Consulting Project

Group 8

Davood Aein

University of San Diego

Master of Science, Applied Data Science

Foundations of Data Science and Data Ethics (ADS-501-01)

2/24/2024

Business Understanding

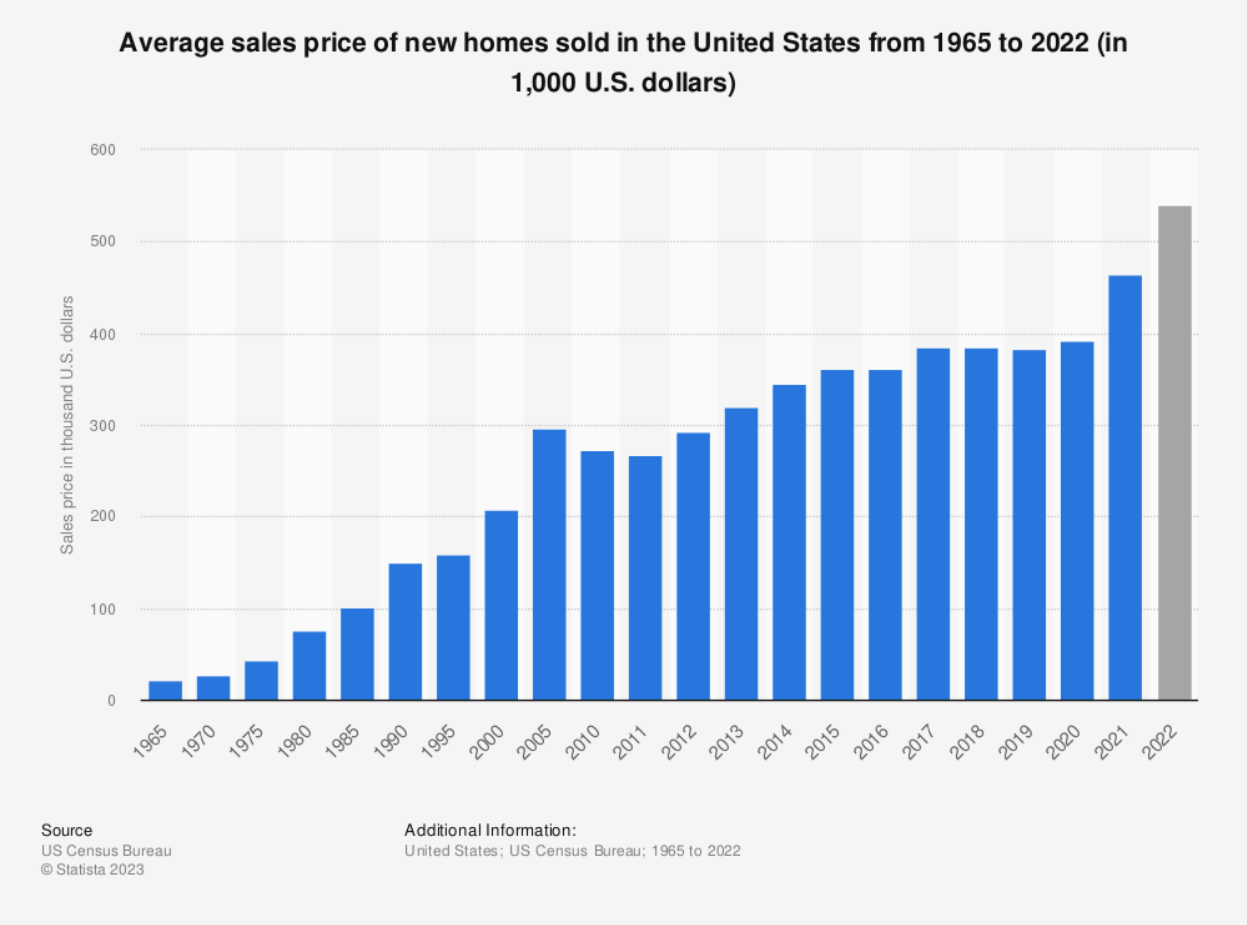
Background

House prices hold immense significance for individuals and society at large, influencing personal wealth, homeownership accessibility, and economic indicators. For homeowners, property values directly impact personal wealth, acting as a financial asset. The affordability of housing, tied to house prices, dictates the ability of individuals and families to achieve homeownership, a key aspect of financial stability. Beyond personal finances, house prices serve as economic indicators, reflecting market health and influencing investment decisions. Real estate investments, influenced by property values, contribute to economic activity. Additionally, house prices affect lending practices, local government revenues through property taxes, and consumer confidence. The stability of the housing market is vital for overall economic stability, impacting employment in the real estate sector and influencing wealth inequality. In essence, house prices are integral to understanding and navigating the economic landscape, with far-reaching implications for individuals and the broader community.

The article primarily explores the connection, between elevated housing prices and their effects on spending habits and the overall macroeconomic situation in China. According to the article, high house prices can potentially have an impact on household consumption by boosting family assets and encouraging families to be more willing to spend. However, there is a viewpoint suggesting that high house prices could potentially hinder consumption for those who do not own a house or are looking for improved living conditions (Fan, 2024).

