Housing Market Consultancy Project

Austin Mallie, Paola Rodriguez, Tommy Barron

University of San Diego

Foundations of Data Science and Data Ethics: ADS-501-01

Erin Cooke

10/21/24

Housing Market Consultancy Project

In the fast-paced world of real estate, where property values fluctuate with market trends and buyer preferences, the ability to predict home prices accurately can make or break investment decisions. If there was a tool that could analyze the details of a property – like its architecture, location, and build quality – and then translate the information into a precise valuation, it would be a game changer. This is the promise of data science in the real estate industry: a way to turn large datasets into actionable insights that can help not only real estate agents, but also investors and homeowners to navigate the complexities of the market with confidence. This consultancy project leverages the power of machine learning and predictive analytics to uncover the hidden patterns within the Ames, Iowa Housing Dataset.

By analyzing features like overall quality, living area, and neighborhood factors, the goal is to develop a model that can not only predict property prices but does so with the accuracy needed for stakeholders to make strategic decisions. Using the industry-standard CRISP-DM framework, this project goes on a journey – from understanding the business needs to building a model that brings clarity to the dynamic world of property valuation. Through this data-driven approach, we aim to bridge the gap between numbers and market realities to ensure that each prediction is accurate and insightful.

Business Objectives and Success Criteria

Business Objectives

The primary business objective of this project is to develop a machine learning model that predicts property sale prices in Ames, Iowa, based on a wide range of features. The model will serve as a tool for real estate stakeholders, enabling them to make accurate pricing decisions.

By using advanced regression techniques, the project aims to bridge the gap between the raw data and actionable market insights, therefore empowering stakeholders to:

- Optimize Pricing Strategies: Enable real estate agents to set competitive prices for properties based on data-driven analysis.
- Assess Investment Opportunities: Help investors identify undervalued properties by comparing predicted prices with market listings.
- Support Mortgage Underwriting: Provide mortgage lenders with accurate property valuations to streamline the loan approval process.

Success Criteria

To determine the success of the predictive model, the following criteria have been established and will be followed:

1. Prediction Accuracy: The model's performance will be evaluated using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on both training and testing datasets. A target RMSE of less than $20,000 is set as a benchmark for acceptable performance, indicating that the predicted values deviate minimally from actual sale prices.
2. Interpretability of the Model: Beyond predictive accuracy, stakeholders need to understand which features are driving the model's predictions. The importance of features like OverallQual, GrLivArea, YearBuilt, and Neighborhood should be clearly interpretable to allow stakeholders to validate and trust the model's recommendations.
3. User Satisfaction: Gathering feedback from end-users (real estate agents, investors, and lenders) will provide qualitative insights into the model's usability and practicality. If

users find the model useful and intuitive, it will be considered successful from a user adoption perspective.

4. Scalability and Generalization: The model should be capable of making accurate predictions even when new data points are introduced (i.e. future property listings). It should generalize beyond the training data, ensuring that it remains relevant in changing market conditions.

Aligning with the CRISP-DM framework, these objectives and criteria ensure that the project remains focused on solving the real-world challenges of the real estate market.

Dataset Understanding and Overview

The Ames Housing Dataset used in this project consists of 1,460 observations and 79 features that describe various characteristics of properties in Ames, Iowa. These features include numerical, categorical, and ordinal variables that provide insights into different aspects of the properties, such as structural details, neighborhood characteristics, and sales conditions. The target variable is SalePrice, representing the final transaction price of each property. The following list consists of key attributes of the dataset:

1. Numerical Features:
    a. GrLivArea: Represents the above-ground living area in square feet. This is expected to have a strong positive correlation with SalePrice, as larger properties tend to be valued higher.
    b. YearBuilt: Indicates the year in which a property was constructed. Generally, newer properties command higher prices due to modern construction standards and lower maintenance needs.

    c. LotArea: The size of the low on which a house is built, measured in square feet. Larger lots can contribute to higher property values, especially in desirable neighborhoods.

2. Categorical Features:
    a. Neighborhood: Captures the location of the property within different areas of Ames. Different neighborhoods have varying desirability, which impacts property values significantly.
    b. HouseStyle: Describes the architectural style of the property (e.g., 1-story, 2-story, split-level). This can influence buyer preferences and the sale price.
3. Ordinal Features:
    a. OverallQual: Ranges from 1 to 10 and represents the overall material and finish quality of a house. This feature is a strong indicator of SalePrice due to the preference for higher-quality materials and finishes among buyers.
    b. ExterQual: Reflects the quality of exterior materials on the property, contributing to the overall appeal and durability of the house.

Initial Analysis and Findings

A correlation analysis of numerical variables revealed that GrLiveArea and OverallQual have strong positive correlations with SalePrice, indicating that larger, high-quality homes tend to sell for more. Visualizations like scatter plots and boxplots highlighted the relationship between these variables and SalePrice, confirming their importance in predicting house prices. The distribution of SalePrice shows a right-skewed patter, with most properties priced below the median but with some high-priced outliers. This suggests the need for potential transformations, such as log transformation, to normalize the data before applying regression models.

