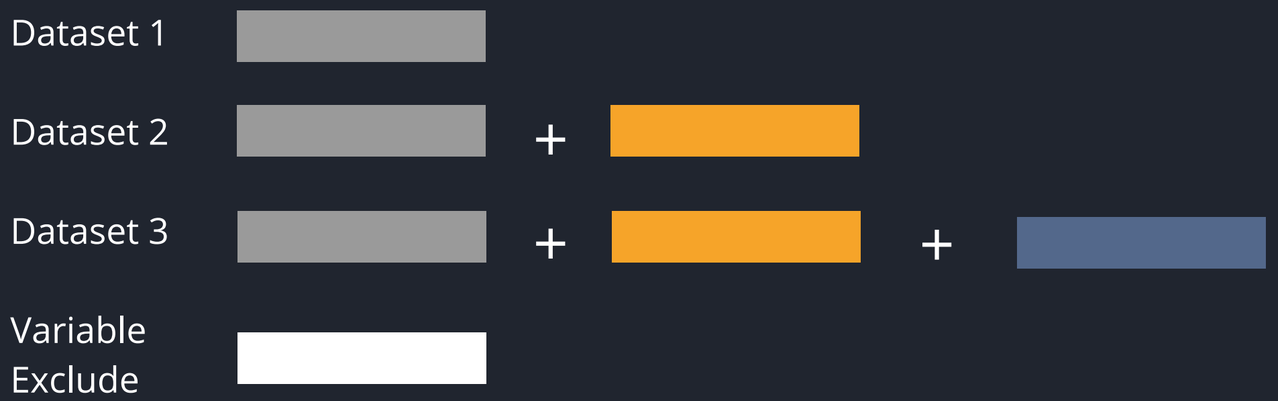
Variables	Description	Type of data
GarageFinish	Interior finish of the garage	level
GarageCars	Size of garage in car capacity	numeric
GarageArea	Size of garage in square feet	numeric
GarageQual	Garage quality	level
GarageCond	Garage condition	level
PavedDrive	**Paved driveaway**	**category**
WoodDeckSF	Wood deck area in square feet	numeric
OpenPorchSF	Open porch area in square feet	numeric
EnclosedPorch	Enclosed porch area in square feet	numeric
3SsnPorch	Three season porch area in square feet	numeric
ScreenPorch	Screen porch area in square feet	numeric
PoolArea	Pool area in square feet	numeric
PoolQC	Pool area in square feet	numeric
Fence	Fence quality	level
MiscFeature	Mescellaneous feature not covered in other categories	category
MiscVal	$value of miscellaneous feature	numeric
MoSold	Month Sold	numeric
YrSold	Year Sold	numeric
SaleType	Type of sale	category
SaleCondition	Condition of sale	category

GENERATE DATA COMPOSITION

In order to ensure the robustness of model, we construct the data composition into three dataset consisting of

1. dataset 1 is using all types of numerical data and level data.
2. dataset 2 is formed by all the first dataset plus several category variables. The category selected in this second dataset consists of columns with the names "Street", "CentralAir", "PavedDrive", and "Alley". These variable names were chosen because it is easy to quantify these data categories compared to other types of category variables
3. dataset 3 is formed by all the second dataset plus using dummy variables for data category.

In order to facilitate understanding of the formation of this composition data, the author illustrates it in color as follows



CATEGORY VARIABELS

Variable	Code	Description
MSZoning	A	Agriculture
	C	Commercial
	FV	Floating Village Residential
	I	Industrial
	RH	Residential High Density
	RL	Residential Low Density
	RP	Residential Low Density Park
	RM	Residential Medium Density
Street	Grvl	Gravel
	Pave	Paved
Alley	Grvl	Gravel
	Pave	Paved
	NA	No alley access
LotShape	Reg	Regular
	IR1	Slightly irregular
	IR2	Moderately irregular
	IR3	Irregular
LandContour	Lvl	Near Flat/Level
	Bnk	Banked - quick and significant rise from street grade to building
	HLS	Hillside - significant slope from side to side
	Low	Depression

Variable	Code	Description
Utilities	AllPub	All public Utilities (E,G,W,& S)
	NoSewr	Electricity, Gas, and Water (Septic Tank)
	NoSeWa	Electricity and Gas Only
	ELO	Electricity only
LotConfig	Inside	Inside lot
	Corner	Corner lot
	CulDSac	Cul-de-sac
	FR2	Frontage on 2 sides of property
	FR3	Frontage on 3 sides of property
Neighborhood	Blmngtn	Bloomington Heights
	Blueste	Bluestem
	BrDale	Briardale
	BrkSide	Brookside
	ClearCr	Clear Creek
	CollgCr	College Creek
	Crawfor	Crawford
	Edwards	Edwards
	Gilbert	Gilbert
	IDOTRR	Iowa DOT and Rail Road
	MeadowV	Meadow Village
	Mitchel	Mitchell
	Names	North Ames

Variable	Code	Description
Neighborhood	NoRidge	Northridge
	NPkVill	Northpark Villa
	NridgHt	Northridge Heights
	NWAmes	Northwest Ames
	OldTown	Old Town
	SWISU	South & West of Iowa State University
	Sawyer	Sawyer
	SawyerW	Sawyer West
	Somerst	Somerset
	StoneBr	Stone Brook
	Timber	Timberland
	Veenker	Veenker
	Artery	Adjacent to arterial street
	Feedr	Adjacent to feeder street
Condition1 & Condition2	Norm	Normal
	RRNn	Within 200' of North-South Railroad
	RRAn	Adjacent to North-South Railroad
	PosN	Near positive off-site feature--park, greenbelt, etc.
	PosA	Adjacent to postive off-site feature
	RRNe	Within 200' of East-West Railroad
	RRAe	Adjacent to East-West Railroad