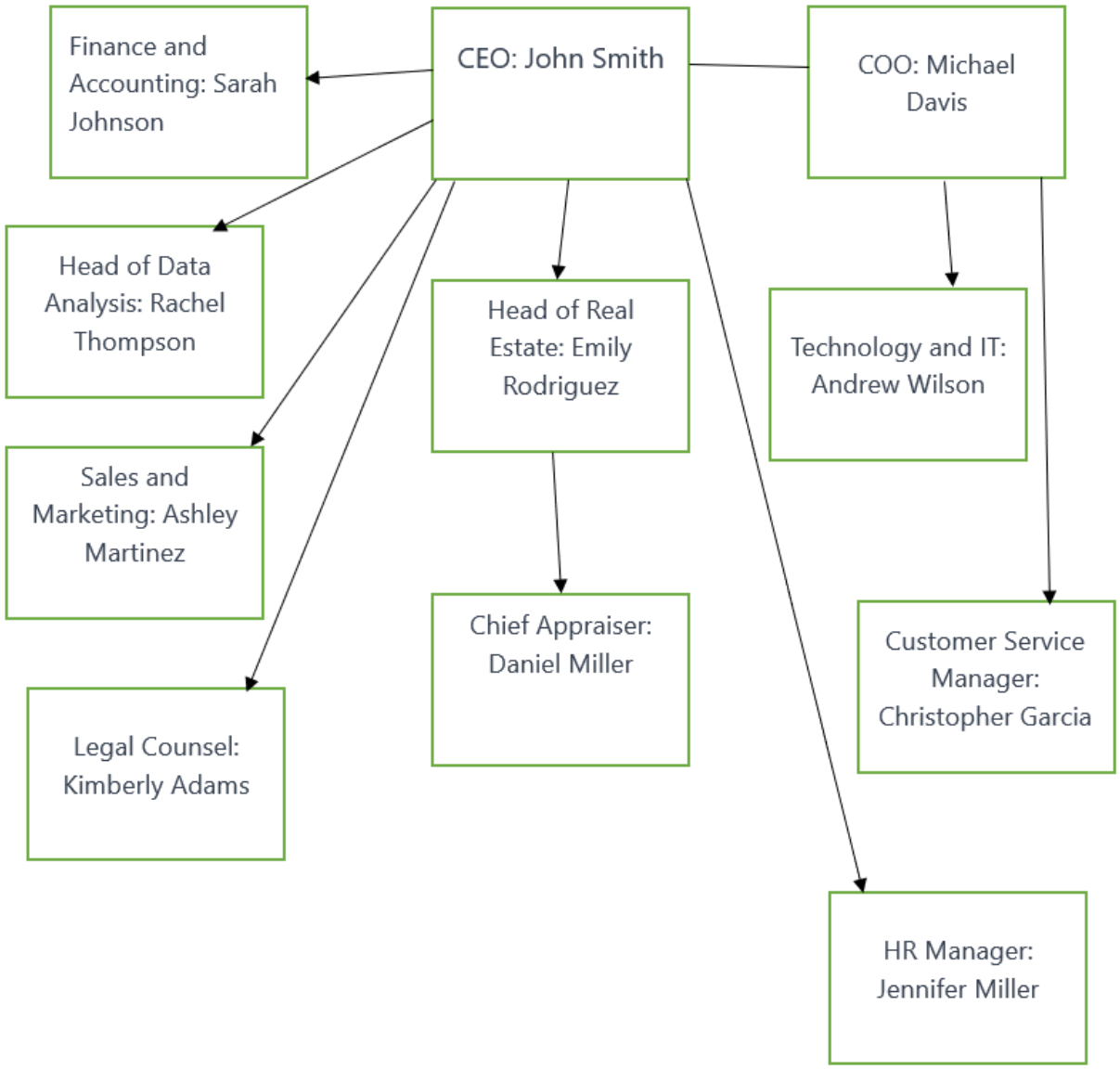
In this graph, we can see the incremental house prices in the US during the past 57 years. It is obvious that the house price has been increasing for more than 50 years. However, there are

some years in which house prices diminished compared to the year before, but in most cases, we can see growth.



(1)

The dataset focuses on predicting the sale prices of properties in dollars. This prediction task is crucial for real estate stakeholders, including property developers, buyers, and sellers. Accurate predictions can aid in strategic decision-making, pricing optimization, and understanding market trends.



(2)

Our company, specializing in house prices, is proud to have Kaggle Inc. as an internal sponsor for our ongoing projects. Kaggle, renowned for hosting skills-based competitions that contribute significantly to advancing the field of data science, plays a vital role in our initiatives.

The company generously provides access to datasets and relevant materials crucial for our work, such as the dataset utilized in this project, originally created by figure-eight, and shared on Kaggle's 'Data for Everyone' website (Kaggle, n.d.). While our organizational chart illustrates the internal structure guiding our efforts in the realm of house prices, Kaggle serves as a key external collaborator, enriching our projects with its wealth of resources and fostering innovation within the data science community. We value this partnership, as it aligns with our commitment to leveraging cutting-edge tools and platforms to enhance the quality of our work in the real estate and data analysis domains.

In the realm of house prices, our business units most significantly impacted by ongoing projects are operations and legal protection. This initiative sheds light on the strengths and weaknesses of our operations unit in efficiently managing and organizing the current flow of content related to real estate on various platforms. It provides a comprehensive overview, enabling the operations team to identify areas of improvement and enhance their processes. Simultaneously, the legal protection division also faces scrutiny, with the project spotlighting their strengths and weaknesses in regulating the posting of appropriate or non-appropriate content to the public. The findings from this project will be instrumental in guiding the vice president of operations and the head of legal, policy, and trust. It serves as a proposal and guide for these leaders to reconsider and potentially revise their approaches, implementing stricter policies that ensure the dissemination of appropriate content. This proactive measure aims to prevent the propagation of content that may instill fear or pose a danger to communities or agencies, reinforcing our commitment to responsible and impactful information sharing in the dynamic landscape of real estate.

The project focusing on house prices encounters several problem areas, notably in the domains of business management, operations, and customer care. These challenges collectively contribute to the overarching issue at hand, which revolves around assessing whether a tweeted message on platforms like Twitter indicates the occurrence or non-occurrence of an emergency related to real estate or housing. The complexity of this task requires a nuanced approach, given the diverse nature of user-generated content.