Inventory of Resources

- Data: The primary data source is the Ames Housing Dataset, provided by Kaggle. This dataset includes extensive property details, making it suitable for developing a comprehensive predictive model.
- Software: The analysis will be conducted using Python libraries, including:
  - Pandas: For data manipulation and cleaning.
  - Seaborn and Matplotlib: For data visualization and exploration.
  - Scikit-Learn: For implementing regression models and performing feature selection.
  - Jupyter Notebook: For iterative analysis and collaboration, which also runs Python 3.9.
- Hardware: This project will utilize computer resources with a minimum of 16GB of RAM and a minimum of Intel i5 CPU capacity to handle the data processing and model training.
- Personnel: The project team includes graduate students from the University of San Diego who are experts in the field of Data Science and real estate analysis. Their expertise will help validate the model's predictions and ensure that business needs are met.
- Documentation and Reporting: Tools like Microsoft Word and Google Docs will be used for documentation, which PowerPoint will be used for presenting findings to stakeholders.

Data Accessibility

The project relies on the availability of the Ames Housing Dataset, including detailed property features. Access to clean, well-documented data is essential for model training, testing, and validation. The project also requires access to Python programming language, data analysis libraries (e.g., Pandas, ScikitLearn, etc.), and data visualization tools like Matplotlib and Seaborn. These are needed to preprocess data, build models, and analyze results. Additionally, feedback from stakeholders, such as real estate professionals and investors, is critical during the model evaluation phase to ensure the predictive model aligns with business needs. Lastly, sufficient processing power is required to handle large datasets, run model training, and perform cross-validation efficiently. Access to cloud computing platforms like Google Collab or local machines with high RAM is important for this purpose.

It is assumed that the Ames Housing Dataset is representative of broader real estate trends in Ames, Iowa. This ensures that the model's predictions will be applicable to future property sales. The relationship between features (e.g., property size, quality, and location) and SalePrice is assumed to be stable over time. This assumes that external factors like major economic shifts or changes in real estate regulations do not significantly alter the market dynamics during the analysis period. It is important to stress that the project assumes that stakeholders' objectives, such as maximizing accuracy in property valuation and improving decision-making processes, will remain consistent throughout the project lifecycle.

Constraints

The project has a defined time frame for completion, which limits the depth of analysis and the exploration of multiple modeling approaches. The presence of missing values and outliers in the dataset requires significant preprocessing efforts. The accuracy of the model could be impacted if data cleaning is not thoroughly conducted, and complex models may require more computational resources and time, potentially exceeding the available hardware capabilities. This necessitates the use of simpler models or feature selection techniques.

RESOLVEDD Strategy

The RESOLVEDD Strategy is a structured approach to ethical decision-making and problem-solving, guiding the project through nine key steps:

1. Review the Facts:
   a. The Ames Housing Dataset contains over 1,460 records of property features and sale prices. Accurate pricing is crucial for stakeholders like real estate agents and investors.
   b. Machine Learning models can provide data-driven insights but require careful handling of data to ensure fairness and accuracy.
2. Estimate the Problem:
   a. Potential problems that could arise ethical concerns include overfitting due to the high number of features, model bias due to unbalanced data, neighborhood bias, outliers, and the risk of not meeting stakeholder expectations if predictions are not accurate enough.
3. List Possible Solutions:
   a. Implement regularization techniques (e.g., Lasso regression) to prevent overfitting.
   b. Use cross-validation to ensure model robustness.
   c. Perform feature selection to reduce dimensionality while maintaining predictive power.
   d. Engage stakeholders throughout the process for feedback and adjustments.
4. State Outcomes:
   a. Improved predictive accuracy in estimating home values.

   b. Enhanced understanding of which property features are most significantly impacting pricing.

   c. Increased stakeholder trust and satisfaction with the predictive tool.

5. Describe Likely Impact:

   a. A well-performing model could lead to more accurate pricing strategies and better investment decisions, potentially improving market efficiency.

   b. Transparent explanations of model predictions will foster greater trust among stakeholders, including lenders and real estate professionals.

6. Explain Values:

   a. Transparency: Ensuring that model decisions are understandable and explainable to stakeholders.

   b. Accuracy: Providing the most precise property valuations possible to support informed decision-making.

   c. Fairness: Avoiding bias in predictions, ensuring that properties are values equitably across different neighborhoods and types.

7. Evaluate Each Solution:

   a. Regularization reduces overfitting buy may require careful parameter tuning to maintain interpretability.

   b. Cross-validation ensures that the model generalizes well but can be computationally intensive.

   c. Feature selection simplifies the model but risks excluding relevant information.

8. Decide and Clarify:
    a. The decision is to use a regularized regression model combined with cross-validation to balance model complexity and performance. Feature selection will focus on the most impactful attributes like OverallQual, GrLiveArea, and YearBuilt.
9. Defend the Decision:
    a. The choice of regularization and cross-validation ensures a robust model that minimizes overfitting while maintaining accuracy.
    b. By focusing on key features, the model remains interpretable, allowing stakeholders to understand and trust its predictions.

Risk and Contingencies

Identified Risks:

Due to the high number of features, there is a risk that the model could perform exceptionally well on training data but poorly on unseen data, which is also known as overfitting. There is also a concern of data quality issues, such as missing values, outliers, and inconsistencies in the dataset that can skew predictions if not properly addressed during preprocessing. There is also a possibility of bias in the dataset as it may have an imbalance across different neighborhoods or property types, which well then lead to biased predictions. Ultimately, with any data analyzation, there is a risk that the final model might not align perfectly with stakeholder expectations, especially if predictive accuracy is lower than desired (OpenAI, 2024). This is why we stress the importance of setting expectations regarding the data and the constraints we are faced with.