CATEGORY VARIABELS

(Continue)

Variable	Code	Description
BldgType	**1Fam**	**Single-family Detached**
	2FmCon	**Two-family Conversion; originally built as one-family dwelling**
	Duplx	**Duplex**
	TwnhsE	**Townhouse End Unit**
	TwnhsI	**Townhouse Inside Unit**
HouseStyle	1Story	One story
	1.5Fin	One and one-half story: 2nd level finished
	1.5Unf	One and one-half story: 2nd level unfinished
	2Story	Two story
	2.5Fin	Two and one-half story: 2nd level finished
	2.5Unf	Two and one-half story: 2nd level unfinished
	SFoyer	Split Foyer
	SLvl	Split Level
RoofStyle	**Flat**	**Flat**
	Gable	**Gable**
	Gambrel	**Gabrel (Barn)**
	Hip	**Hip**
	Mansard	**Mansard**
	Shed	**Shed**
RoofMatl	ClyTile	Clay or Tile
	CompShg	Standard (Composite) Shingle
	Membran	Membrane

Variable	Code	Description
RoofMatl	**Metal**	**Metal**
	Roll	**Roll**
	Tar&Grv	**Gravel & Tar**
	WdShake	**Wood Shakes**
	WdShngl	**Wood Shingles**
Exterior1st & Exterior2nd	AsbShng	Asbestos Shingles
	AsphShn	Asphalt Shingles
	BrkComm	Brick Common
	BrkFace	Brick Face
	CBlock	Cinder Block
	CemntBd	Cement Board
	HdBoard	Hard Board
	ImStucc	Imitation Stucco
	MetalSd	Metal Siding
	Other	Other
	Plywood	Plywood
	PreCast	PreCast
	Stone	Stone
	Stucco	Stucco
	VinylSd	Vinyl Siding
	WdSdng	Wood Siding
	WdShing	Wood Shingles

Variable	Code	Description
MasVnrType	**BrkCmn**	**Brick Common**
	BrkFace	**Brick Face**
	CBlock	**Cinder Block**
	None	**None**
	Stone	**Stone**
Foundation	BrkTil	Brick & Tile
	CBlock	Cinder Block
	PConc	Poured Concrete
	Slab	Slab
	Stone	Stone
	Wood	Wood
Heating	**Floor**	**Floor Furnace**
	GasA	**Gas forced warm air furnace**
	GasW	**Gas hot water or steam heat**
	Grav	**Gravity furnace**
	OthW	**Hot water or steam heat other than gas**
	Wall	**Wall furnace**
Central Air	N	No
	Y	Yes
Electrical	**SBrkr**	**Standard Circuit Breakers & Romex**
	FuseA	**Fuse Box over 60 AMP and all Romex wiring (Average)**
	FuseF	**60 AMP Fuse Box and mostly Romex wiring (Fair)**
	FuseP	**60 AMP Fuse Box and mostly knob & tube wiring (poor)**
	Mix	**Mixed**

CATEGORY VARIABELS

(Continue)

Variable	Code	Description
GarageType	2Types	More than one type of garage
	Attchd	Attached to home
	Basment	Basement Garage
	BuiltIn	Built-In (Garage part of house - typically has room above garage)
	CarPort	Car port
	Detchd	Detached from home
	NA	No Garage
PavedDrive	Y	Paved
	P	Partial Pavement
	N	Dirt/Gravel
MiscFeature	Elev	Elevator
	Gar2	2nd Garage (if not described in garage section)
	Othr	Other
	Shed	Shed (over 100 SF)
	TenC	Tennis Court
	NA	None

Variable	Code	Description
SaleType	WD	Warranty Deed - Conventional
	CWD	Warranty Deed - Cash
	VWD	Warranty Deed - VA Loan
	New	Home just constructed and sold
	COD	Court Officer Deed/Estate
	Con	Contract 15% Down payment regular terms
	ConLw	Contract Low Down payment and low interest
	ConLI	Contract Low Interest
	ConLD	Contract Low Down
	Oth	Other
SaleCondition	Normal	Normal Sale
	Abnorml	Abnormal Sale - trade, foreclosure, short sale
	AdjLand	Adjoining Land Purchase
	Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
	Family	Sale between family members
	Partial	Home was not completed when last assessed (associated with New Homes)

CODING ON LEVEL VARIABELS

Name Variable	Level of Variable	Description of Level	Coding numeric
LandSlope	**Gtl**	**Gentle Slope**	**1**
	Mod	**Moderate Slope**	**2**
	Sev	**Severe Slope**	**3**
ExterQual	Po	Poor	1
	Fa	Fair	2
	TA	Average/Typical	3
	Gd	Good	4
	Ex	Excellent	5
ExtrCond	**Po**	**Poor**	**1**
	Fa	**Fair**	**2**
	TA	**Average/Typical**	**3**
	Gd	**Good**	**4**
	Ex	**Excellent**	**5**
BsmtQual	Po	Poor	1
	Fa	Fair	2
	TA	Average/Typical	3
	Gd	Good	4
	Ex	Excellent	5
BsmtExposure	**NA**	**No Basement**	**0**
	No	**No Exposure**	**1**
	Mn	**Minimum Exposure**	**2**
	Av	**Average Exposure**	**3**
	Gd	**Good Exposure**	**4**