In the realm of estimating house prices, the current landscape encompasses a blend of traditional real estate valuation practices and contemporary data-driven methodologies. Traditional approaches often involve seasoned appraisers conducting comparative market analyses, evaluating property values by juxtaposing similar properties within a specific geographic area. Concurrently, emerging technologies have ushered in data analytics and machine learning to the real estate sector. Modern platforms employ algorithms that scrutinize an extensive array of data, including property features, local market trends, economic indicators, and recent sales data, aiming to predict house prices. These advanced models leverage statistical techniques to enhance accuracy and provide dynamic and nuanced estimates.

One notable advantage lies in the accuracy and precision of traditional methods, particularly when conducted by experienced professionals who bring a wealth of industry knowledge to their assessments. Data-driven insights from modern approaches offer another advantage, as these algorithms can uncover patterns and trends that may elude traditional methods. Moreover, efficiency is a key benefit, especially with automated valuation models (AVMs) that swiftly generate estimates, streamlining the valuation process.

However, challenges persist within the current landscape. Traditional methods, at times, can introduce subjectivity into the valuation process, with appraisers' judgments potentially

leading to variability in property valuations. Data-driven models face their own set of limitations, heavily relying on the quality and completeness of available data. Incomplete or biased data can compromise the reliability of predictions. Moreover, the lack of transparency in complex machine learning algorithms may pose a challenge to user acceptance, with some stakeholders finding it difficult to trust or understand the intricacies of the predictive models.

The level of acceptance among users exhibits variability within the industry. Traditional methods, deeply rooted in established practices, often enjoy a higher degree of trust from certain stakeholders such as real estate agents and appraisers. Conversely, newer data-driven solutions may be embraced by more tech-savvy users seeking dynamic and data-driven insights. The acceptance level is contingent upon the transparency and explain ability of the methods employed for predicting house prices. Ongoing feedback collection from users and stakeholders is paramount in gauging and enhancing the overall acceptance of the current solutions.

Business Objectives and Success Criteria

The overarching goal of this project is to refine pricing strategies within the housing market, aiming to optimize revenue and attract potential buyers effectively. By leveraging data-driven insights, the objective is to enhance decision-making processes related to house pricing.

In pursuit of the primary business objective, several key questions need addressing. These include understanding factors influencing customer decisions to move to competitors, determining the impact of lower fees on specific customer segments, analyzing market trends, assessing competitors' pricing strategies, identifying customer preferences, exploring regional variances, and evaluating the impact of economic factors on house prices. Some questions to solve the problems would include the following:

- How do various factors influence customers' decisions to switch to competing housing options?
- If lower fees are implemented, how might this impact specific customer segments, and could it lead to an increase in market share?
- What are the current market trends affecting house prices, and how can these trends be leveraged for competitive advantage?
- In what ways can the analysis of competitors' pricing strategies inform our own pricing decisions?
- How can customer preferences be accurately identified and integrated into pricing strategies?
- What types of housing features or characteristics significantly influence customer choices in the market?
- Are there regional variations in housing demand that necessitate customized pricing strategies?
- How can regional nuances be incorporated into pricing decisions to maximize market responsiveness?
- To what extent do economic factors such as interest rates and unemployment influence house prices?
- How can the business model adapt to economic fluctuations while maintaining competitive pricing?
- How would adjustments in pricing affect the income generated on a regular basis (e.g., monthly, annually)?

- What divisions or departments within the housing market business would be most affected by changes in pricing and how?
- To what extent can innovative research methods and technological solutions, such as machine learning, be employed to enhance pricing strategies?
- How willing is the organization to invest in technology and data science tools to improve access to data, analytics, and research methods?
- What ethical considerations should be considered when developing pricing strategies?
    - How can equitable principles be integrated into pricing models to ensure fairness in housing markets?

To ensure the success of the project, it is imperative to establish specific business requirements. These include the need to retain current customers during pricing adjustments, maintain competitiveness in the market, adhere to data privacy regulations to safeguard customer information, and guarantee the accuracy of the predictive model for effective decision-making. Anticipated benefits of this project include increased revenue through optimized pricing strategies, enhanced customer retention by understanding and meeting customer preferences, a competitive advantage achieved by staying ahead of market trends, efficient resource allocation based on demand-driving factors, and compliance with data privacy regulations to maintain customer trust. This project addresses the complex challenge of optimizing house pricing strategies within the dynamic housing market. Through a data-driven approach, the aim is to provide actionable insights into customer behavior, market trends, and competitive landscapes. The model developed in this project will play a pivotal role in ensuring that pricing decisions align with business objectives, facilitating a balance between competitiveness, profitability, and customer satisfaction.

The success of the project will be measured through specific and measurable criteria. Firstly, the project aims to achieve a notable increase in revenue, leveraging optimized pricing strategies. The target percentage increase will be determined based on thorough analysis of historical data and current market conditions. Additionally, a key focus is on reducing customer churn by a predetermined percentage, ensuring that existing customers are retained even in the face of pricing adjustments. Market share growth is another quantifiable criterion, with the goal of attracting new buyers while maintaining a competitive stance within the housing market.

In addition to quantitative metrics, subjective criteria are integral to the project's success. Generating useful and insightful information into the relationships between pricing strategies, customer preferences, and market trends is a subjective yet critical goal. This will be assessed by a team consisting of senior management and data science experts. The project's adaptability to changing market conditions will be evaluated collaboratively by cross-functional teams, including marketing, sales, and research. Ethical considerations, crucial in the housing market, will be assessed by legal, compliance, and executive teams. Furthermore, the user-friendliness of the implemented pricing strategies will be gauged through user feedback and satisfaction surveys.