Contingency Plans

As a result of these identified risks, we have several contingency plans in place to ensure we achieve the best results possible. By applying regularization techniques, like Lasso and Ridge regression to control model complexity and using cross-validation to assess model performance on multiple subsets of the data, we can ensure that the model generalizes well and mitigate any overfitting.

Addressing Data Quality

By implementing imputation methods (e.g., mean or median imputation) for missing values, we can use these methods as outlier detection techniques to identify and handle extreme values that could distort model results. We then follow this process by conducting exploratory data analysis (EDA) to ensure data integrity before modeling.

Mitigating Bias

By using stratified sampling during model training, we can ensure that different neighborhoods and property types are represented proportionally, this mitigating any bias present. Additionally, it is important to perform fairness checks to validate that the model is not systematically underestimating or overestimating prices for specific groups.

Managing Stakeholder Expectations

It is important to maintain open communication throughout the project, and this can be achieved by holding regular meetings with stakeholders to provide updates and gather feedback. Even taking the steps to developing a communication plan to explain the model’s predictions and limitations clearly can help ensure that stakeholders understand the potential variance in predictions and that everyone is on the same page.

Terminology

The glossaries below are common terms that are used throughout the domain of business and data mining. The business glossary describes terms that are commonly used in the real estate market. The data mining glossary outlines common terms used for data science, data visualization, and statistical mathematics.

Business Glossary

Homeowner: A person who owns the house or property in which they live.

House Style: The style of dwelling that includes how many floors and if the second level is finished

Investor: An investor is an individual or entity that allocates money or other resources with the expectation of receiving financial returns or profit over time.

Overall Quality: This feature variable ranks the condition of the house from 1-10 where 10 is very excellent and 1 is very poor.

Property: Land and any structures or improvements attached to it.

Real Estate Agent: A licensed professional who assists individuals and businesses by buying, selling, or renting properties.

Sale Price: The final monetary amount that closed the deal on a property transaction.

Scalability: The capacity to be changed in size or scale.

Zoning Classifications: Identification of the general zone type of the sale.

Data Mining Glossary

Accuracy: How well the model predicts or classifies data compared to the actual outcomes.

Boxplot: A graphical representation used to summarize the dataset’s key statistical measures, making it easier to identify trends, outliers, and the overall spread of the data.

Categorical: Data that can be classified into distinct categories or groups.

CRISP-DM: A widely used framework for managing data mining and data science projects.

Data Cleaning: The process of fixing or removing incorrect, corrupt, incorrectly formatted, duplicate, or incomplete data within a data set.

Data Normalization: The process of scaling down the data set such that the normalized data falls between zero and one.

Data Preparation: The process of collecting, transforming, and enriching raw data to make it suitable for analysis.

Exploratory Data Analysis: The process of analyzing and investigating data sets, summarizing the main characteristics, and using data visualization methods to gain insights into the data.

Explanatory Variable: A variable in statistical analysis that is used to explain or predict the variation in another variable.

Fairness Checks: The evaluation of whether a model, algorithm, or system treats different groups or individuals fairly.

Feature Engineering: The process of selecting, modifying, or creating new features from raw data to improve the performance of machine learning models.

Feature Importance: Indication of how much each feature/variable contributes to the model prediction.

Key Performance Indicators: Specific and measurable metrics used to track the performance, quality, or impact of data within an organization.

Log Transformation: A mathematical technique used to stabilize variance and make data more normally distributed, which can improve the performance of statistical models and analyses.

Machine Learning: A branch of artificial intelligence that focuses on developing algorithms and models that enable computers to learn from and make decisions or predications based on data, without being explicitly programmed to perform a specific task.

Mean Absolute Error: The measurement of the absolute difference between the predicted values and the actual values.

Mean Squared Error: The measurement of the average of the squares of the difference between predicted values and actual values.

Numerical: A value that represents a quantity and can be expressed as a number.

Ordinal: A type of categorical data that represents the order of ranking of items but does not convey the exact differences between them.

Overfitting: A common problem in machine learning and statistical modeling where a model learns not only the underlying patterns in the training data but also the noise and outliers.

Precision: A performance metric used in classification tasks to evaluate the accuracy of a model in identifying positive instances.

Predictor/Feature Variable: A variable that is used in statistical models and machine learning algorithms to predict or explain the value of another variable.

R-Squared: The proportion of the variance in the dependent variable that can be explained by the independent variable(s) in the model.

Regression: A statistical and machine learning technique used to model the relationship between a dependent variable (also called the response variable) and one or more independent variable(s) (also called predictors or features).

Response Variable: The variable that is being predicted or explained in a statistical model or experiment.

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

Scatterplot: A data visualization that displays the relationship between two quantitative variables.

Skew: The asymmetry of the probability distribution of a real-values random variable.

Standard Deviation: A statistical measure of variability that indicates the average amount that a set of numbers deviates from their mean.

Stratified Sampling: A sampling technique that involves dividing the population into distinct subgroups based on certain characteristics.

Data Mining Goals and Success Criteria

Data Mining

The objective of this project is to create an algorithm that closely predicts the final closing cost of a property sale given various characteristics of physical features. The team will be creating a predictive model using weighted regression to evaluate the impact of each explanatory variable on the response variable. As we know, different parts of a property are favored more by the buyers than others, such as available utilities, lot size, etc. Therefore, it makes sense to value some of these variables more than others. In a weighted regression model, we can adjust the strength of each variable to fine tune the accuracy of the results.

