Advanced Regression House Prices

Team 1

Gerard Corrales Fernandez, Sahil Wadhwa, Anahit Shekikyan

Shiley-Marcos School of Engineering, University of San Diego

Master of Science, Applied Data Science

Foundations of Data Science and Data Ethics

Erin Cooke

May 15, 2024

Business Understanding

Background

In the dynamic and complex world of real estate, accurately predicting house prices is not only a challenge but an essential requirement. Traditional valuation methods have often relied on historical data and expert judgment, but the advent of advanced regression techniques and machine learning has revolutionized this process. These modern approaches offer a data-driven means to uncover complex patterns and variables that influence property prices, providing a more detailed and accurate assessment framework.

Real estate markets are influenced by a multitude of factors beyond simple supply and demand dynamics. Location, property characteristics, economic trends, and demographic shifts all play pivotal roles in determining property values. This study explores the predictive power of advanced regression techniques, seeking to analyze these influences and measure their impact on residential property prices.

Drawing from extensive datasets sourced from platforms like Kaggle, which host competitions such as "House Prices: Advanced Regression Techniques," this research explores how machine learning algorithms can effectively predict housing prices based on a rich array of property attributes. These attributes range from basic structural features such as square footage and number of rooms to more detailed factors like condition scores and features.

The methodology includes rigorous data preprocessing techniques to ensure data integrity, exploratory data analysis (EDA) to uncover hidden patterns, and advanced feature engineering to extract meaningful insights. By transforming raw data into actionable intelligence, researchers can build regression models that not only forecast prices accurately but also enhance understanding of the underlying drivers of property values.

The significance of this research extends beyond academic interest. Accurate price predictions are invaluable to various stakeholders in the real estate ecosystem: prospective buyers seeking fair deals, sellers aiming to maximize returns, and investors evaluating market opportunities. By leveraging the power of machine learning, this study aims to empower decision-makers with reliable tools for navigating the complexities of real estate transactions in an increasingly data-driven world.

In conclusion, the fusion of advanced regression techniques with real estate assessment represents a transformative shift in how property prices are determined and understood. As technology continues to advance, so does our ability to refine these models, offering ever-improving accuracy and insights into the multifaceted dynamics of housing markets. This research serves as a testament to the transformative potential of data analytics in shaping the future of real estate assessment and investment strategies.

Business Objectives and Success Criteria

The primary goal is to develop a predictive model that accurately estimates house prices in Ames, Iowa, which will boost buyer confidence. By providing reliable price estimates, buyers can make informed decisions, fostering trust in the listings. Additionally, this model will assist sellers in determining optimal listing prices, ensuring they attract serious buyers and receive fair market value. Real estate agents will also benefit from this tool, as it will improve their ability to provide data-driven advice to clients, improving service quality and client satisfaction.

A successful predictive model will achieve a data, placing it in the top 10% of the Kaggle leaderboard, as measured by Root Mean Squared Error (RMSE) on the test dataset. The model must comply with all relevant real estate regulations to maintain credibility and avoid legal issues. Stakeholder satisfaction is essential, with a target satisfaction rating of 90% or higher from buyers, sellers, and real estate agents in follow-up surveys. The model should be implemented through a user-friendly interface or dashboard, ensuring simplicity of use for all stakeholders. Continuous improvement is essential, with mechanisms in place to gather user feedback and update the model regularly to adapt to market chcanges and evolving user needs.

Educational outreach is important to ensure real estate agents and other stakeholders can effectively utilize the predictive model. Conducting workshops and training sessions will facilitate this. Furthermore, regularly assessing the model's impact on the Ames, Iowa real estate market will help in understanding and enhancing its effectiveness.

Inventory of Resources

The hardware that is being used for this project is a Mac Desktop (iMac) which is used with macOS and has a RAM capacity up to 128 GB. This high RAM capacity ensures enough power to handle large datasets and complex mathematics. The development environment being used is either Juptyer notebooks or VSCode which runs Python 3.14. The versatile VSCode is a great IDE for coding because it offers many extensions and integrations. And google docs is used to share information between all the members for this project.

The data that is being used was collected from the Ames City Assessor's Office with the help of someone with the name of Dean De Cock who collected data of individual residential property in Ames, Iowa. This data ranged between the years of 2006 to 2010. The project team consists of data science students with enough knowledge to do a complete data analysis, feature engineering, and modeling.

Some tools and libraries are being used for efficiency and effectiveness in the project such as pandas, seaborn, matplotlib, scikit-learn, and XGBoost. Git and GitHub are used for version control enabling all the team members to share their code. This project also benefits from Kaggle community which provides tutorials, code, data, and community support. Additionally, other resources are being used such as YouTube, online courses, and research papers. So, using these resources effectively, the team should be able to predict house prices accurately.

Requirements, Assumptions, Constraints, And RESOLVEDD Strategy

The "House Prices - Advanced Regression Techniques" data project from Kaggle aims to predict the final sale price of houses in residential areas in Ames, Iowa, using various variables and factors. The target group for this project includes residential houses in Ames, Iowa. As of April 2024, the cost of living in Ames is 5% lower than the national average. House prices in Ames have increased by 4.6% from the previous year, with a median price of $293,000. The model must be accurate and easy to understand, as incorrect predictions could support buyer confidence and trust. While rapid implementation is not critical, results should be delivered promptly to maintain relevance. The model should be easy to maintain and repeat for future use, with only external factors like changes in the local economy potentially affecting the process. Model performance will be evaluated using metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE). Lower values of these metrics indicate higher prediction accuracy.

There are no significant legal, security, or privacy concerns, as the dataset used is public and does not require protection. However, it is essential to acknowledge the dataset’s licensing terms as provided by Kaggle, ensuring compliance with any specified usage restrictions or attribution requirements.

The project concludes that the dataset is representative of the current housing market in Ames, Iowa, that economic factors affecting the housing market remain relatively stable, and that data quality is acceptable for building an accurate model after appropriate preprocessing.

Challenges include the need to handle missing values appropriately to ensure the quality of the analysis and model, while significant interruptions in deployment can affect the model's relevance and accuracy. Changes in the wider economic environment or local market conditions may impact the model's predictions.

The RESOLVEDD strategy for this project focuses on accurately predicting house prices in Ames, Iowa, using a variety of features. It involves utilizing Exploratory Data Analysis (EDA) to understand the data distribution, identify correlations, and handle missing values. Log transformations are applied to normalize skewed distributions. Relevant features are chosen based on EDA and correlation analysis, with a focus on features like "OverallQual," "GrLivArea," and "TotalBsmtSF," which show strong correlations with "SalePrice." Preprocessing steps include the imputation of missing values, encoding of categorical variables, and scaling of numerical features. The model is trained using regression techniques and evaluated using MSE or RMSE. Project timelines are managed to ensure timely delivery of results, and the model is kept straightforward to facilitate updates and maintenance. Regular assessments of model performance using cross-validation and relevant metrics ensure that the model’s predictions align with current market trends. The model is deployed for predicting house prices, with a process in place for periodic updates to incorporate new data and adjust for changing market conditions. Thorough documentation of the data preprocessing steps, model training process, and evaluation metrics ensures consistency and transparency. Based on model performance and evaluation, a final deployment strategy is decided, with continuous monitoring and improvement to maintain accuracy.

Risks And Contingencies

As we start with the House Price Prediction Project with the Ames Housing dataset, it is very important to know the importance of possible risks to mitigate them. Some factors that might influence the outcome of this data mining project are data quality issues, and modeling difficulties. For data quality issues the dataset contains some “NaN” values indicating the absence of some features and will be handled by substituting “No” to ensure proper classification into categorical columns. Also, the numerical column contains missing values, and they will be substituted by the mean. If new data is added in the future, similar preprocessing steps will be applied to ensure consistency. In modeling challenges, exists the potential risk of overfitting the model due to the size of the dataset and the complexity of the features. To minimize the risk, methods such as cross-validation and pruning decision trees should be used but these methods are out of the scope of this research. In addition, using solutions like Random Forest or Gradient Boosting would improve the prediction model but this research will focus only on using linear regression as a predicting method.

Other external factors that could influence future real estate prices are economic shifts, changes in housing market dynamic, or policy changes. These factors will not influence our dataset, but it is important to know that our model might not work properly in the current year. To prevent this, incorporating more data about market trends and economic indicators could potentially make our model more robust for future predictions.