Inventory of Resources

Leading the House Prices Project is Alex Realty, serving as the Project Manager and Lead Data Scientist. Supporting the data-related aspects is Olivia Homefield, the Data Engineer, along with Ethan Estates overseeing database administration. Jordan Infrastructure provides IT support, and the project benefits from the analytical insights of market analysts Sofia Marketson, Max Property, and Lily Valuables. The data mining expertise comes from Derek Datascope, Laura Insights, and Ryan Regression. Business insights are provided by Emma Homewise and

Nathan RealtyPro, while statisticians Isaac Statsman, Chloe Chartwell, and Owen Analytics contribute their statistical knowledge. Victoria Legality heads the Legal and Compliance Team.

The computational infrastructure allocated for the House Prices Project encompasses a diverse set of hardware resources. These include access to an Excel CSV dataset sourced from Kaggle, coupled with the utilization of Microsoft Excel and Kaggle platforms for data processing. The hardware ensemble comprises a central processing unit (CPU) for robust computational capabilities. Additionally, a MacBook PC, equipped with a Linux operating system, forms an integral part of the hardware setup. The storage capacity is facilitated by a Local Disk C boasting over 1 TB of space, complemented by Random Access Memory (RAM) for efficient data handling. Cloud storage solutions are integrated, with Microsoft OneDrive facilitating seamless data storage and retrieval. The software ecosystem is enriched by the presence of Anaconda Navigator, featuring installations of Python 3 and Jupyter Notebooks to empower the project with advanced programming and analytical capabilities.

The data sources are the user sample data that were collected and included in Kaggle Inc. Examples of data sources included for this project are a test Excel CSV file with 443 KB, a train CSV file with 450 KB, and a sample submission CSV file with 32 KB.

Requirements, Assumptions, Constraints, And RESOLVEDD Strategy

To successfully achieve the projects' objective, it is crucial for all participants to acknowledge and adhere to requirements, assumptions, and limitations. The primary necessity for this undertaking would be the cooperation of all individuals involved well as a signed licensing agreement, from each participant who can be identified through the data. This agreement will highlight the importance of all parties avoiding any form of harm whether it be verbal or physical. It also emphasizes that no one should manipulate or tamper with any data

related to the project, a row. Furthermore, it states that the gathered data and private information should not be shared outside the project's scope. Lastly, it emphasizes the need for cooperation throughout the duration of the project. If any participant breaches any of these terms it will put the objective of this project at risk. They will be required to compensate for any damage caused to the sponsor, Kaggle. Any individual who is reported to be violating a condition or multiple conditions at once will be removed from the project for the duration, which could also jeopardize the project if crucial participants and executives responsible for evaluating the data are absent. All employees assigned to the data mining department are responsible for evaluating and managing the collected data and potential trends for this project. The goal is to create a model that can be easily understood by individuals. Data scientists are primarily responsible for evaluating and analyzing the data. However, it is the supervisor of the data science team who ultimately decides whether a model should be approved and directed to the CEO.

The assumptions towards this project include the following:

- The participants involved in this house prices project are assumed to predominantly belong to the age group between 25 and 60 years, as this demographic is likely to be actively engaged in the real estate market.
- It is assumed that the data obtained for house prices is accurate and available, with reliable records of property transactions. Any discrepancies or errors in the data are expected to be minimal, allowing for a robust analysis.
- The analysis assumes a stable economic environment without major fluctuations or crises that could significantly impact housing prices. Factors such as inflation, interest rates, and employment rates are considered relatively stable during the project duration. However, recession in 2008 is included.

- The project assumes that the housing market is competitive but not overly saturated, allowing for meaningful insights into price trends and influencing factors. Extreme competitiveness or lack thereof may affect the model's predictive accuracy.
- It is assumed that there are no major technological breakthroughs during the project timeline that could drastically alter the dynamics of the real estate market. Technological advancements that do occur are not anticipated to disproportionately affect housing prices.
- The project assumes that relevant data is readily accessible, subject to compliance with legal and ethical standards. Any restrictions on data accessibility are expected to be manageable.
- Assumptions are made regarding the necessity of presenting the model and its results to stakeholders, including senior management or sponsors. The assumption is that a clear and comprehensible presentation style will be effective in conveying the model's insights and recommendations.
- The cost assumption includes the affordability of tools and resources required for the analysis. For instance, any software or data sources necessary for the project are assumed to be within a reasonable budget, not exceeding predetermined cost limits.

These assumptions guide the entire duration of the project, ensuring that each phase of the analysis is aligned with the expectations outlined. Verification and reassessment of these assumptions are integral components of the data analysis process to maintain the accuracy and relevance of the insights generated.

The constraints that are to be addressed throughout the project include the following:

- The project needs to follow all the obligations concerning data collection, analysis, and reporting in the field of estate. This involves adhering to data protection laws, privacy regulations and any applicable local or national regulations regarding estate.
- The project operates within a predefined budget, and any additional costs, such as obtaining specialized datasets, tools, or hiring external expertise, need to be carefully considered. Overspending could impact the overall financial feasibility of the project.
- There is a strict timeline for completing the project, with the expectation of delivering actionable insights within a specified timeframe. This constraint necessitates efficient project management, data processing, and analysis to meet deadlines.
- The project team must manage computer resources effectively, considering limitations such as the capacity of Random Access Memory (RAM) and storage devices. Ensuring data integrity and preventing loss in case of system shutdowns or technical issues is crucial.
- Access to relevant data sources, including property listings, transaction records, and market trends, requires formal agreements between involved parties. Any restrictions, passwords, or authentication processes for accessing data must be addressed and documented.
- The project team must ensure compatibility with different operating systems, data management systems, and file formats to avoid complications during data

analysis. Compatibility issues could hinder the effectiveness of the data mining process.

- The project team acknowledges that certain knowledge or insights related to housing market trends may be limited. The scope of understanding is constrained to the information available in the provided dataset, and hidden clues or external resources may not be accessible.