To measure the success of our model, we will use historical data from previous property sales and compare the accuracy of our algorithm to the real life equivalent. To confirm our model is robust, we will use statistical evaluations to analyze and refine the results. One measurement we will use is the r-squared value. This value ranges from zero to one where a value of one can accurately determine the results one hundred percent of the time. There is no perfect cutoff for an r-squared value that is good or bad, but generally, the higher, the better (Bobbitt, 2019). The other evaluation metrics are the mean absolute error, root mean squared error, and mean squared error, or MAE RMSE, and MSE, respectively. These metrics will provide us with a calculation of the difference between the actual sale price and the predicted sale price. Typically, the lower these metrics are, the closer the model's output is to the actual value (Kumar, 2024). The team will be aiming for these metrics to be as low as $20,000, so accurate and confident business actions may be taken. By evaluating all the features presented in the data set, the team should get a strong understanding of which models will offer the best performance. If provided with the proper allocation of funds and timeframe, we will be able to create an algorithm that will accurately appraise property and allow us to offer the best real estate service at a competitive value.

Business Success

The statistical metrics mentioned previously align with theoretical and calculatable benchmarks, but the business will require different key performance indicators to assess the impact of the project. One of these KPIs will be monthly recurring revenue (Atlan, 2023). This metric will have the most direct impact on the organization because it follows the month-over-month revenue growth that the real estate agents generate from sales and commissions. Ideally, the success of the project relies on significant growth of revenue before and after the project implementation. Another KPI that can be used to measure the business impact of this project is operational efficiency through the length of inventory turnover (Atlan, 2023). By tracking the life cycle of a property sale, we can accurately measure the optimal time to market adjust the sale price to get the best commission for the business. Additionally, with increased inventory turnover, more sales will occur, and the overall monthly recurring revenue rate will increase for the company.

Project Plan with Gantt Chart

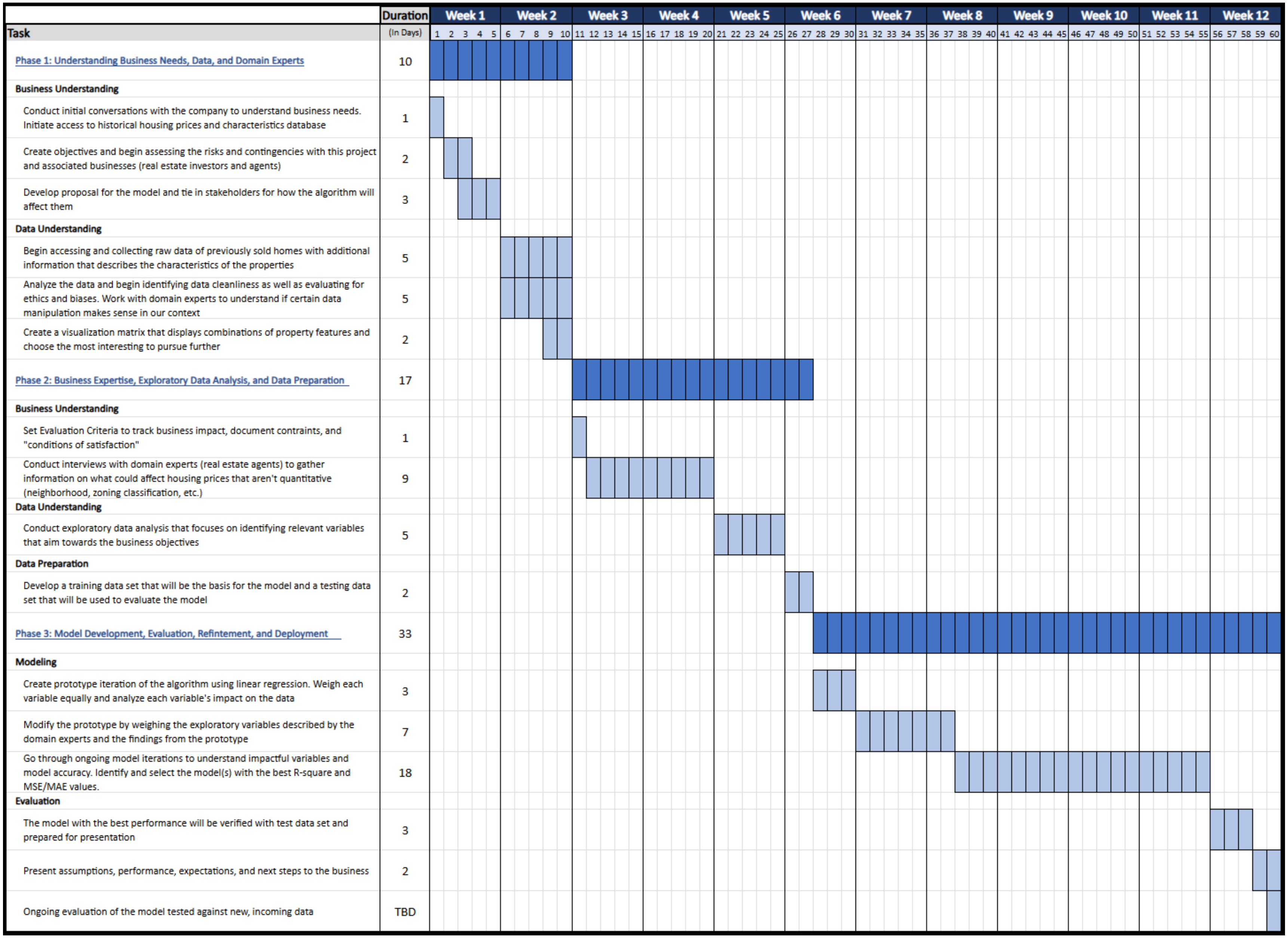
The project will be approximately 12 weeks, broken into three phases. The first phase will start the conversation for business needs, data access, and domain expertise. The team will begin by communicating with the stakeholders and outline any risks and contingencies that might develop over the course of the project. Next, data will be collected from data archines that will be evaluated for relevancy, ethics, and potential biases that may affect the model outcome. If needed, domain experts may be consulted to understand the data. Using python visualization tools, the team will create rudimentary visualizations that aim to guide the next steps of the project.