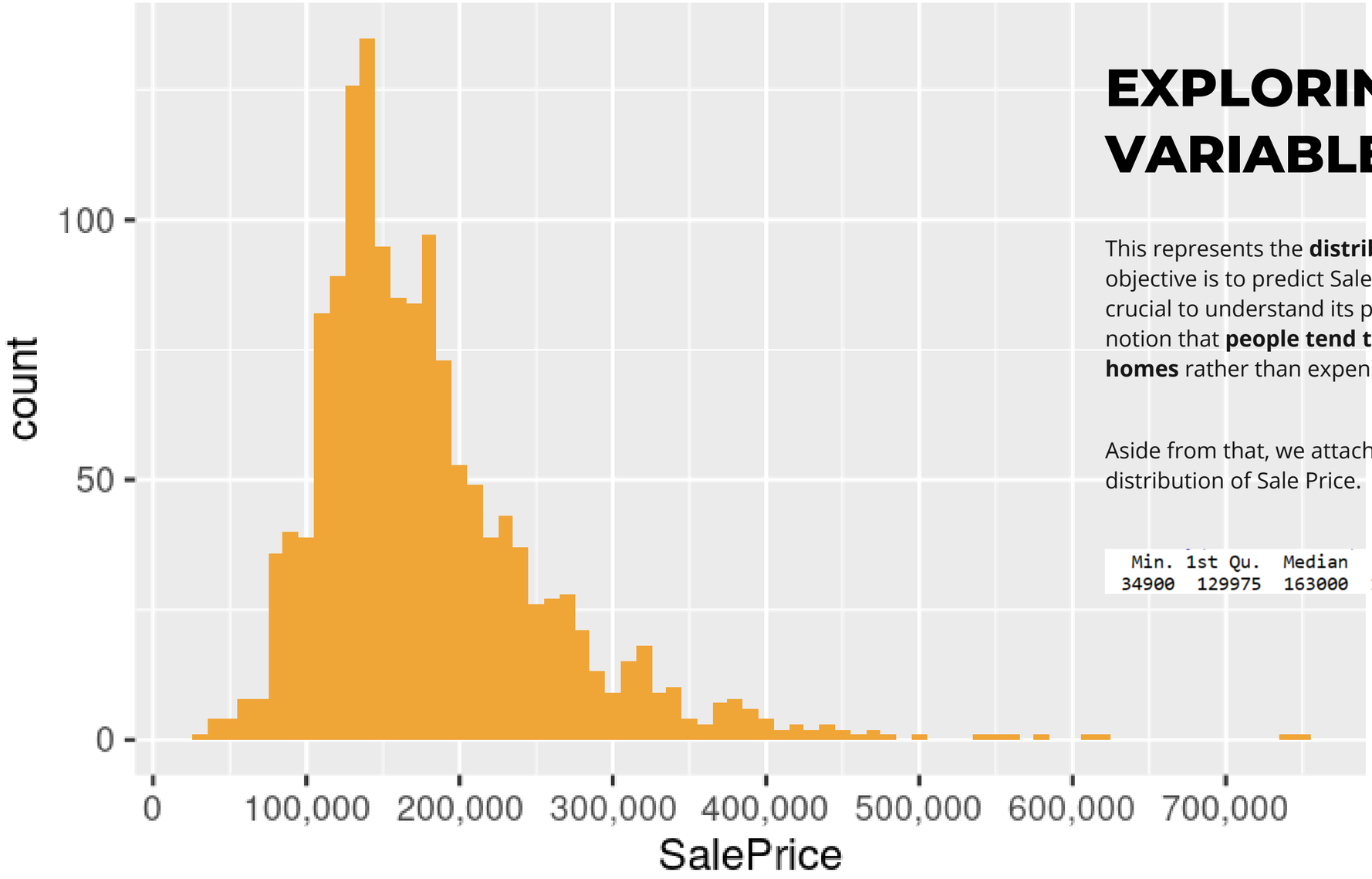
Name Variable	Level of Variable	Description of Level	Coding numeric
BsmtFinType1	**NA**	**No Basement**	**0**
	Unf	**Unfinished**	**1**
	LwQ	**Low Quality**	**2**
	Rec	**Acerage Rec Room**	**3**
	BLQ	**Below Average Living Quarters**	**4**
	ALQ	**Average Living Quarters**	**5**
	GLQ	**Good Living Quarters**	**6**
BsmtFinType2	NA	No Basement	0
	Unf	Unfinished	1
	LwQ	Low Quality	2
	Rec	Acerage Rec Room	3
	BLQ	Below Average Living Quarters	4
	ALQ	Average Living Quarters	5
	GLQ	Good Living Quarters	6
HeatingQC	**Po**	**Poor**	**1**
	Fa	**Fair**	**2**
	TA	**Average/Typical**	**3**
	Gd	**Good**	**4**
	Ex	**Excellent**	**5**
KitchenQual	Po	Poor	1
	Fa	Fair	2
	TA	Average/Typical	3
	Gd	Good	4
	Ex	Excellent	5

CODING ON LEVEL VARIABELS (Continue)