These constraints are essential considerations throughout the project's lifecycle, guiding decision-making and ensuring that the analysis remains within legal, financial, and technical boundaries. The RESOLVEDD strategy, emphasizing ethical decision-making, may be employed to address any ethical dilemmas that may arise during the project.

Risks And Contingencies

The project faces the risk of increased market competition, where a competitor may introduce more effective pricing models or strategies. To counter this, a Housing Market Monitoring Team will be established. This team's role is to continually analyze competitors' strategies and promptly adapt the company's pricing models to maintain a competitive edge in the housing market.

A potential financial risk arises if the project's funding for further data mining depends solely on initial results. To address this, the project will diversify funding sources, collaborating closely with finance teams to secure alternative financial support and ensure the project's continued progression. The technical risk of poor-quality or incomplete data impacting pricing model accuracy is acknowledged. Contingency plans include the implementation of robust data cleansing processes, collaboration with data providers to enhance data quality, and the

establishment of validation checks to ensure data accuracy. Unforeseen technical challenges during the implementation of data science tools and models pose a potential risk. To mitigate this, a technical support team will be developed, regular training sessions will be conducted, and open communication channels with IT specialists will be maintained to address and resolve technical challenges promptly.

The risk of unexpected power or internet-related system outages that could disrupt data mining processes is recognized. Contingency plans involve coordinating with IT to establish wireless hotspots, saving regular checkpoints, and conducting IT assessments to maintain system stability and continuity. Potential risks related to privacy breaches and data integrity are identified. To address these concerns, integrity agreements will be drafted and enforced, signed by all project members. The legal team, led by Vijaya Gadde, will oversee compliance, ensuring strict consequences for any violations to safeguard data privacy and integrity. In the event of key personnel, such as data scientists, being absent due to personal reasons, a contingency team will be established. Project supervisors will oversee technical operations, and task-sharing strategies among team members will be implemented to ensure a reliable and flexible work environment, allowing the project to proceed seamlessly despite potential staff shortages.

Terminology

Business Terminology:

Property Assessment: The process of determining the worth of a property based on factors, like its location, size, condition, and market trends.

MLS (Multiple Listing Service): A platform used by real estate professionals to exchange information about properties detailing their features and prices.

Comparable Sales: Recently sold properties that share similarities with the property in question used to establish its market value.

Property Appraisal: An evaluation of a property's value carried out by an appraiser often required by lenders for mortgage approvals.

Asking Price: The initial price at which a property is listed for sale.

Market Research: An analysis of real estate market conditions and trends to assist in determining property values and pricing strategies.

House Inspection: A inspection of a property's condition conducted by an expert inspector to identify any issues impacting its value.

Initial Deposit: The payment made by a buyer when purchasing a property is usually calculated as a percentage of the purchase price.

Closing Expenses: costs incurred during the finalization of a real estate deal, including fees, title insurance and property taxes. The rate of interest set by a lender for a mortgage loan.

Data Mining Terminology:

Characteristic: Quality within a dataset utilized for making forecasts or identifying trends.

Regression Analysis: A technique that investigates the connection between a factor such as home cost and one or more independent factors like square footage, number of bedrooms.

Clustering: A data mining method that categorizes data points together based on attributes aiding in the recognition of trends in property categories or market segments.

Ensemble Learning: An approach in machine learning that merges forecasts from models to enhance accuracy and effectiveness.

Reduction of Dimensionality: Methods employed to diminish the quantity of variables within a dataset enhancing model efficiency and comprehensibility.

Feature Engineering: The procedure of generating characteristics or adjusting existing ones to enhance a model's forecasting capability.

Cross-Validation: A method for evaluating a model's performance by dividing the dataset into subsets training the model on some and testing it on others.

Outlier: An observation that significantly differs from data points in a dataset and may necessitate consideration during analysis.

Hyperparameter: A configuration setting external to a model that can be fine-tuned to optimize its performance.

Time Series Analysis: involves examining and forecasting patterns in data over time which could be valuable for gaining insights into how house prices change over periods.