The second phase of the project will draw in consultation with domain experts, conducting exploratory data analysis, and preparing the data for modeling. The team will be conducting interviews with domain experts to gather additional information regarding potential confounding variables that aren't quantitative and may negatively impact on the model (i.e. neighborhoods, zoning classifications, etc.). Using the knowledge gained from the experts, criteria will be set to document business impact and evaluate stakeholder satisfaction. After this, extensive exploration data analysis will be used to shape the data in a way that is relevant to the business criteria. Once this is completed, the exploratory data will be split into a training set and test set to measure the accuracy and ethical nature of the model.

Finally, the third phase of the plan is to develop the model, evaluate the model based on varying statistical metrics, and prepare the model for deployment to the business. To start, the team will create a prototype model with equally weighted variables. Then, using modified regression techniques, the model will be adjusted by applying various degrees of weights onto the variables of the model. The process will go through different iterations of development to understand the optimal variable weights and assess the accuracy of the models using statistical metrics. Once the final model has been developed and chosen, it will be tested with the test data set to ensure the validity of training results. Finally, the model will be showcased to the business and stakeholders and will undergo the next steps for the business.

Figure 1

Gantt Chart for Project Plan



Task	Duration (In Days)
Phase 1: Understanding Business Needs, Data, and Domain Experts	10
Business Understanding	
Conduct initial conversations with the company to understand business needs. Initiate access to historical housing prices and characteristics database	1
Create objectives and begin assessing the risks and contingencies with this project and associated businesses (real estate investors and agents)	2
Develop proposal for the model and tie in stakeholders for how the algorithm will affect them	3
Data Understanding	
Begin accessing and collecting raw data of previously sold homes with additional information that describes the characteristics of the properties	5
Analyze the data and begin identifying data cleanliness as well as evaluating for ethics and biases. Work with domain experts to understand if certain data manipulation makes sense in our context	5
Create a visualization matrix that displays combinations of property features and choose the most interesting to pursue further	2
Phase 2: Business Expertise, Exploratory Data Analysis, and Data Preparation	17
Business Understanding	
Set Evaluation Criteria to track business impact, document contraints, and "conditions of satisfaction"	1
Conduct interviews with domain experts (real estate agents) to gather information on what could affect housing prices that aren't quantitative (neighborhood, zoning classification, etc.)	9
Data Understanding	
Conduct exploratory data analysis that focuses on identifying relevant variables that aim towards the business objectives	5
Data Preparation	
Develop a training data set that will be the basis for the model and a testing data set that will be used to evaluate the model	2
Phase 3: Model Development, Evaluation, Refintement, and Deployment	33
Modeling	
Create prototype iteration of the algorithm using linear regression. Weigh each variable equally and analyze each variable's impact on the data	3
Modify the prototype by weighing the exploratory variables described by the domain experts and the findings from the prototype	7
Go through ongoing model iterations to understand impactful variables and model accuracy. Identify and select the model(s) with the best R-square and MSE/MAE values.	18
Evaluation	
The model with the best performance will be verified with test data set and prepared for presentation	3
Present assumptions, performance, expectations, and next steps to the business	2
Ongoing evaluation of the model tested against new, incoming data	TBD

Data Description

The dataset for this project is derived from Ames, Iowa, and is designed to predict house sales prices based on a wide array of features. It includes 2,919 records, with 1,460 records in the training set and 1,4,59 in the test set. These records provide a comprehensive view of various factors influencing property values, including structural characteristics, environmental conditions, and amenities. The dataset contains 81 features, with data types ranging from integers (26 int64), floats (12 float64), to categorical variables (43 object types).

The target variable, SalePrice, is the focus of the regression model, aiming to forecast house prices by leveraging the relationships between numerous features. These features range from OverallQual, which measures the overall material and finish quality of the house, to Neighborhood, reflecting the geographical location of the property. Other important attributes include YearBuilt, LotFrontage, and GarageType, all of which provide further insight into the physical and functional aspects of each house.

By understanding these features, the dataset allows for a detailed exploration of factors that affect property valuation in Ames, aligning with the broader business goal of providing accurate price predictions. This will serve as a valuable tool for real estate stakeholders, including homebuyers, sellers, and appraisers, as they make more informed decisions based on the underlying data.