Since this is a real estate dataset showing only numbers and categories based on one location, no potential biases could have been identified. It is crucial to keep monitoring for biases that could affect our model fairness. To make sure that our model is unbiased, tools such as fairness-aware machine learning and bias detection should be employed to reduce any favoritism.

Terminology

In this section, many vocabulary words are introduced encompassing various aspects of the real estate market, statistics, and data science techniques to have a better understanding of this project.

RMSE (Root Mean Squared Error): The square root of the average of the squared differences between the predicted values and the actual values.

- **Training RMSE:** 0.1519 - Indicates the model's prediction error on the training dataset.
- **Validation RMSE:** 0.1582 - Indicates the model's prediction error on the validation dataset.

R^2 (R-squared or Coefficient of Determination): A statistical measure representing the proportion of the variance for a dependent variable explained by an independent variable or variables in a regression model.

- **Training R^2:** 0.8486 - Reflects how well the independent variables explain the variability of the dependent variable for the training data.
- **Validation R^2:** 0.8658 - Reflects how well the independent variables explain the variability of the dependent variable for the validation data.

SalePrice Predictions:

- **Id:** A unique identifier for each property.
- **SalePrice:** The model's predicted sale price for the property.

The 'SalePrice' predictions are the output of the model after being trained and validated with the respective RMSE and R^2 metrics, providing an estimate of property values.

Data Mining Goals and Success Criteria

In this data mining project, the primary goal is to predict the sales price for each house in the data set, focusing on the SalePrice variable. This challenge involves analyzing the dataset from Ames, Iowa, with 79 variables that describe characteristics of residential homes. These variables have very specific details like white-picket fence, or alley that they are not obvious, but it shows the complexity of negotiations in the real estate market. This project will be measured using the Root-Mean-Squared-Error (RMSE) between the predicted and observed sale prices, ensuring that errors in predicting both expensive and cheaper homes are weighted equally. To achieve this, will be employed a linear regression machine learning algorithm such as the Ordinary Least Squares (OLS) (e.g. sm.OLS().fit()). For data mining, the selected model with the highest R^2 value will be used to predict the target variable, the sale price. A high R^2 value in the prediction model indicates that a significant proportion of the variation in prices is explained by the descriptive features included in the model (Devore, 2016). Using reliable computer systems in data mining project will provide with enough advanced information and knowledge to properly estimate sale prices in real estate.

As Kelleher (2020) mentioned, measurements like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are metrics to evaluate the performance of models. And this improvement in model performance leads to increased accuracy in predicting the exact sale prices, which is a crucial aspect of the modeling process. To evaluate the efficacy, several prediction models will be used in predicting the goal variable. After analyzing all descriptive features, the data mining team will select the model with the greatest R^2 value and lowest error. To extract insights from data, multidimensional visualizations will be used and shared with other team members for feedback. Strategic modifications to the models are anticipated to increase the R^2 value.

Data Mining Goals	Data Mining Success Criteria
1. Factors that affect sale prices are identified using data analysis techniques.	• The team gathers all the relevant data for the study from Kaggle website and stores it in Google drive. • The database is processed using python for statistics to have a preliminary understanding of the data. • Handling categorical features using One-Hot Encoding. • Feature engineering is needed to create new columns to optimize calculations. • Some visualizations are created to see the relationship between the

	dependent variable and independent variables.
2. Create several models and select those with better performance	- A heatmap shows some high correlations between features, but the rest would have to be removed. - Select features with significant correlation with the target variable. - Using all the features the R^2 in the prediction model is 0.933. - Using many methods of error MSE, RMSE, and MAE to evaluate performance. - Selecting the best performance model with the highest R^2.
3. Understanding the best model in reproducing data	- Use the model to make predictions displaying the data frame with the results - Comparing the results with the actual data.
4. Sharing the results with the team	- Evaluating the results with the team to keep improving the prediction model until it meets the goal.

5. Model deployment	• All team members agreed on the same model. • The model predicts successfully the outcome

Project Plan

This project is structured in five distinct phases and it is represented in Figure 1. The first phase involves accessing, collecting, processing, and visualizing the Housing dataset. The team will do an initial data exploration to understand the dataset's structure and quality, identify missing values, and generate preliminary visualizations. After that, some goals would have to be modified, risks and contingencies will be drafted, and necessary resources will be collected. Data visualization will be used to create different types of graphs using Python, and raw data will be analyzed carefully for completeness and accuracy. This phase aims to be completed in about 2 weeks.

During phase 2, the team will start exploratory data analysis (EDA) evaluating the criteria to track the project's business impact, constraints, and define success goals. Understanding more about what influences house prices and other possible techniques to apply in EDA to identify significant variables and features. This phase will last 2 weeks.

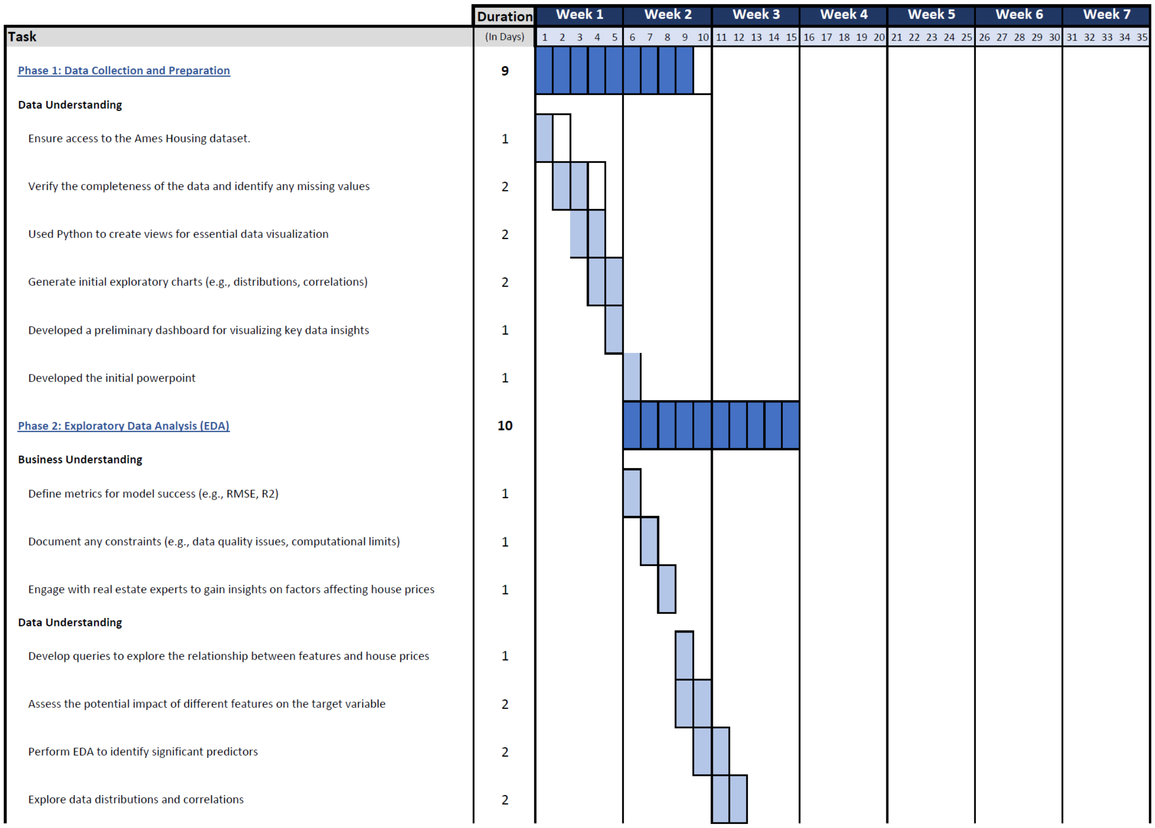
Phase 3 requires developing feature engineering and selection, where new features will be created based on EDA findings. Features that are relevant will be selected for modeling, and the final dataset version will be ready. This phase is expected to take 2 weeks.

Phase 4, modeling and validation will be executed. Preparing initial models to evaluate the performance with the data. So, the highest R-squared models will be selected for the final step. This phase will take approximately 2 weeks.