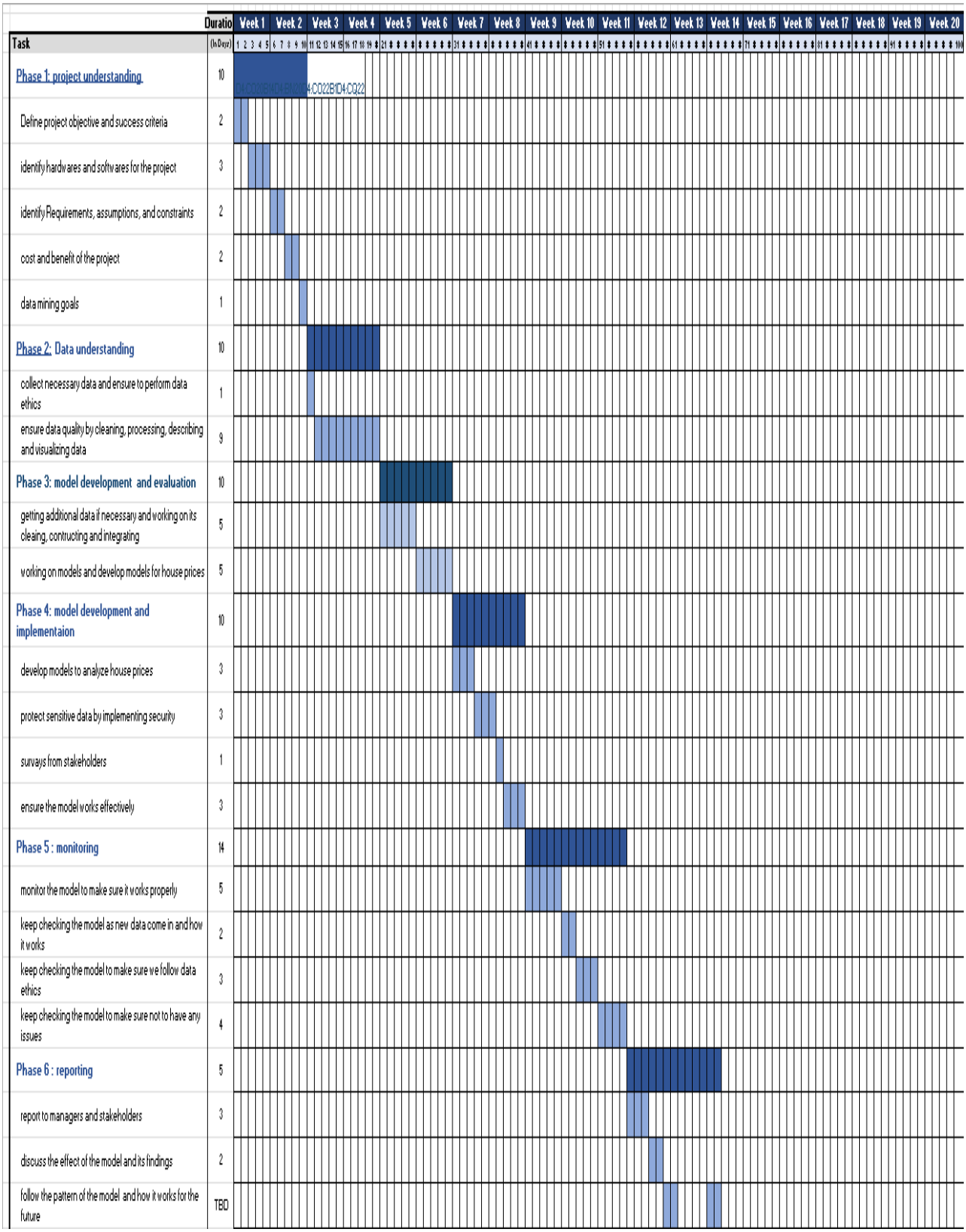
Data Mining Goals and Success Criteria

The primary business objective for our housing market analysis company is to enhance the accuracy of property price predictions. In business terms, this translates to providing clients with more precise forecasts regarding future property values, enabling informed decision-making for buying, selling, or investing in real estate. The corresponding data mining goal is to develop a predictive model that accurately estimates house prices based on various factors such as historical sales data, demographic information, and other relevant variables. This involves creating a sophisticated algorithm that can analyze patterns within the data and generate precise

predictions for property prices in the housing market. A specific data mining goal is to segment the housing market into distinct categories based on various criteria such as location, property type, and economic indicators. This segmentation will facilitate targeted insights for different market segments, enabling our clients to tailor their strategies based on specific market dynamics. The primary problem type for this data mining project is prediction. The goal is to predict future house prices accurately, allowing stakeholders to make well-informed decisions about their real estate transactions.

The success criteria for the project include achieving a high level of predictive accuracy in the model, ensuring that the estimated house prices closely align with the actual market values. Additionally, the model's performance in terms of speed and efficiency will be crucial to ensure real-time applicability in the dynamic housing market. Defined benchmarks will serve as key indicators for success, comparing the model's performance against industry standards. This includes evaluating the model's accuracy against established metrics and benchmarks to validate its effectiveness and reliability. Subjective criteria, such as model explainability and the insights provided, are essential for a holistic evaluation. The success of the project relies not only on accurate predictions but also on the model's ability to provide meaningful insights into the factors influencing house prices. Considering deployment from the project's inception is crucial. Success is not solely measured by accurate predictions but also by the practical implementation of the model in real-world scenarios. Deployment planning activities will involve ensuring seamless integration into existing systems and processes.

Project Plan/ Order of Tasks (Insert Gantt Chart)



Data Understanding

Initial Data Collection Report

The project utilized datasets that were acquired and accessed from Kaggle using Excel. The dataset collection comprises three files: a training dataset in CSV format, a testing dataset in CSV format, and a sample dataset in CSV format. These files form the basis for the analysis and development of models supplying the required data for training and testing models. The dataset for this project encompasses a range of features related to real estate or property values. The primary target variable is "SalePrice," denoting the property's sale price in dollars. This variable serves as the focal point for prediction. The dataset consists of both numerical and categorical features, providing a comprehensive view of various aspects related to properties. Numerical features include details such as building class ("MSSubClass"), linear feet of street connected to the property ("LotFrontage"), lot size in square feet ("LotArea"), original construction date ("YearBuilt"), and many others. These features capture quantitative aspects, from the dimensions of the property to the number of rooms and bathrooms, providing a rich set of data for analysis. On the categorical side, features like "MSZoning" (general zoning classification), "Street" (type of road access), and "Neighborhood" (physical locations within Ames city limits) offer insights into the qualitative characteristics of properties. These categorical attributes help to categorize and differentiate properties based on various criteria, such as location, zoning regulations, and architectural styles.

Some attributes may hold particular significance in influencing property prices. For example, features like "Neighborhood" and "OverallQual" (overall material and finish quality) may be of substantial importance. Understanding the relative importance of these attributes is crucial, and consultation with domain experts or stakeholders can aid in prioritizing certain

features over others in the analysis. A critical consideration in the data collection process is assessing the quality of both individual data sources and the merged dataset. Inconsistencies between sources may lead to challenges do not present in individual datasets. Therefore, a thorough evaluation of data quality, including aspects like consistency, completeness, and correctness, is essential. Data cleaning and preprocessing steps may be necessary to address any identified issues. The combination of numerical and categorical features, coupled with an understanding of attribute importance and data quality considerations, forms the basis for informed analysis and modeling in the real estate domain.