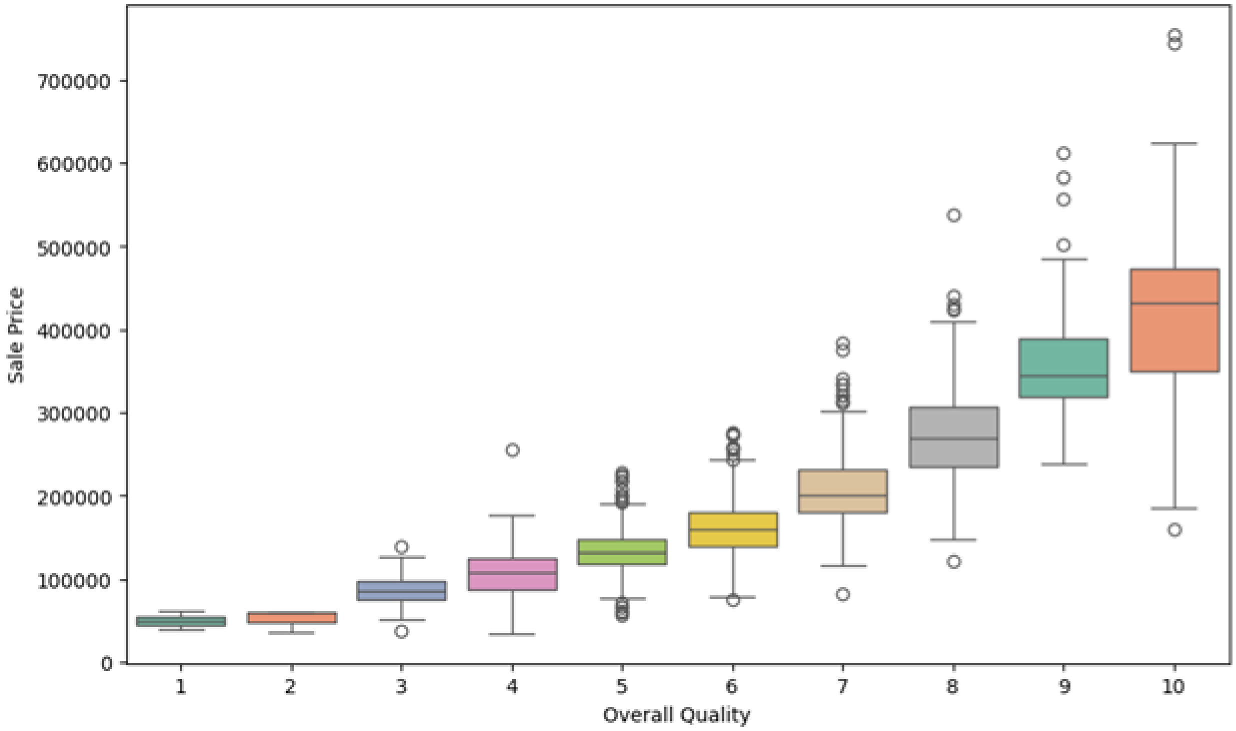
Data Exploration

In line with the CRISP-DM model’s focus on understanding the data, an exploratory data analysis was conducted to uncover patterns, trends, and key relationships between the features and SalePrice. The average sale price of a house in this dataset is $180,921, with a significant standard deviation of $79,442, indicating a wide range in property values. The most expensive home is priced at $755,000, while the least expensive was sold for $34,900, reflecting the diversity in housing stock within Ames.

Several features show strong correlations with SalePrice, which are critical for building an effective predictive model. OverallQual, which measures the overall quality of the house, has the strongest positive correlation with a value of 0.791, indicating that higher-quality homes generally command higher prices. As seen in Figure 2, the price for homes increases per quality level. Houses with the quality level of one will more often than not have the lowest sales price while houses with a quality level of ten will have the highest sales price.