Name Variable	Level of Variable	Description of Level	Coding numeric
Functional	**Sal**	**Salvage Only**	**0**
	Sev	**Severly Damaged**	**1**
	Maj2	**Major Deductions 2**	**2**
	Maj1	**Major Deductions 1**	**2**
	Mod	**Moderate**	**3**
	Min2	**Minor Deduction2**	**4**
	Min1	**Minor Deduction1**	**4**
	Typ	**Typical Functional**	**5**
FireplaceQu	NA	No Fireplace	0
	Po	Poor	1
	Fa	Fair	2
	TA	Average/Typical	3
	Gd	Good	4
	Ex	Excellent	5
GarageFinish	**NA**	**No Garage**	**0**
	Unf	**Unfinished**	**1**
	Rfn	**Rough Finished**	**2**
	Fin	**Finished**	**3**
GarageQual	NA	No Fireplace	0
	Po	Poor	1
	Fa	Fair	2
	TA	Average/Typical	3
	Gd	Good	4
	Ex	Excellent	5

Name Variable	Level of Variable	Description of Level	Coding numeric
GarageCond	**NA**	**No Fireplace**	**0**
	Po	**Poor**	**1**
	Fa	**Fair**	**2**
	TA	**Average/Typical**	**3**
	Gd	**Good**	**4**
	Ex	**Excellent**	**5**
PoolQc	NA	No Fireplace	0
	Fa	Fair	1
	TA	Average/Typical	2
	Gd	Good	3
	Ex	Excellent	4
Fence	**NA**	**No Fence**	**0**
	MnWw	**Minimum Wood/Wire**	**1**
	GdWo	**Good Wood**	**2**
	MnPrv	**Minimum Privacy**	**3**
	GdPrv	**Good Privacy**	**4**

Exploratory Data Analysis

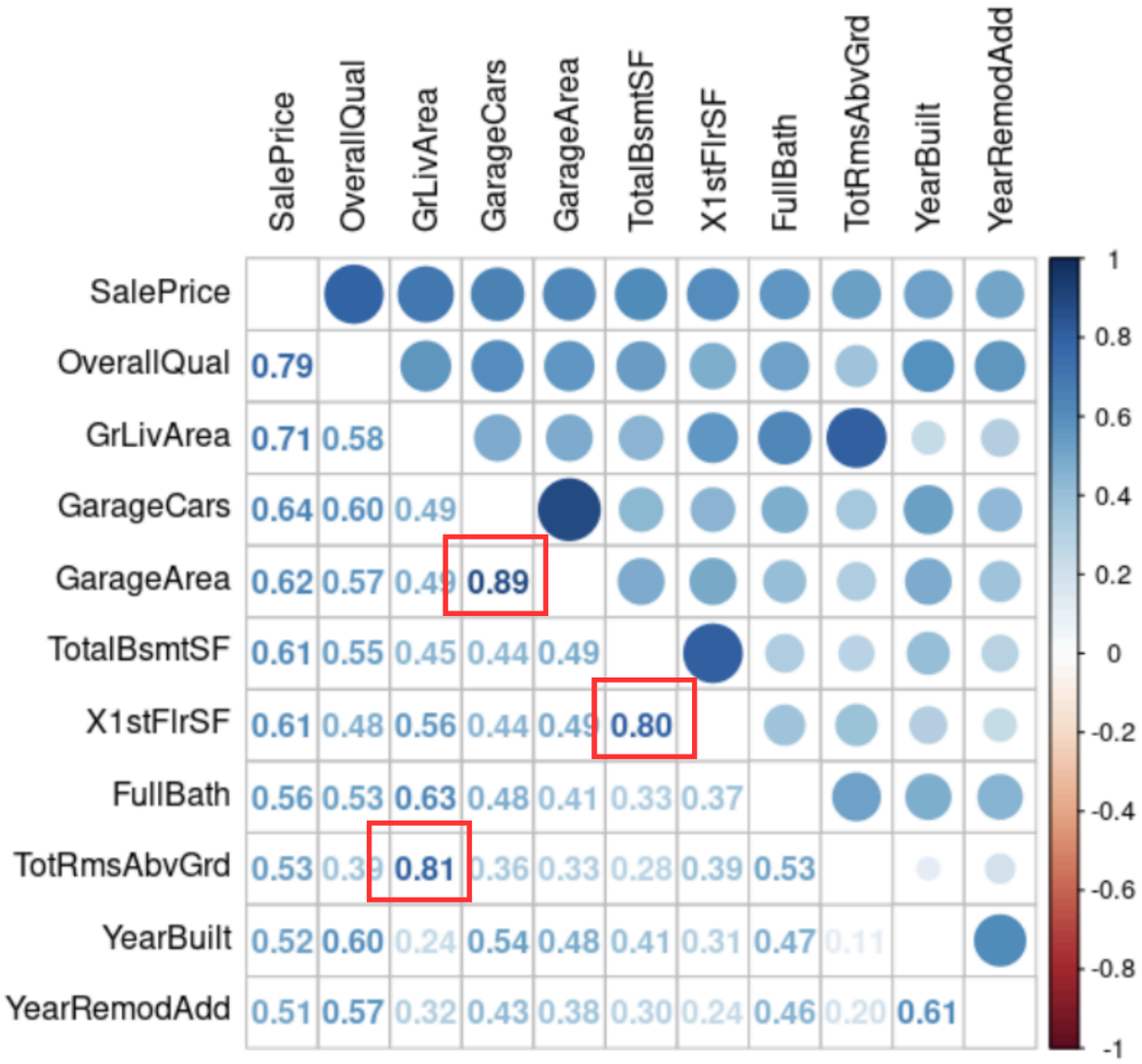


EXPLORING RESPONSE VARIABLE SALEPRICE

This represents the **distribution of SalePrice**. Since our objective is to predict SalePrice based on these features, it's crucial to understand its position. The findings support the notion that **people tend to afford more affordable homes** rather than expensive ones.