In the final phase 5, the model is ready to be deployed in the production environment, present it to class how to use it, and be ready to keep track of the model performance. Some regular updates should be needed to improve the model's performance. This phase will last 2 weeks.

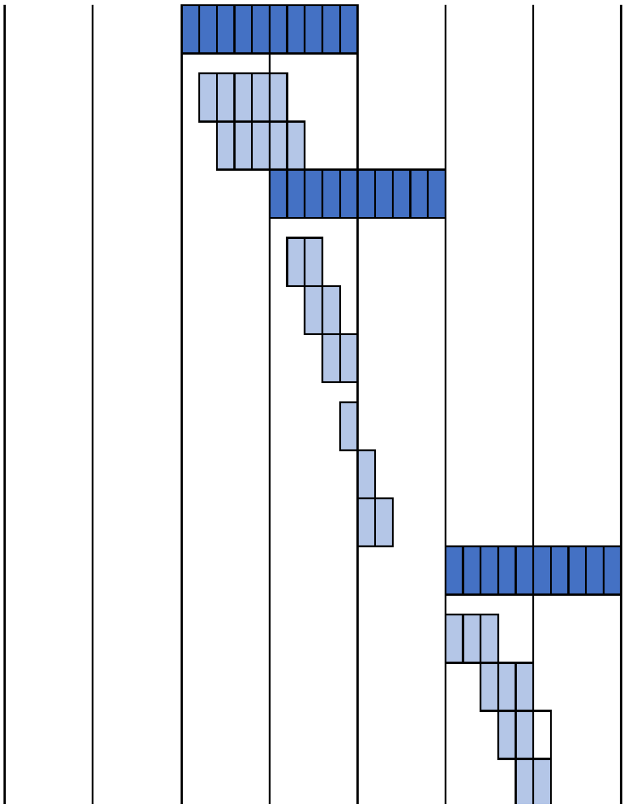
Figure 1

Project Plan



Task	Duration (In Days)
Phase 1: Data Collection and Preparation	9
Data Understanding	
Ensure access to the Ames Housing dataset.	1
Verify the completeness of the data and identify any missing values	2
Used Python to create views for essential data visualization	2
Generate initial exploratory charts (e.g., distributions, correlations)	2
Developed a preliminary dashboard for visualizing key data insights	1
Developed the initial powerpoint	1
Phase 2: Exploratory Data Analysis (EDA)	10
Business Understanding	
Define metrics for model success (e.g., RMSE, R2)	1
Document any constraints (e.g., data quality issues, computational limits)	1
Engage with real estate experts to gain insights on factors affecting house prices	1
Data Understanding	
Develop queries to explore the relationship between features and house prices	1
Assess the potential impact of different features on the target variable	2
Perform EDA to identify significant predictors	2
Explore data distributions and correlations	2

Phase 3: Feature Engineering and Selection	**10**
Data Preparation	
Engineer new features based on domain knowledge	5
Select relevant features for modeling and create the final dataset	5
Phase 4: Model Development and Validation	**10**
Modeling	
Develop and train initial baseline models (linear regression)	2
Generate predictions and prepare the output dataset for visualization	2
Improving models through validation	2
Evaluation	
Assess model performance using validation metrics	1
Share results with stakeholders and gather feedback	1
Iterate on the model based on feedback	2
Phase 5: Model Deployment and Monitoring	**10**
Deployment	
Implement the final model in the production environment	3
Updated powerpoint	3
Tracking model performance	2
Schedule regular updates	2



Data Understanding

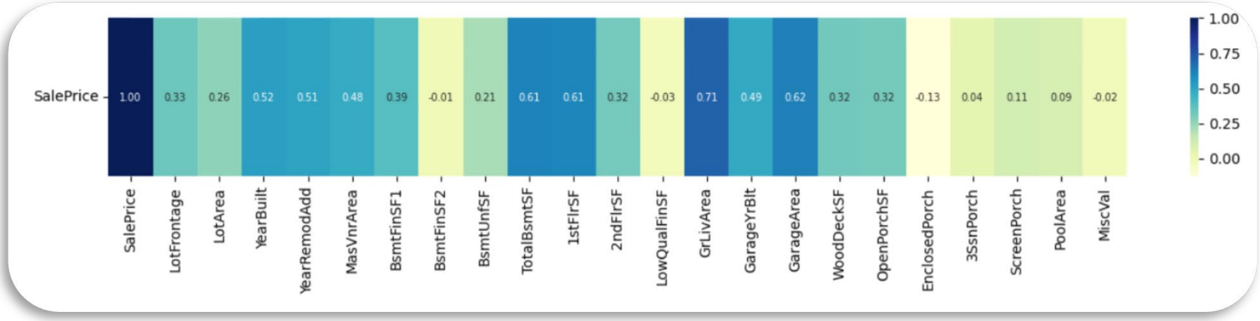
Initial Data Collection Report

The team focused on studying the data from “House Prices” dataset that can be found at Kaggle website where they will collect, describe, and explore data to understand which trends and patterns exist in the data. The business objective is how to predict house prices using the target variable “SalePrice”. The data was imported from csv files downloaded from Kaggle and Python was used to perform tasks. The database contains 81 descriptive features, totaling approximately 460 kB. The initial observations revealed several missing values in features such as ‘LotFrontage’, ‘Alley’, ‘MasVnrType’, ‘MasVnrArea’, ‘BsmtQual’, ‘BsmtCond’, ‘BsmtExposure’, ‘BsmtFinType1’, ‘BsmtFinType2’, ‘Electrical’, ‘FireplaceQu’, ‘GarageType’, ‘GarageYrBlt’, ‘GarageFinish’, ‘GarageQual’, ‘GarageCond’, ‘PoolQC’, ‘Fence’, and ‘MiscFeature’. Only 16 columns contained empty values indicating that they do not have the specific attribute, so data missing imputation was used to assign labels as “no”. No duplicate records were found in the dataset. Understanding how these data attributes interact between them to affect the sale price will be very useful to support decision-making.

As Devore (2016) mentions, features in the dataset with strong correlations above or equal to |70%| are needed to predict our dependent variable. And those attributes with correlations less or equal to |30%| must be excluded from our models. But at the same time considering not to eliminate too many features to not lose essential information (Kelleher, 2020). Therefore, according to Figure 2 many variables are in negative or below |30%| correlation, so they will be dropped as Devore mentioned.

Figure 2

Heatmap Showing Correlations Between the Target Variable and Linear Regression Variables



Data Description Report

This report provides a detailed description of the house pricing dataset used in the House Prices - Advanced Regression Techniques project. The dataset, sourced from Kaggle, includes various features that describe the physical characteristics and sale conditions of residential properties in Ames, Iowa. The goal is to predict the final sale price of each property based on these features. The dataset is provided in CSV format, which allows for easy manipulation and analysis using various data analysis tools.

Let's overview the dataset, it contains the following features:

MSSubClass: Identifies the type of dwelling involved in the sale.

- **20:** 1-STORY 1946 & NEWER ALL STYLES
- **30:** 1-STORY 1945 & OLDER
- **40:** 1-STORY W/FINISHED ATTIC ALL AGES
- **45:** 1-1/2 STORY - UNFINISHED ALL AGES

- **50:** 1-1/2 STORY FINISHED ALL AGES
- **60:** 2-STORY 1946 & NEWER
- **70:** 2-STORY 1945 & OLDER
- **75:** 2-1/2 STORY ALL AGES
- **80:** SPLIT OR MULTI-LEVEL
- **85:** SPLIT FOYER
- **90:** DUPLEX - ALL STYLES AND AGES
- **120:** 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
- **150:** 1-1/2 STORY PUD - ALL AGES
- **160:** 2-STORY PUD - 1946 & NEWER
- **180:** PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
- **190:** 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

- **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

LotFrontage: Linear feet of street connected to property.

LotArea: Lot size in square feet.

Street: Type of road access to property.

- **Grvl:** Gravel
- **Pave:** Paved

Alley: Type of alley access to property.

- **Grvl:** Gravel
- **Pave:** Paved
- **NA:** No alley access

LotShape: General shape of property.

- **Reg:** Regular
- **IR1:** Slightly irregular
- **IR2:** Moderately Irregular
- **IR3:** Irregular

LandContour: Flatness of the property.