Figure 2

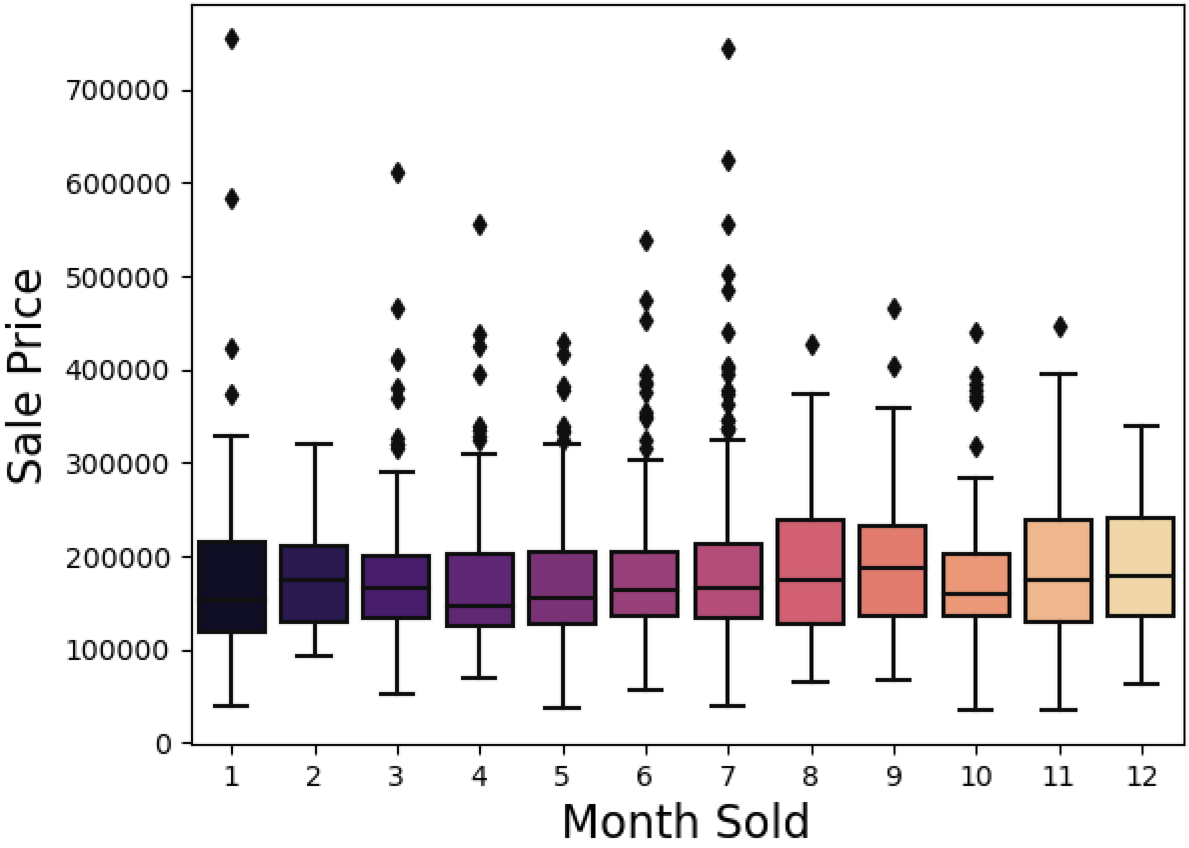
Overall Quality by Sale Price



Within the data set, another relationship between the sale price and the month sold was also found. The boxplot in Figure 3 shows the varying sale price of houses between the twelve months of the year. It is interesting to discover that the best months for sale prices were September, or month 9, and December, month 12, due to their higher median sale price compared to the other months of the year. The beginning months of the year show smaller median house prices and are clustered in the same area. The months of March, June, and July have a lot of outliers indicating that a few large values houses were sold and are increasing the median sales price for those months.

Figure 3

Month Sold by Sale Price



Data Quality

Ensuring high-quality data is a crucial step in the CRISP-DM model, particularly within the data preparation phase. The dataset is mostly complete, but several features exhibit missing values that need to be addressed before proceeding with the modeling. For instance, variables such as Alley, Fence, MiscFeature, and PoolQC have substantial missing data. A structured imputation strategy will be applied, where missing values for features like LotFrontage can be filled in using the median lot size within each neighborhood. Similarly, missing garage-related fields may be imputed with “No Garage” to represent properties that lack such amenities.

Table 1

Missing Values of Figure Characteristics

Feature	PoolQC	MiscFeature	Alley	Fence	FireplaceQu
Missing Values	1,453	1,406	1,369	1,179	690
Feature	LotFrontage	GarageType	GarageYrBlt	GarageFinish	GarageQual
Missing Values	259	81	81	81	81
Feature	GarageCond	BsmtExposure	BsmtFinType2	BsmtFinType1	BsmtCond
Missing Values	81	38	38	37	37
Feature	BsmtQual	MasVnrArea	MasVnrType	Electrical	
Missing Values	37	8	8	1	

The most affected columns are those that relate to specific property features such as pools, garages, and basement conditions. These missing values will require appropriate imputation, depending on whether the absence of data reflects a meaningful “no feature” case (e.g., no garage or no pool) or if it needs to be inferred from other similar data points.