Aside from that, we attached a **summary** of the distribution of Sale Price.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
34900	129975	163000	180921	214000	755000	1459



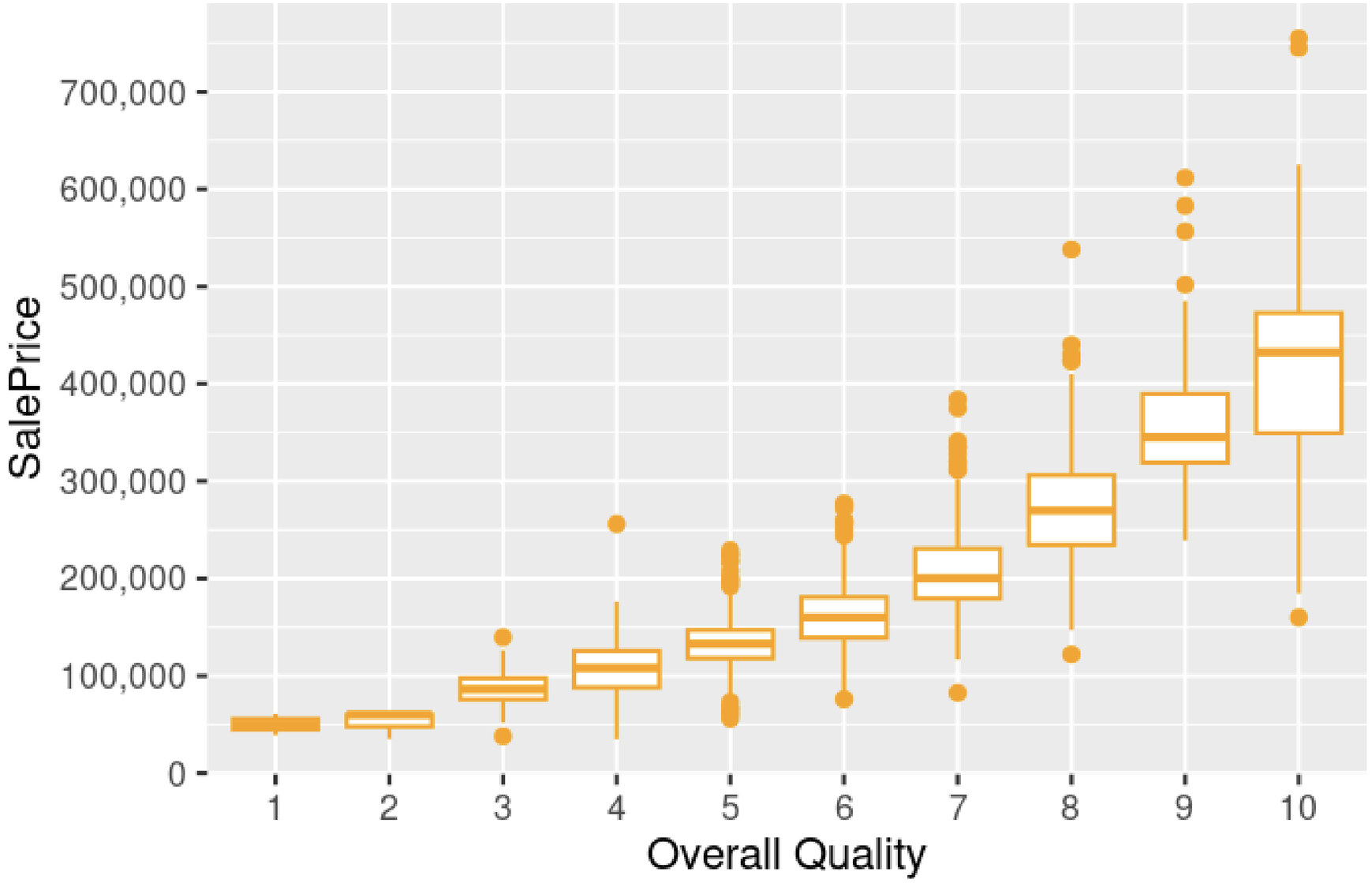
OVERALL CORRELATION PLOT

In order to construct the most effective model possible, we need to utilize **features that exhibit strong correlations** with the predictor variable (SalePrice in this case)

To identify these influential features, we generated a **correlation plot**, where darker blue shades indicate stronger relationships

In accordance with **SalePrice**, the **order of features** are: OverallQual (0.79), GrLivArea (0.71), GarageCars (0.64), GarageArea (0.62), TotalBsmtSF (0.61), X1stFlrSF (0.61), FullBath (0.56), TotRmsAbvGrd (0.53), YearBuilt (0.52), & YearRemodAdd (0.51)

The **top three overall correlations** are observed between: GarageCars and GarageArea, TotalBsmtSF and 1stFlrSF, and GrLivArea and TotRmsAbvGrd.



OVERALL QUALITY

Finding the overall quality is crucial for predictive modeling and data analysis in real estate as it strongly correlates with the sale price and **provides valuable insights** for feature selection.