- **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

Utilities: Type of utilities available.

- **AllPub:** All public Utilities (E, G, W, & S)
- **NoSewr:** Electricity, Gas, and Water (Septic Tank)
- **NoSeWa:** Electricity and Gas Only
- **ELO:** Electricity only

LotConfig: Lot configuration.

- **Inside:** Inside lot
- **Corner:** Corner lot
- **CulDSac:** Cul-de-sac
- **FR2:** Frontage on 2 sides of property
- **FR3:** Frontage on 3 sides of property

LandSlope: Slope of property.

- **Gtl:** Gentle slope
- **Mod:** Moderate Slope

- **Sev:** Severe Slope

Neighborhood: Physical locations within Ames city limits.

- **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
- **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

Condition1: Proximity to various conditions.

- **Artery:** Adjacent to arterial street
- **Feedr:** Adjacent to feeder street
- **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 positive off-site feature
- **RRNe:** Within 200' of East-West Railroad

- **RRAe:** Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present).

- **Artery:** Adjacent to arterial street
- **Feedr:** Adjacent to feeder street
- **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 positive off-site feature
- **RRNe:** Within 200' of East-West Railroad
- **RRAe:** Adjacent to East-West Railroad

BldgType: Type of dwelling.

- **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: Style of dwelling.

- **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

OverallQual: Rates the overall material and finish of the house.

- **10:** Very Excellent
- **9:** Excellent
- **8:** Very Good
- **7:** Good
- **6:** Above Average
- **5:** Average
- **4:** Below Average
- **3:** Fair
- **2:** Poor

- **1:** Very Poor

OverallCond: Rates the overall condition of the house.

- **10:** Very Excellent
- **9:** Excellent
- **8:** Very Good
- **7:** Good
- **6:** Above Average
- **5:** Average
- **4:** Below Average
- **3:** Fair
- **2:** Poor
- **1:** Very Poor

YearBuilt: Original construction date.

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions).

RoofStyle: Type of roof.

- **Flat:** Flat
- **Gable:** Gable
- **Gambrel:** Gabrel (Barn)

- **Hip:** Hip
- **Mansard:** Mansard
- **Shed:** Shed

RoofMatl: Roof material.

- **ClyTile:** Clay or Tile
- **CompShg:** Standard (Composite) Shingle
- **Membran:** Membrane
- **Metal:** Metal
- **Roll:** Roll
- **Tar&Grv:** Gravel & Tar
- **WdShake:** Wood Shakes
- **WdShngl:** Wood Shingles

Exterior1st: Exterior covering on house.

- **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
- **Wd Sdng:** Wood Siding
- **WdShing:** Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material).

- **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
- **Wd Sdng:** Wood Siding
- **WdShing:** Wood Shingles

MasVnrType: Masonry veneer type.

- **BrkCmn:** Brick Common
- **BrkFace:** Brick Face
- **CBlock:** Cinder Block
- **None:** None
- **Stone:** Stone

MasVnrArea: Masonry veneer area in square feet.

ExterQual: Evaluates the quality of the material on the exterior.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Average/Typical
- **Fa:** Fair
- **Po:** Poor

ExterCond: Evaluates the present condition of the material on the exterior.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Average/Typical
- **Fa:** Fair
- **Po:** Poor

Foundation: Type of foundation.

- **BrkTil:** Brick & Tile
- **CBlock:** Cinder Block
- **PConc:** Poured Contrete
- **Slab:** Slab

- **Stone:** Stone
- **Wood:** Wood

BsmtQual: Evaluates the height of the basement.

- **Ex:** Excellent (100+ inches)
- **Gd:** Good (90-99 inches)
- **TA:** Typical (80-89 inches)
- **Fa:** Fair (70-79 inches)
- **Po:** Poor (<70 inches)
- **NA:** No Basement

BsmtCond: Evaluates the general condition of the basement.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Typical - slight dampness allowed
- **Fa:** Fair - dampness or some cracking or settling
- **Po:** Poor - Severe cracking, settling, or wetness
- **NA:** No Basement

BsmtExposure: Refers to walkout or garden level walls.

- **Gd:** Good Exposure

- **Av:** Average Exposure (split levels or foyers typically score average or above)
- **Mn:** Minimum Exposure
- **No:** No Exposure
- **NA:** No Basement

BsmtFinType1: Rating of basement finished area.

- **GLQ:** Good Living Quarters
- **ALQ:** Average Living Quarters
- **BLQ:** Below Average Living Quarters
- **Rec:** Average Rec Room
- **LwQ:** Low Quality
- **Unf:** Unfinshed
- **NA:** No Basement

BsmtFinSF1: Type 1 finished square feet.

BsmtFinType2: Rating of basement finished area (if multiple types).

- **GLQ:** Good Living Quarters
- **ALQ:** Average Living Quarters
- **BLQ:** Below Average Living Quarters
- **Rec:** Average Rec Room

- **LwQ:** Low Quality
- **Unf:** Unfinshed
- **NA:** No Basement

BsmtFinSF2: Type 2 finished square feet.

BsmtUnfSF: Unfinished square feet of basement area.

TotalBsmtSF: Total square feet of basement area.

Heating: Type of 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

HeatingQC: Heating quality and condition.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Average/Typical
- **Fa:** Fair

- **Po:** Poor

CentralAir: Central air conditioning.

- **N:** No
- **Y:** Yes

Electrical: Electrical system.

- **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

1stFlrSF: First Floor square feet.

2ndFlrSF: Second floor square feet.

LowQualFinSF: Low quality finished square feet (all floors).

GrLivArea: Above grade (ground) living area square feet.

BsmtFullBath: Basement full bathrooms.

BsmtHalfBath: Basement half bathrooms.

FullBath: Full bathrooms above grade.

HalfBath: Half baths above grade.

BedroomAbvGr: Bedrooms above grade (does NOT include basement bedrooms).

KitchenAbvGr: Kitchens above grade.

KitchenQual: Kitchen quality.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Typical/Average
- **Fa:** Fair
- **Po:** Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms).

Functional: Home functionality rating.

- **Typ:** Typical Functionality
- **Min1:** Minor Deductions 1
- **Min2:** Minor Deductions 2
- **Mod:** Moderate Deductions
- **Maj1:** Major Deductions 1
- **Maj2:** Major Deductions 2
- **Sev:** Severely Damaged
- **Sal:** Salvage only

Fireplaces: Number of fireplaces.

FireplaceQu: Fireplace quality.

- **Ex:** Excellent - Exceptional Masonry Fireplace
- **Gd:** Good - Masonry Fireplace in main level
- **TA:** Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
- **Fa:** Fair - Prefabricated Fireplace in basement
- **Po:** Poor - Ben Franklin Stove
- **NA:** No Fireplace

GarageType: Garage location.

- **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

GarageYrBlt: Year garage was built.

GarageFinish: Interior finish of the garage.

- **Fin:** Finished
- **RFn:** Rough Finished
- **Unf:** Unfinished
- **NA:** No Garage

GarageCars: Size of garage in car capacity.

GarageArea: Size of garage in square feet.

GarageQual: Garage quality.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Typical/Average
- **Fa:** Fair
- **Po:** Poor
- **NA:** No Garage

GarageCond: Garage condition.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Typical/Average
- **Fa:** Fair

- **Po:** Poor
- **NA:** No Garage

PavedDrive: Paved driveway.

- **Y:** Paved
- **P:** Partial Pavement
- **N:** Dirt/Gravel

WoodDeckSF: Wood deck area in square feet.

OpenPorchSF: Open porch area in square feet.

EnclosedPorch: Enclosed porch area in square feet.

3SsnPorch: Three season porch area in square feet.

ScreenPorch: Screen porch area in square feet.

PoolArea: Pool area in square feet.

PoolQC: Pool quality.

- **Ex:** Excellent
- **Gd:** Good
- **TA:** Average/Typical
- **Fa:** Fair

Fence: Fence quality.

- **GdPrv:** Good Privacy
- **MnPrv:** Minimum Privacy
- **GdWo:** Good Wood
- **MnWw:** Minimum Wood/Wire
- **NA:** No Fence
- **NA:** No Pool

MiscFeature: Miscellaneous feature not covered in other categories.