Data Description Report

File 1: test.csv

The dataset acquired from Kaggle, accessed through Excel, contains 1459 rows and 80 columns. The variables that are included in this file cover all aspects that affect house prices such as alley, street, area, quality of different part of a house and many more. The id numbers range from 1461 to 2919 and it is a numerical identifier for each house. Also, it has numerical patterns. The strongest positive correlation in this file goes for GarageArea and GarageCars which is (0.90). On the other hand, there is no strong negative correlation for this file, and we can see lack of correlation for some variables. There are missing values over 1000 for Alley, PoolQC, Fence, and MiscFeature. However, these variables are not effective for target variables. SalePrice is our target variable which we are trying to predict, and it is not included in this file. LotArea has the most outliers.

File 2: train.csv

The train csv file, which was also accessed through Excel from Kaggle, contains 1460 rows and 81 columns. All variables are the same except SalePrice column which added to this dataset as it is target variable. The Id value, which ranges from 1 to 1460, follows a specific numerical pattern. This file includes target variable. Given the significance of each row and column in the utilization and analysis of the project, it is imperative not to remove any rows containing null values. Additionally, it is crucial not to delete any data that could contribute to the project's foundation. Instead, the recommended approach is to strike a balance between accuracy and insights from each variable and the information content within those variables. To address cells with null values, they will be replaced with mean, mode, or median. This approach ensures that valuable information is retained and considered, fostering a comprehensive analysis.

Data Exploration Report

The first action towards this section of exploring the data was creating a heatmap (Figure 3) through Jupyter that analyzed correlation for selected variables. This process, which focuses on seven variables (OverallQual, GrLivArea, GarageCars, GarageArea, SalePrice, 1stFlrSF, TotalBsmtSF) was made to gain correlation between variables and analysis of the amount we gain. The project's objective is focused towards identifying SalePrice which is target variable. In this heatmap, we can see that SalePrice has a strong positive relationship with OverallQual (0.79). It is obvious that OverallQual, which contains Overall material and finish quality has a linear relationship with SalePrice. This is the only variable which has a relationship with SalePrice. Moreover, GarageCars and GarageArea have a strong positive relationship (0.88). These variables must have a strong positive relationship together since we need more space for more cars. Also, that is the strongest relationship in this dataset.