Among the numeric variables, Overall Quality exhibits the **strongest correlation with SalePrice** (0.79) from the previous correlation plot, representing the overall material and finish of the house on a scale from 1 (very poor) to 10 (very excellent).

Results

METHODOLOGY

Linear Regression

Linear regression stands as a highly renowned algorithm in both statistics and machine learning. Its primary purpose is to establish a connection between one or more features (independent/ explanatory/ predictor variables) and a continuous target variable (dependent/ response variable). When dealing with one feature, the model adopts the form of simple linear regression, while with multiple features, it transforms into multiple linear regression (Manasa et al., 2020).

$$Y = a + bX$$

Where

Y = Independent Variable

b = Slope Line

a = intercept

X = Explanatory Variable

The formulation of linear model is:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots\ldots\ldots + \alpha_p X_p$$

The assumptions in the model are: the error terms are normally distributed, the error terms have constant variance, the model carries out a linear relationship between the target variable and the functions.

Manasa, J., Gupta, R., & Narahari, N. S. (2020). Machine Learning based Predicting House Prices using Regression Techniques. ***In 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*** **(pp. 624-630).**

METHODOLOGY

Random Forest

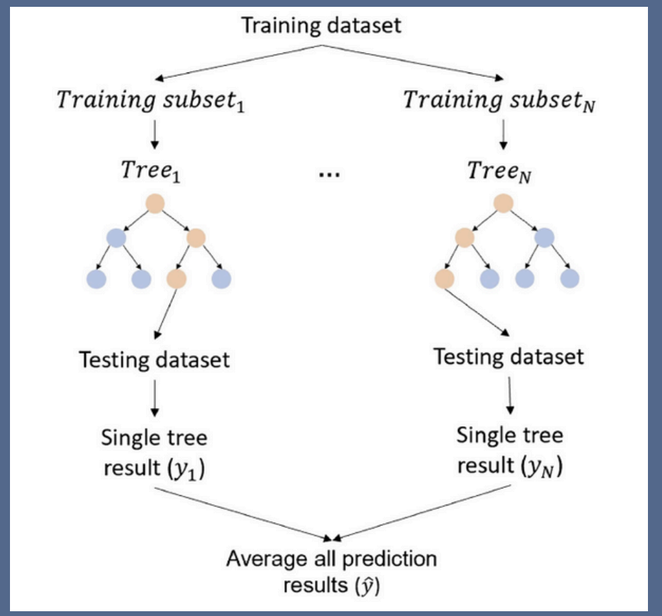
Random Forest is a kind of ensemble model that combines the prediction of multiple decision trees to create a more accurate final prediction and is a verified powerful tool, this model combines the classification and regression tree and bagging method (Breiman, 2001; Truong, 2019).

The random forest algorithm can be summarized in the following steps:

The training dataset is first divided into training subsets, which are chosen at random. Second, using the training subsets, trees are generated randomly and trained. The parent node divides into two daughter nodes, and the resulting information impurity can be expressed as follows:

$$\Delta g(N) = g(N) - P_L g(N_L) - P_R g(N_R)$$

$$\hat{y} = \frac{1}{N_{trees}} \sum_{i=1}^{N_{trees}} y_i$$



(Andariesta & Wasesa, 2022)

Andariesta, D. T., & Wasesa, M. (2022). Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach. Journal of Tourism Futures.

Breiman, L. (2001). Random Forests. SpringerLink.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 174, 433-442.

METHODOLOGY

Stepwise Regression

The stepwise regression algorithm serves as an automated method for statistical model selection when dealing with a substantial number of potential explanatory variables and a lack of underlying theory to guide the selection. While primarily used in regression analysis, the fundamental approach is adaptable to various forms of model selection (Sharma & Yu, 2015).

Sharma MJ, Yu SJ (2015) Stepwise regression data envelopment analysis for variable reduction. Appl Math Comput 253:126–134

MODEL EVALUATION EXPLANATION

01

02

03

$$\text{RMSE (units)} = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}(H_{mi} - H_{ei})^2\right]}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}.$$

$$\text{MAPE} = \frac{\sum^{\frac{|y1 - yt'|}{y1}}}{n} \text{ x } 100\ \%$$

RMSE

The root mean square error (RMSE) is a measure of the differences between predicted values and observed values. It is a frequently used measure of the error of a model in predicting quantitative data. The RMSE is calculated by taking the square root of the average of squared errors. The formula for RMSE is:

MAE

The mean absolute error (MAE) is a metric used to evaluate the performance of a regression model. It is defined as the average absolute difference between the predicted values of the model and the true values of the data

MAPE

The mean absolute percentage error (MAPE) is a metric used to measure the accuracy of predictions in various industries, such as finance and economic forecasts. It is defined as the average absolute percentage difference between predicted values and actual values.