- **Elev:** Elevator
- **Gar2:** 2nd Garage (if not described in garage section)
- **Othr:** Other
- **Shed:** Shed (over 100 SF)
- **TenC:** Tennis Court
- **NA:** None

MiscVal: $Value of miscellaneous feature.

MoSold: Month Sold (MM).

YrSold: Year Sold (YYYY).

SaleType: Type of sale.

- **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: Condition of sale.

- **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)

The dataset includes 80 features that cover various aspects of the residential properties, including physical characteristics, quality, and sale conditions. Each feature provides valuable information that can be used to predict the sale price of the properties. This comprehensive dataset allows for detailed analysis and modeling, which is crucial for developing accurate predictive models in real estate.

Data Exploration Report

The "House Prices - Advanced Regression Techniques" dataset from Kaggle.com is a comprehensive dataset used to predict the final sale price of homes. It includes numerous features describing different aspects of residential properties in Ames, Iowa, making it an excellent candidate for exploring advanced regression techniques.

The dataset comprises two primary files: train.csv and test.csv. The train.csv file contains 1,460 records with 81 columns, including the target variable “SalePrice”, while the test.csv file contains 1,459 records with the same 80 features, but without the “SalePrice” column. This mix of numerical and categorical features is essential for building predictive models.

Some key features in the dataset are “LotArea”, “OverallQual”, “YearBuilt”, “TotalBsmtSF”, “GrLivArea”, “FullBath”, “GarageCars”, and the target variable “SalePrice”. These features provide a diverse set of attributes that capture various aspects of the properties, from lot size and construction quality to living area and garage capacity.

To ensure high-quality subsequent analysis or machine learning models, several preprocessing steps were taken. Missing values for numerical variables were imputed with the mean or median if the proportion of missing values was manageable. For categorical columns, the mode or a placeholder indicating the absence of a feature was used. Columns with a very high percentage of missing values were dropped if considered unimportant. Categorical variables were then converted into numerical format using techniques like one-hot encoding.

Numerical features were scaled using standardization or normalization techniques to ensure they had a mean of 0 and a standard deviation of 1. Additionally, the “SalePrice” variable was log-transformed to normalize its distribution, which is beneficial for many statistical analyses and machine learning algorithms that assume normality.

Exploratory Data Analysis (EDA) revealed several important insights. The distribution of “SalePrice” (see Figure 3) showed that most houses are sold within the $100,000 to $200,000 price range, with a peak around $150,000. The distribution is right-skewed, indicating fewer sales as the price increases, with some outliers representing high-value properties.

A heatmap (see Figure 4) was created to visualize relationships among all variables, revealing strong positive correlations of features like “OverallQual”, “GrLivArea”, and “TotalBsmtSF” with “SalePrice”. A scatter plot (see Figure 5) of “GrLivArea” against “SalePrice” showed a clear positive relationship, indicating larger living areas tend to command higher prices. A box plot (see Figure 6) of “OverallQual” against “SalePrice” demonstrated higher overall quality ratings correlate with increased sale prices. Similarly, a scatter plot (see Figure 7) of “YearBuilt” against “SalePrice” illustrated newer homes typically fetch higher prices, though some older homes also command high prices due to renovations or historical value. For a comprehensive assessment of each feature’s relationship with the target variable across the dataset, a scatter plot matrix was created (see Figure 9), enabling the identification of potential patterns.

Applying a logarithmic transformation to the “SalePrice” variable resulted in a more bell-shaped distribution (see Figure 8), reducing skewness and making the data more suitable for modeling. This transformation improved the normality of the data, which is beneficial for many statistical analyses and machine learning algorithms.

The summary statistics of the log-transformed “SalePrice” (see Figure 10) are as follows: the mean is 12.024, the median is 12.001, the range is 3.074, the variance is 0.159, the standard deviation is 0.399, the skewness is 0.121, and the kurtosis is 0.810. The first quartile (Q1) is 11.775, the third quartile (Q3) is 12.274, and the interquartile range (IQR) is 0.499.

Visualizations created during the EDA include histograms of the “SalePrice” distribution and its log-transformed counterpart, a correlation heatmap, scatter plots of “GrLivArea” vs. “SalePrice”, and box plots of “OverallQual” vs. “SalePrice”. These visualizations help to understand the relationships between different features and the target variable, guiding further analysis and model development.

Initial data exploration indicates that features like “OverallQual”, “GrLivArea”, and “TotalBsmtSF” are strong predictors of “SalePrice”. Addressing missing values and transforming variables improves the dataset's suitability for predictive modeling. Future steps include further feature engineering, model training, and validation to develop a robust predictive model for house prices. This report provides a foundational understanding of the dataset, guiding further analysis and model development.

Data Quality Report

The "House Prices - Advanced Regression Techniques" dataset from Kaggle consists of 2,919 observations and 80 explanatory variables used to predict house prices (SalePrice). This dataset captures various aspects of residential properties in Ames, Iowa, including 24 ordinal, 23 nominal, 14 discrete, and 19 continuous variables.

A significant part of ensuring data quality involves addressing missing values. The dataset contains 16 variables with missing values. For categorical variables, missing values are replaced with "No," indicating the absence of a feature. For numerical variables, mean imputation is employed, replacing missing values with the mean of the respective variable. This approach maintains the dataset’s statistical properties without introducing additional variability.

Exploratory Data Analysis (EDA) reveals several key insights into the dataset. The target variable “SalePrice” exhibits a right-skewed distribution. To normalize this, a log transformation is applied, resulting in a more bell-shaped distribution and improving the suitability of the data for regression modeling. A correlation analysis using a heatmap identifies strong relationships between certain variables and “SalePrice”. For instance, variables like “OverallQual”, “GrLivArea”, and “TotalBsmtSF” show high positive correlations with “SalePrice”, indicating their predictive value.

Outliers in “SalePrice” are identified through visualizations such as scatter plots and box plots. These outliers represent properties sold at significantly higher prices than the majority, possibly due to unique features or attractive locations. While outliers can skew analysis, they also provide valuable insights and should be carefully considered.

To improve data quality and usability, several transformations are performed. Numerical features are standardized to have a mean of 0 and a standard deviation of 1, ensuring that all features contribute equally to the analysis and improving the performance of many machine learning algorithms. Categorical variables are converted into numerical format using one-hot encoding, creating binary columns for each category and allowing algorithms to process categorical data effectively. Additionally, the “SalePrice” variable is log-transformed to reduce skewness and make its distribution more normal. This transformation aids in meeting the assumptions of many statistical models and improves the robustness of predictions.

Summary statistics of the log-transformed “SalePrice” show a mean of 12.024, a median of 12.002, a range of 3.074, a variance of 0.159, a standard deviation of 0.399, a skewness of 0.121, a kurtosis of 0.810, a first quartile (Q1) of 11.775, a third quartile (Q3) of 12.274, and an interquartile range (IQR) of 0.499. These statistics indicate that the log transformation has successfully normalized the distribution of “SalePrice”, making it more suitable for analysis.

Several visualizations are employed to explore and validate the data, including histograms to visualize the distribution of “SalePrice” and its log-transformed version, heatmaps to examine correlations between variables, scatter plots to investigate relationships between key variables (such as “YearBuilt” vs. “SalePrice”), and box plots to compare “SalePrice” across different neighborhoods or quality ratings.

In conclusion, the dataset is well-structured with a variety of features providing comprehensive information about the properties. The missing values are effectively handled, and appropriate transformations are applied to ensure data quality. The insights from EDA and the visualizations will inform the development of robust predictive models for house prices.

Figure 3

Histogram Providing a Summary of the 'SalePrice' Distribution and Overall Housing Market Trends.

