Project Proposal & Outline

CSP 571 Data Preparation and Analysis

Project Group Members

- Abhiram Ravipati – A20539084
- Sumanth Kalyan Bandigupthapu – A20544342

1. Project Proposal

1.1 Description:

Our project focuses on analyzing real estate data to gain insights into market trends, property characteristics, and historical sales data. The dataset includes information about the real estate agency or individual responsible for property sales, the current status of property listings, listed prices, bedroom and bathroom counts, property size, address specifics such as street name, city, state, and postal code, as well as the size of the house and the date of the property's previous sale. Through meticulous analysis, we aim to uncover patterns and correlations within the data, providing valuable insights for decision-making in the real estate industry.

1.2 Research Goal:

The primary goal is to understand the dynamics of the US real estate market and explore factors influencing property prices and market demand. This analysis will provide valuable insights for stakeholders such as investors, real estate developers, policymakers, and homebuyers to make informed decisions.

1.3 Research Questions:

1) What is the distribution of property prices across different cities and states within our dataset?

2) How do the number of bedrooms and bathrooms correlate with property prices in the real estate market?
3) Is there any relationship between the size of the house and its listing price?
4) Do properties with larger acre area tend to demand higher selling prices?
5) How does the previous sold date impact the current listing price of properties?

1.4 Proposed Methodology:

1.4.1 Data Collection:

- Gather the real estate dataset from reliable sources such as Zillow, Realtor.com, or public databases.
- Ensure the dataset includes relevant attributes such as property prices, housing inventory, demographic factors, and economic indicators.

1.4.2 Data Cleaning:

- Removal of duplicates within the dataset to ensure data integrity.
- Handle missing values in columns such as price, number of bedrooms, bathrooms, and acreage.
- Address outliers and inconsistencies that may cause issue in analysis results.
- Perform data transformation and normalization to prepare the dataset for analysis.

1.4.3 Exploratory Data Analysis (EDA):

- Conduct a comprehensive exploration of the dataset to understand its characteristics and distributions.
- Visualize the distribution of property prices across different cities and states using histograms, box plots, or heatmaps.
- Analyze the relationships between key variables such as bedrooms, bathrooms, house size, and property prices using scatter plots and correlation matrices.
- Identify any trends or patterns in the data that may inform further analysis.

1.4.4 Feature Engineering:

- Creating new features such as total number of rooms (bedrooms + bathrooms) or price per square foot.

- Transform categorical variables into numerical representations using techniques like one-hot encoding.
- Engineer features that capture the temporal aspects of the data, such as age of the property based on previous sold dates.

1.4.5 Model Building:

- Split the dataset into training and testing sets to evaluate model performance.
- Select appropriate machine learning algorithms based on the nature of the problem (regression or classification).
- Train regression models using techniques such as linear regression, decision trees, random forests.

1.4.6 Conclusion and Recommendations:

- Summarize the key findings from the analysis, including significant insights and trends observed in the real estate market.
- Providing actionable recommendations for stakeholders based on the analysis results, such as areas for investment or factors to consider when pricing properties.
- Discuss the limitations of the analysis and suggesting potential areas for future research or data collection.

1.5 Metrics to Measure Analysis:

- Evaluate the performance of the trained models using relevant metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared value for regression analysis.
- Assess the accuracy, precision, recall, and F1-score for classification models.
- Compare the performance of different models to identify the most effective approach for predicting property prices.

2. Project Outline

2.1 Literature Review and Related Work:

- Choy, Lennon & Ho, Winky. (2023). The Use of Machine Learning in Real Estate Research. 12. 740. 10.3390/land12040740.

In this Paper they discuss how computer algorithms, called machine learning, are changing the way we study real estate. They explain how these algorithms can help predict property prices, understand market trends, and make better investment choices. The authors also talk about important things to consider, like how to pick the right data and make sure the predictions are fair and accurate. Their paper is a helpful guide for anyone interested in using technology to improve real estate research and decision-making.

- Gale, H., Roy, S.S. Optimization of United States Residential Real Estate Investment through Geospatial Analysis and Market Timing. Appl. Spatial Analysis 16, 315–328 (2023). Gale and Roy (2023) explore how to make smart decisions when investing in houses in the United States. They use maps and information about when to buy and sell to figure out the best times and places to invest. By studying these factors, they help investors understand where and when they can make the most money from buying and selling homes.
- Kumar, Sailaja, Kameshwari Soundarya, Harshitha R., Evangelin Geetha D., and T.V. Suresh (2018) conducted a study on analyzing real estate data using principal component analysis (PCA) and the programming language R. The researchers applied PCA, a statistical technique, to identify the most important factors influencing real estate trends. By utilizing R, a popular tool for statistical computing, they were able to efficiently analyze and visualize the data. Their findings provide insights into understanding the key drivers of real estate dynamics, offering valuable information for stakeholders in the industry.

2.2 Data Sources: (Dataset is obtained from Kaggle)

https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data

2.3 Dataset Description:

S. No	Column Name	Column Type	Description
1	brokered_by	num	This column identifies the real estate agency, broker, or individual responsible for managing

			the sale of the property. It indicates who is facilitating the transaction between the seller and potential buyers.
2	status	chr	This column indicates the current status of the property listing. For example, "for_sale" means the property is currently available for purchase, while other statuses like "ready_to_build" would indicate that the property has already been sold.
3	price	num	This column displays the listed price of the property in US dollars. It represents the amount that the seller is asking for the property.
4	bed	int	The number of bedrooms in the property. Bedrooms are defined spaces within a house intended primarily for sleeping, typically containing a bed and often other furnishings like dressers or closets.
5	bath	int	The number of bathrooms in the property. Bathrooms are rooms containing a bathtub or shower, toilet, and sink, used for personal hygiene activities such as bathing and toileting.
6	acre_lot	num	This column represents the size of the property in acres. An acre is a unit of area commonly used in real estate to measure the size of land.
7	street	num	The name or identifier of the street where the property is located. It specifies the physical address or location of the property within a city or town.
8	city	chr	The name or identifier of the street where the property is located. It specifies the physical address or location of the property within a city or town.
9	state	chr	The name or abbreviation of the state where the property is located. It indicates the specific region or jurisdiction within a country where the property is situated.
10	zip_code	int	The postal code or ZIP code of the area where the property is located. ZIP codes are numerical codes used by postal services to facilitate the sorting and delivery of mail within a specific geographic area.

11	house_size	num	This column displays the size of the house in square feet. It represents the total interior living space of the property, including all rooms and levels.
12	prev_sold_date	chr	This column indicates the date when the property was previously sold. It provides historical information about the property's sales history, including when it was last transferred from one owner to another if applicable.

The dataset contains 1,048,576 rows of property information. It includes details such as the entity responsible for brokering the sale, the current status of the property, listed price, number of bedrooms and bathrooms, property size, street name, city, state, ZIP code, house size, and the previous sold date. Additionally, there are missing values present in some entries that are needed to be addressed during analysis.

2.4 Software Packages and Tools:

- R Programming
- RStudio
- R Packages: ggplot2, glm, caret, shiny, dplyr, tidyverse, leaflet, caret, randomForest, etc.,.