Figure 3

Correlation Heatmap for Selected Variables

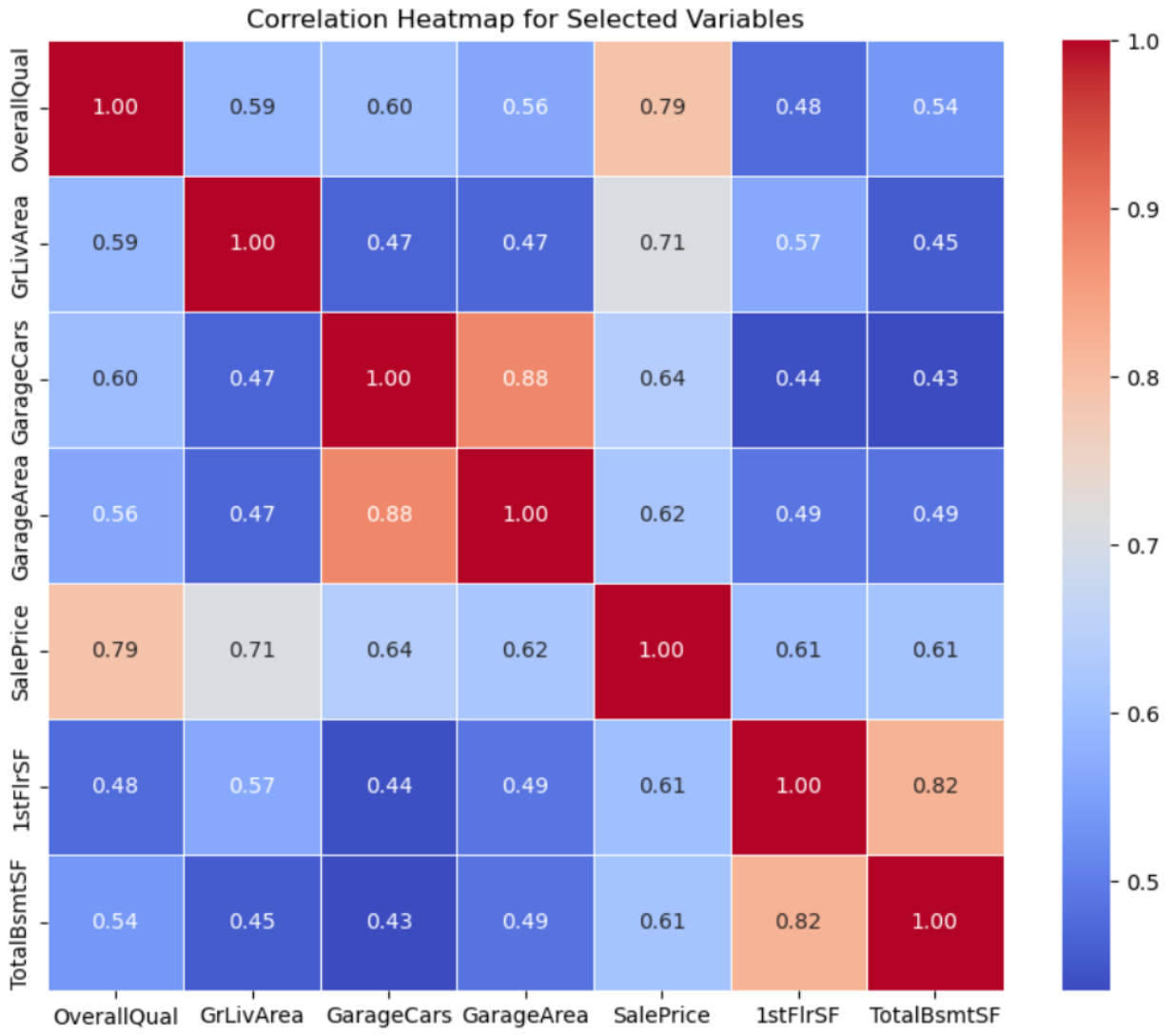
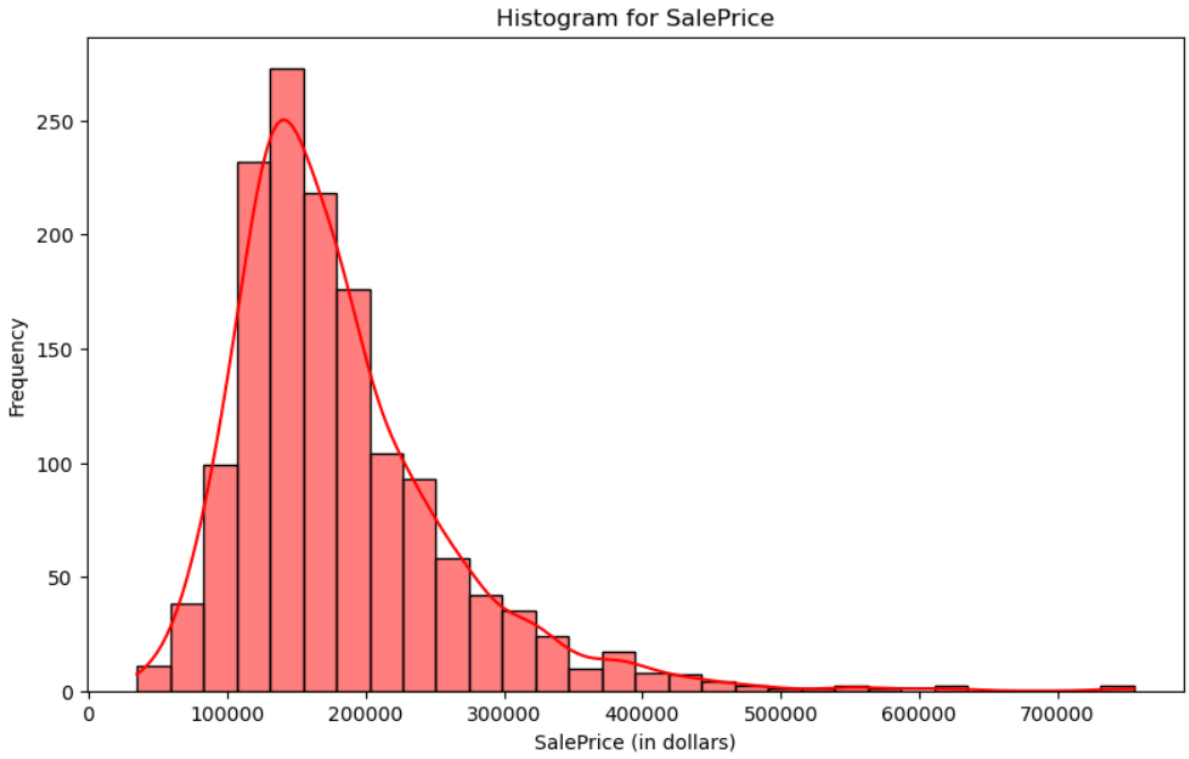


Figure 4 is the next action toward exploring data through Jupyter which analyzes SalePrice frequency. The histogram displays how prices are distributed within a dataset. On the x axis you can see the sale prices in dollars while the y axis represents the frequency of sales at each price point. The sales distribution is skewed towards the right indicating that there are sales occurring at lower prices compared to higher prices. In other words, there are some outliers in higher prices. The majority of sales fall within the range of $100,000 to $200,000. However,

there are instances where sales have been recorded with prices exceeding $500,000. Sales of items priced over $500,000 are not frequently encountered, they can yield profits for the company. On the other hand, sales of items priced below $50,000 are more commonplace though they might have profit margins.

Figure 4

Histogram for SalePrice



Data Quality Report

The train.csv dataset provides every variable and value needed to establish a machine learning algorithm. The dataset contains all aspects that need to be considered for a target variable such as SalePrice. The Id column is the unique column which represents primary key correctly and follows a precise numerical order. All columns are expressed in the correct unit and

make sense within the context of the dataset. However, there are concerns regarding the quality of the gathered data that must be resolved to guarantee the error performance of the ultimate model. Among the categorical variables, the Alley column is the only categorical variable which has 1369 missing values. However, it is not a main variable to affect target variable. There are some missing values within the dataset under Masonry veneer area in square feet, Walkout or garden level basement walls, Quality of second finished area, Year garage was built, Height of the basement, Quality of basement finished area, Interior finish of the garage, General condition of the basement, Garage location, Garage quality, Garage condition variables. These variables do not have many missing values and they can be replaced by mean, median, or mode. On the other hand, some variables like Linear feet of street connected to property, Type of alley access, Masonry veneer type, Fireplace quality, Pool quality, Fence quality, Miscellaneous feature not covered in other categories have too many missing values. One possible solution to address this problem is to assign the values a label of "Unknown." This way the other related variables and data rows can still be utilized for information without being removed.

Reference

https://www.statista.com/statistics/240991/average-sales-prices-of-new-homes-sold-in-the-us/

Fan, W., He, Y., Hao, L., & Wu, F. (2024). Do high house prices promote the development of China's real economy? empirical evidence based on the decomposition of real estate price. *PLOS ONE*, *19*(1). https://doi.org/10.1371/journal.pone.0295311