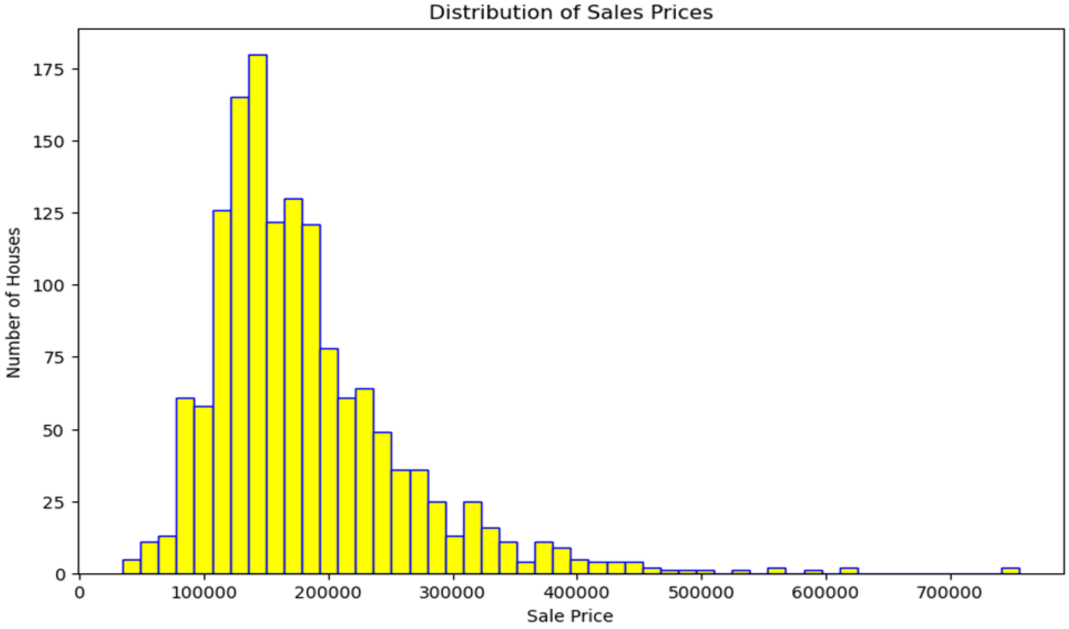


Figure 4

Heatmap providing correlation between the variables

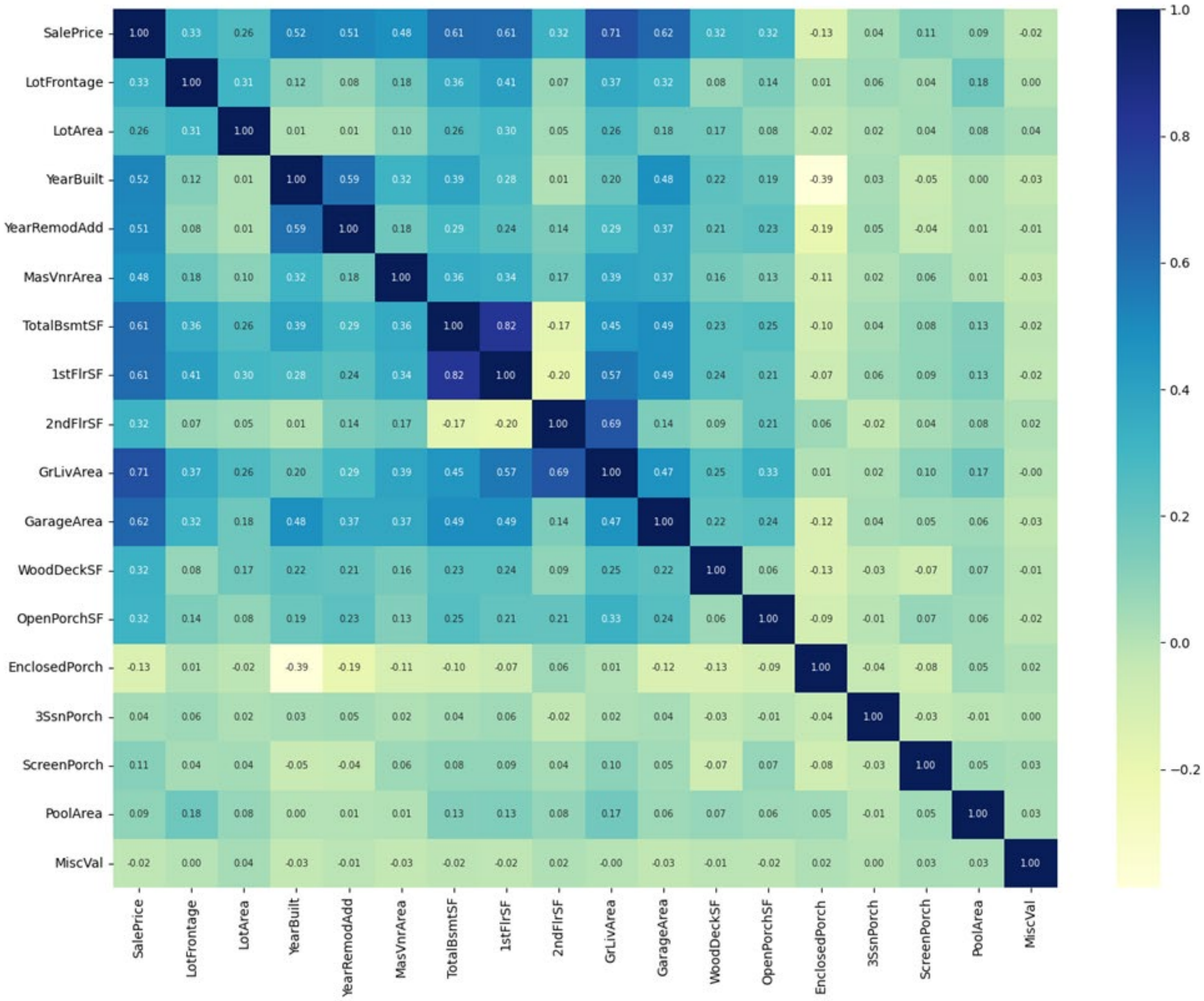


Figure 5

Provides correlation between "SalePrice" and "GrLivArea" and calculates how the size of the living area might affect the sale price of a house

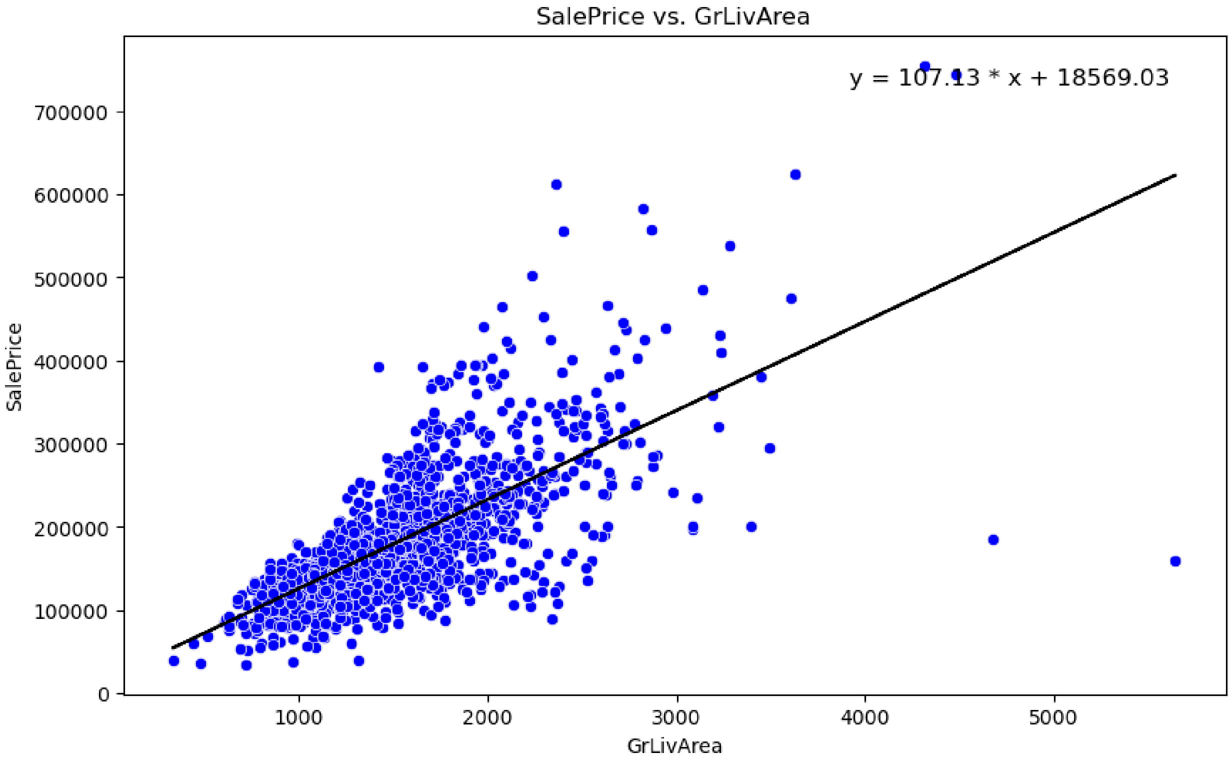


Figure 6

Boxplots of "OverallQual" against "SalePrice"