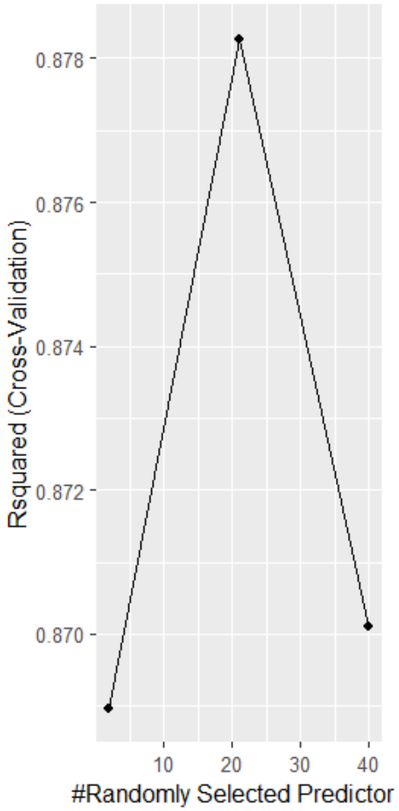
01 **Kambezidis, H. D. (2012). The solar resource. In Elsevier eBooks (pp. 27–84). https://doi.org/10.1016/b978-0-08-087872-0.00302-4**

02 **Pontius, R. G., Thontteh, O., & Chen, H. (2007). Components of information for multiple resolution comparison between maps that share a real variable. Environmental and Ecological Statistics, 15(2), 111–142. https://doi.org/10.1007/s10651-007-0043-y**

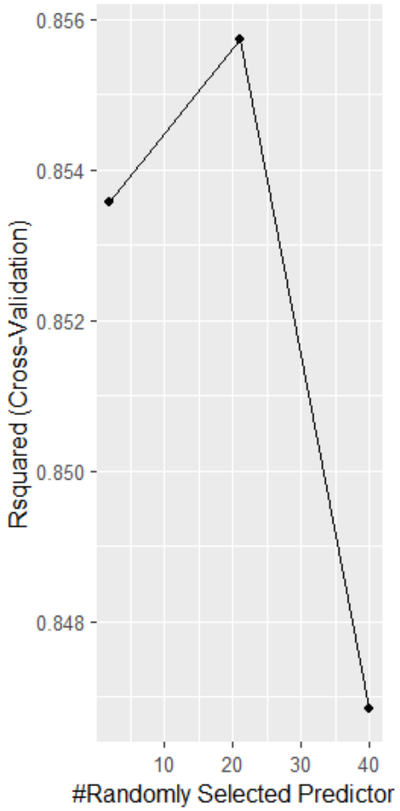
03 **Khair, U., Fahmi, H., Hakim, S. A., & Rahim, R. (2017). Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error. Journal of Physics, 930, 012002. https://doi.org/10.1088/1742-6596/930/1/012002**