Table 2

Descriptive Statistics of Feature Characteristics

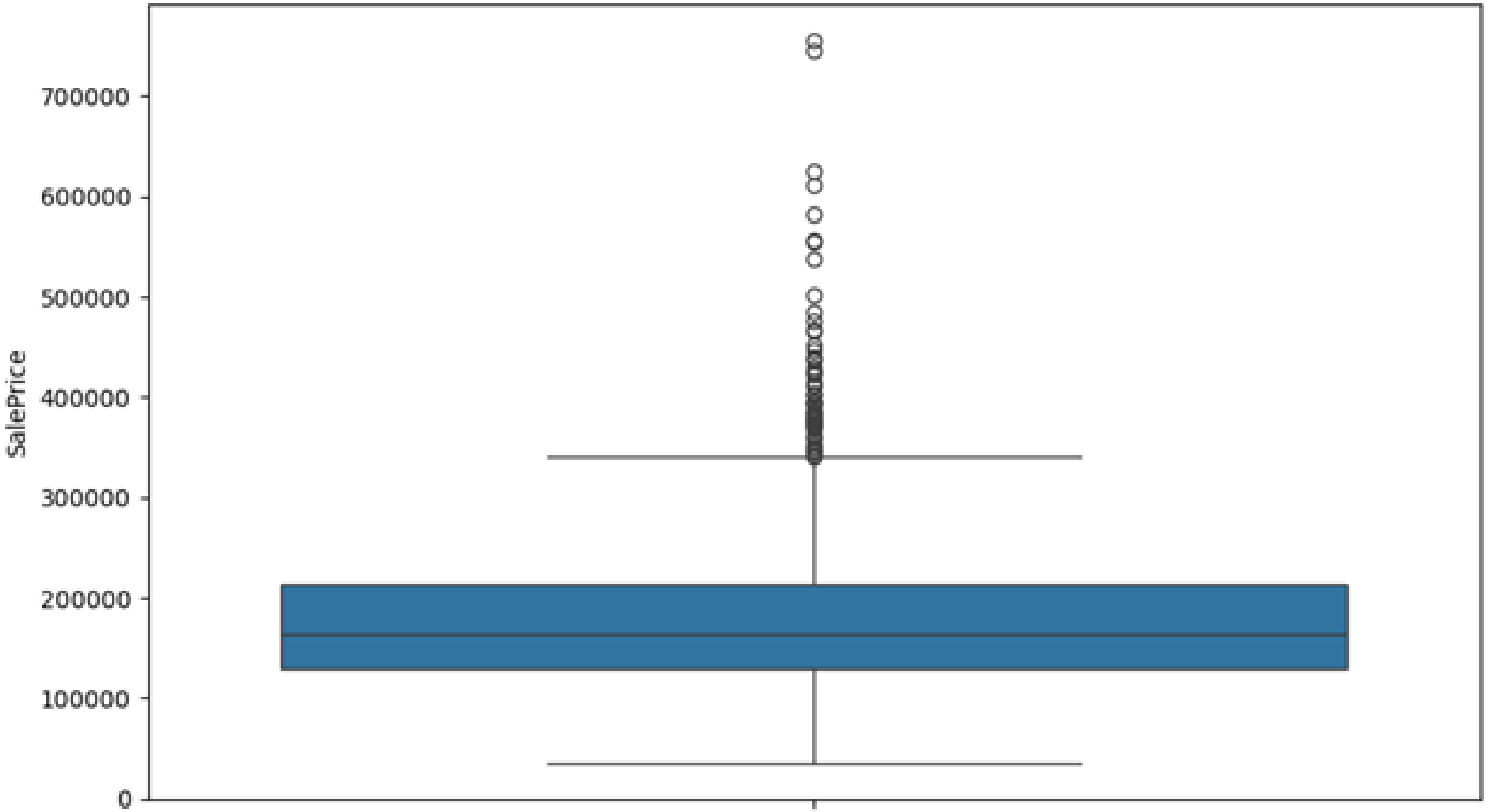
					1st	3rd	
Variable	Mean	Std Dev	Median	Min	Quartile	Quartile	Max
1stFlrSF	1162.63	386.59	1087	334	882	1391.25	4692
2ndFlrSF	346.99	436.53	0	0	0	728	2065
3SsnPorch	3.41	29.32	0	0	0	0	508
BsmtFinSF1	443.64	456.10	383.5	0	0	712.25	5644
BsmtFinSF2	46.55	161.32	0	0	0	0	1474
BsmtUnfSF	567.24	441.87	477.5	0	223	808	2336
EnclosedPorch	21.95	61.12	0	0	0	0	552
GarageArea	472.98	213.80	480	0	334.5	576	1418
GrLivArea	1515.46	525.48	1464	334	1129.5	1776.75	5642
LotArea	10516.83	9981.26	9478.5	1300	7553.5	11601.5	215245
LotFrontage	70.05	24.28	69	21	59	80	313
LowQualFinSF	5.84	48.62	0	0	0	0	572
MasVnrArea	103.69	181.07	0	0	0	166	1600
MiscVal	43.49	496.12	0	0	0	0	15500
OpenPorchSF	46.66	66.26	25	0	0	68	547
PoolArea	2.76	40.18	0	0	0	0	738
ScreenPorch	15.06	55.76	0	0	0	0	480
TotalBsmtSF	1057.43	438.71	991.5	0	795.75	1298.25	6110
WoodDeckSF	94.24	125.34	0	0	0	168	857

This table provides more comprehensive insights into the distribution of continuous variables, highlighting that many features (such as PoolArea, 3SsnPorch, and ScreenPorch) have a large number of zeros, reflecting properties that lack these amenities. The minimum and maximum values for variables such as LotArea and GarageArea are within reasonable ranges, though outliers in the upper end will need to be examined to determine if they represent rare but valid property features or if they indicate potential data errors.

Outliers were also identified, particularly at the extremes of SalePrice, where properties priced exceptionally high or low may represent special cases, such as luxury homes or distressed sales as shown in the figure below. The number of outliers increases in the middle prices of homes and decreases as you get to the larger and smaller prices for homes. These outliers will be carefully reviewed, and depending on their nature, may be removed or adjusted to ensure they do not unduly influence the model.

Figure 4

Boxplot of SalePrice



References

Atlan. (2023, September 22). KPIs for Every Data Team: A 2024 Guide! Atlan. https://atlan.com/kpis-for-data-team/#:~:text=KPI%20stands%20for%20Key%20Performance%20Indicator%2C%20and%20in,quality%20management%2C%20data%20analytics%2C%20or%20even%20data-driven%20decision-making

Bobbitt, Z. (2019, February 24). What is a Good R-squared Value? Statology. https://www.statology.org/good-r-squared-value

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Chearer, C., & Rüdiger, W., (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS.

Montoya, A., & DataCanary. (2016). House Prices – Advanced Regression Techniques. Kaggle. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

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.

Kumar, A. (2024, August 18). MSE vs RMSE vs MAE vs MAPE vs R-Squared: When to Use? Analytics Yogi. https://vitalflux.com/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/#Root_Mean_Squared_Error_EMSE

OpenAI. (2024). ChatGPT (October 2023 version). https://chat.openai.com/

Siegel, E. (2016). Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons, Hoboken, New Jersey.