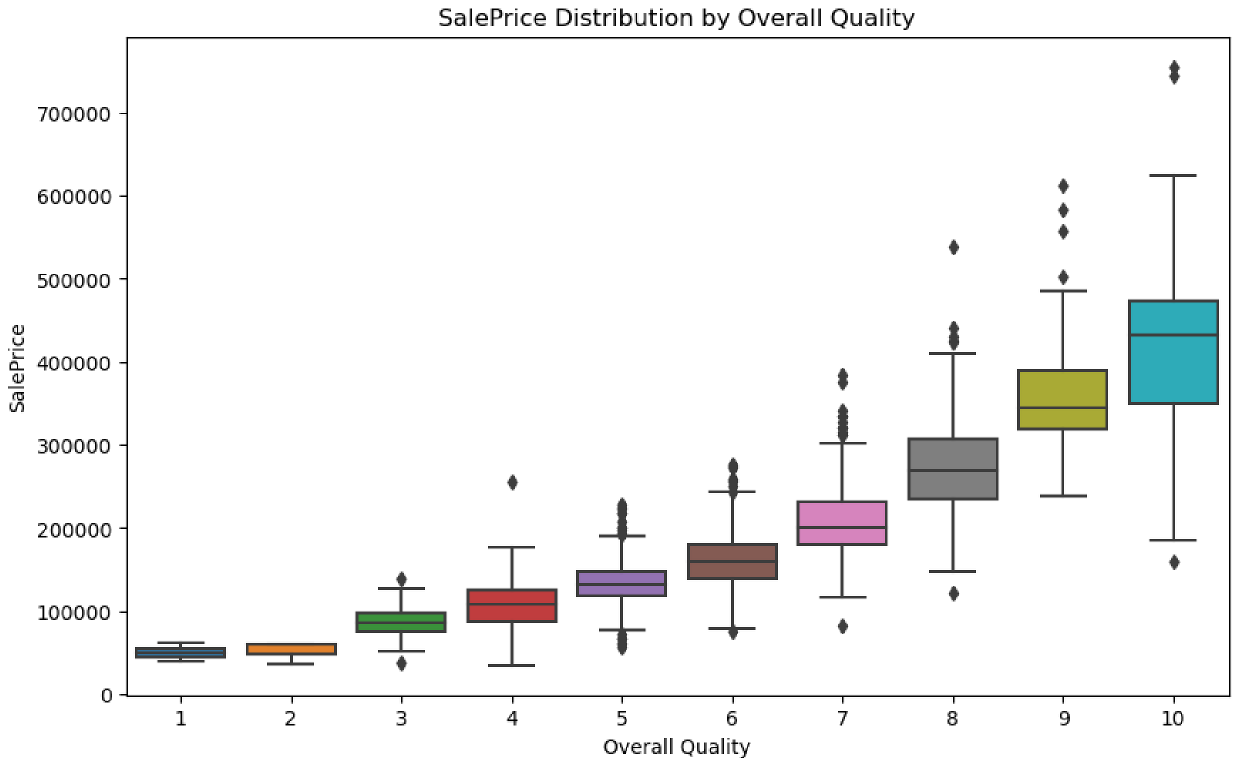


Figure 7

Correlation of "YearBuilt" against "SalePrice

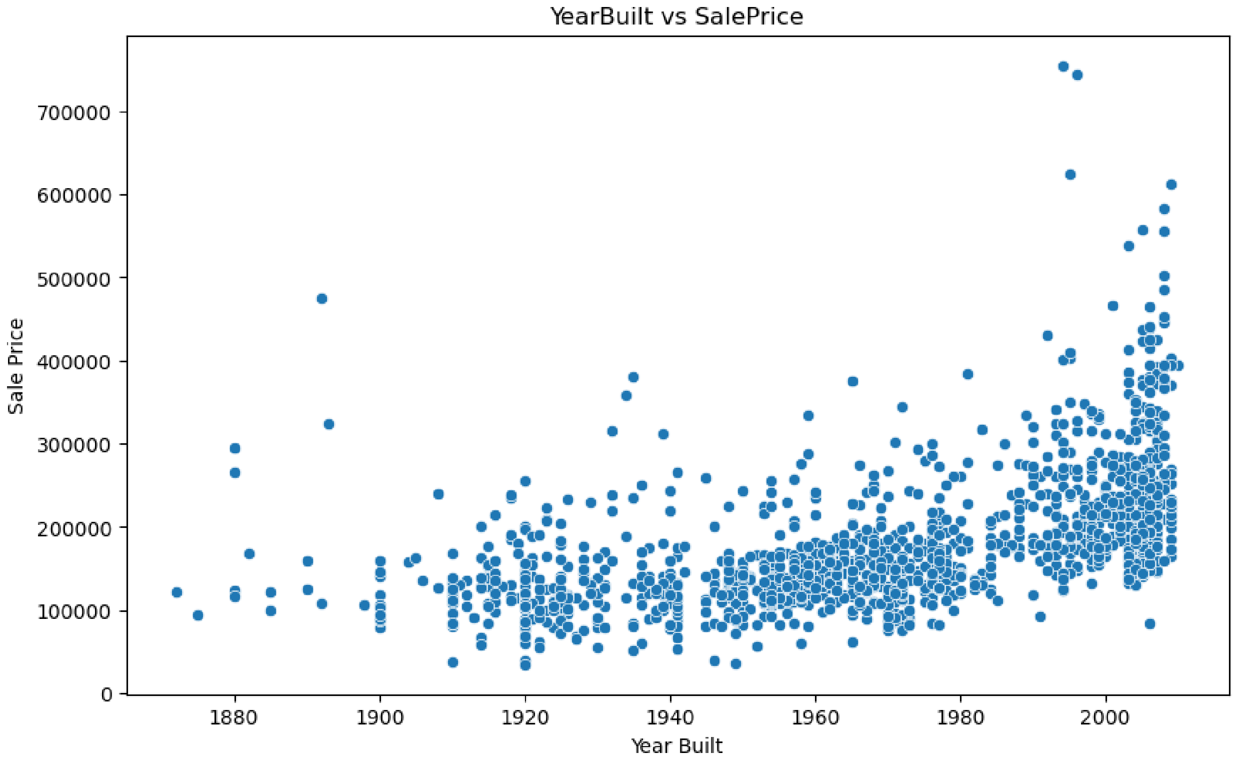


Figure 8

logarithmic Transformed Distribution of Sale Price

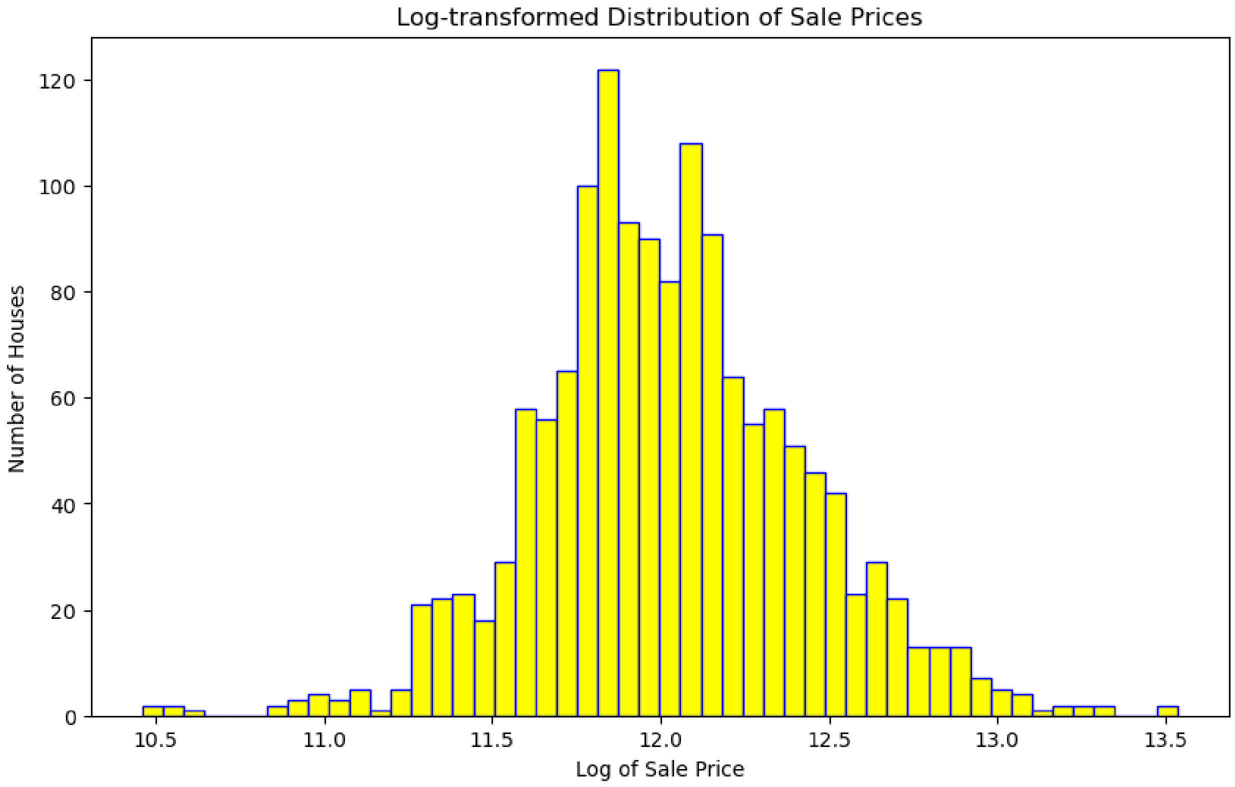


Figure 9

Scatterplot matrix with each subplot showing a selected feature versus 'SalePrice'

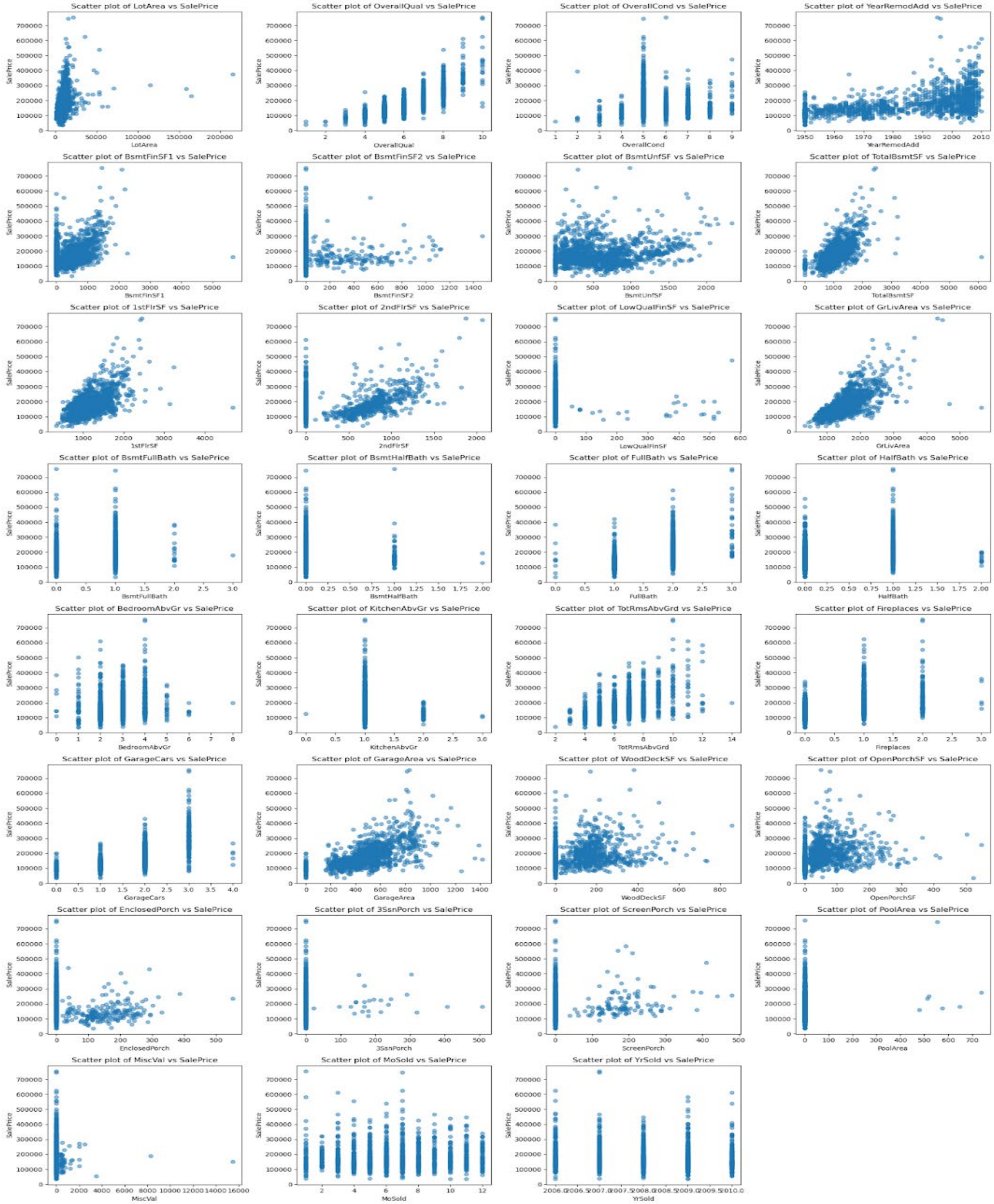
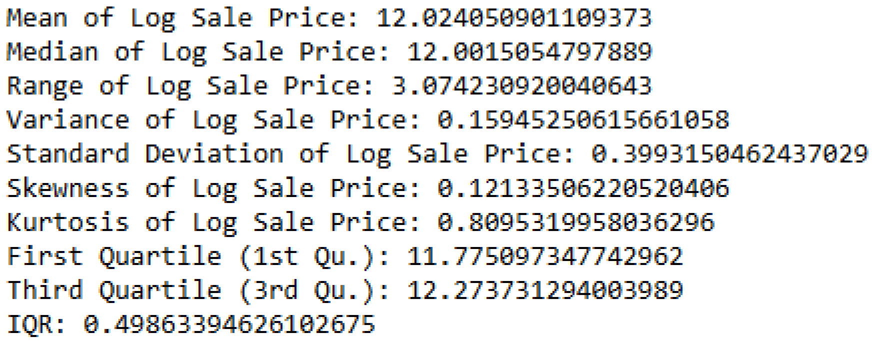


Figure 10

Summary statistics for transformed 'log_sale _price' data



```
Mean of Log Sale Price: 12.024050901109373
Median of Log Sale Price: 12.0015054797889
Range of Log Sale Price: 3.074230920040643
Variance of Log Sale Price: 0.15945250615661058
Standard Deviation of Log Sale Price: 0.3993150462437029
Skewness of Log Sale Price: 0.12133506220520406
Kurtosis of Log Sale Price: 0.8095319958036296
First Quartile (1st Qu.): 11.775097347742962
Third Quartile (3rd Qu.): 12.273731294003989
IQR: 0.49863394626102675
```

References:

Kelleher, J., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (2nd ed.). The MIT Press.

Devore, J. (2016). *Probability and statistics for engineering and the sciences* (9th ed.). Cengage Learning. Kaggle. (n.d.). *House Prices - Advanced Regression Techniques*. Retrieved from https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Federal Housing Finance Agency. (n.d.). *House Price Index*. Retrieved June 20, 2024, from https://www.fhfa.gov/data/hpi#ReleaseDates

Cho, K. (2018, February 11). *Predicting housing prices using advanced regression techniques*. *Towards Data Science*. https://towardsdatascience.com/predicting-housing-prices-using-advanced-regression-techniques-8dba539f9abe