Plot RFE (Recursive Feature Elimination)
MODEL RANDOM FOREST

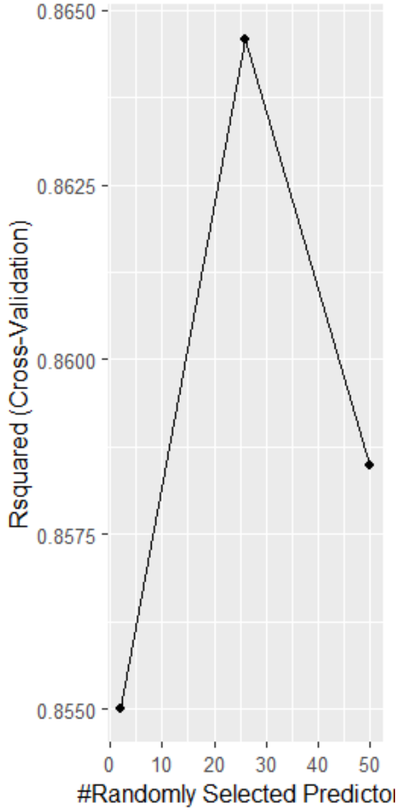
RMSE



01

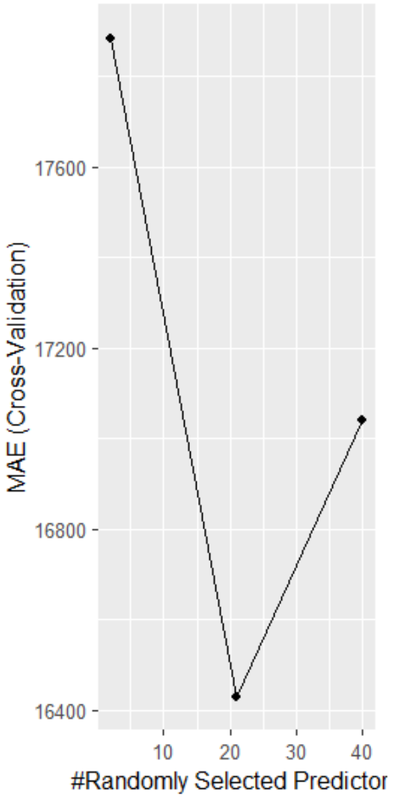


02

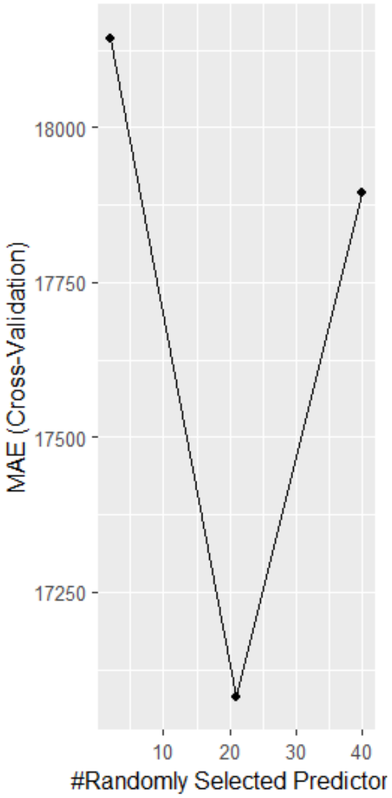


03

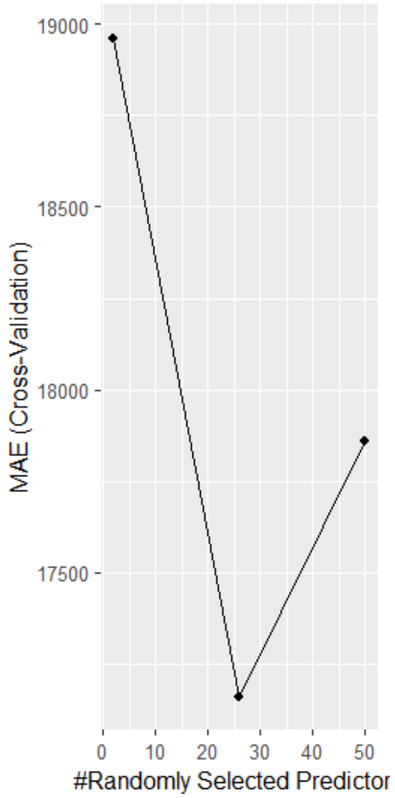
MAE



01



02



03

Dataset/data composition

Dataset/data composition

01

Table Evaluation Model

LM 1 → "Type of Model" — "Dataset/data composition"

	Rsquared	RMSE	MAE	MAPE (%)
LM 1	0,8089	34605,630	20865,032	11,34
LM 2	0,810	34068,216	20840,828	11,35
LM 3	0,789	37755,077	20196,718	44,48
RF 1	**0,878**	**27713,807**	**16427,269**	**3,76**
RF 2	0,856	29140,918	17080,192	44,02
RF 3	0,865	29489,294	17158,335	43,54

LM : Linear Model
RF : Random Forest

02

Table Hyperparameter for Random Forest Model

	Mtry	N Trees	Maximum nodes
RF 1	21	11	10
RF 2	22	11	10
RF 3	23	11	10

From **Table evaluation model,** show the Rsquare, RMSE, MAE, and MAPE of the all model that generated. Color Blue indicate the best performing model across all model. The italics on LM2 indicate the best performing model on linear model using composition dataset 2 (numeric dataset variables + level dataset variables, and several categories dataset)
MAPE of 3% indicates that, on average, the predictions of the model are off by 3% from the actual values.
MAE value provides an easy-to-understand representation of the average prediction error

Conclusion & Implications

CONCLUSION

The conclusion drawn from the table highlights the superior performance of Random Forest (RF) models over Linear Regression (LM) models based on the evaluation metrics of RMSE, MAE, and MAPE. The RF models, represented by RF 1, RF 2, and RF 3, consistently outperform their LM counterparts, LM 1, LM 2, and LM 3, across all three error metrics. The RF models exhibit lower RMSE and MAE values, indicating their predictions are closer to the actual values, resulting in smaller prediction errors. Additionally, the RF models demonstrate lower MAPE percentages, indicating a more accurate representation of the relative prediction errors. These results affirm that the Random Forest approach provides a better fit to the data and yields more accurate predictions compared to the Linear Regression models, making it a preferred choice for modeling and prediction tasks in the given context.

CONCLUSION

The use of a composition of dummy variables with numeric and level data for linear regression models appears to have a negative impact on the quality of the prediction model. As observed in the table, the Linear Regression models (LM 1, LM 2, and LM 3) incorporating dummy variables show lower performance metrics compared to the Random Forest (RF) models.

However, the situation is different when considering the Random Forest models. In the case of RF 2 and RF 3, when comparing their RMSE values, there is an increase in RF 3, implying that the composition of dummy variables in the Random Forest model is more adaptive and beneficial for dummy variables. This suggests that the Random Forest approach can effectively handle the incorporation of dummy variables, leading to improved prediction accuracy and a better representation of the housing price data.

In conclusion, the comprehensive model that includes dummy variables with numeric and level data proves to be less effective for linear regression models but shows better adaptability and performance in the Random Forest models. Therefore, researchers and practitioners should consider using Random Forest with dummy variables for better prediction of housing prices. Further investigation and experimentation may be needed to understand the underlying reasons for the observed differences and to optimize the model's performance.

IMPLICATIONS

FOR RESEARCHERS

- The research would contribute to the advancement of predictive modeling techniques for housing prices, helping researchers explore and refine various methodologies to achieve better accuracy and performance.
-
- By comparing two different modeling approaches, researchers would gain insights into the strengths and limitations of each method, leading to a deeper understanding of the underlying data and the behavior of different variables.

FOR PRACTICIONERS

- Practitioners, such as data analysts and real estate professionals, would gain insights into the best modeling approach to predict housing prices. They can choose between the comprehensive model and the simplified model, based on their specific requirements and resources.

- Understanding the implications of model interpretability and complexity trade-offs would enable practitioners to select models that align with the practical needs of stakeholders and facilitate effective communication of results.

FOR RESEARCHERS

- Accurate price predictions would aid businesses in mitigating risks associated with property investment and development, leading to more informed and data-driven decision-making.

- Adopting the most suitable modeling approach can result in better efficiency and reduced costs for data analytics and housing price prediction tasks.

- The research could lead to a deeper understanding of factors influencing housing prices, enabling businesses to better cater to customer demands and preferences.


THANK YOU