King County, USA

Real Estate Price Prediction and Analysis

Project Objective:

The primary goal of this project is to analyze factors that influence real estate prices and build predictive models that estimate house prices based on various property features. By understanding these factors, the project aims to provide insights that could benefit potential home buyers, real estate professionals, and investors in making informed decisions about property investments. Specifically, the project will focus on identifying key drivers of property prices, exploring patterns and trends in house sales, and using predictive model techniques to predict house prices based on historical data.

Context:

The real estate market is highly dynamic, influenced by a variety of factors such as property size, location, condition, and local economic conditions. Predicting house prices accurately is valuable not only for real estate agents and homebuyers but also for urban planners, real estate developers, and policy makers. In a competitive market, small changes in property attributes can lead to significant fluctuations in price, making it important to quantify these impacts.

The dataset used for this project comes from Kaggle's House Sales in King County, USA dataset, which includes house sales from King County (Seattle), Washington. The dataset contains over 21,000 entries with features ranging from basic characteristics like the number of bedrooms and bathrooms to more complex attributes like the presence of a waterfront or recent renovations.

Key Questions:

1. What factors most significantly impact price of a property?
2. Which variables (e.g., square footage, number of bedrooms, year built, etc.) have the highest correlation with property prices?
3. Can we build an accurate predictive model for estimating house prices?
4. Which machine learning algorithms (e.g., Linear Regression, Decision Trees, Random Forest, etc.) perform best in predicting house prices?
5. Are there seasonal effects or trends across the one-year sales data?
6. Do properties with recent renovations have higher price tags?

Data:

The dataset used for this project is publicly available on Kaggle and contains house sale prices for King County, including Seattle. It includes houses sold between May 2014 and May 2015. The real estate dataset contains 21,613 records, each representing a property sale in King County, Washington.

https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data

Explanation of Variables

Variable	Explanation	Data Type
id	Unique ID for each home sold	integer
date	Date of the home sale	datetime
price	Price of each home sold	float
bedrooms	Number of bedrooms	integer
bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower	float
sqft_living	Square footage of the apartments interior living space	integer
sqft_lot	Square footage of the land space	integer
floors	Number of floors	float
waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not	integer
grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.	integer
sqft_above	The square footage of the interior housing space that is above ground level	integer
sqft_basement	The square footage of the interior housing space that is below ground level	integer
yr_built	The year the house was initially built	integer
yr_renovated	The year of the house’s last renovation	integer
zipcode	What zipcode area the house is in	integer
lat	Lattitude	float
long	Longitude	float
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors	integer
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors	integer
view	An index from 0 to 4 of how good the view of the property was: 0 = No view, 1 = Fair 2 = Average, 3 = Good, 4 = Excellent	integer
condition	An index from 1 to 5 on the condition of the apartment: 1 = Poor- Worn out, 2 = Fair- Badly worn, 3 = Average, 4 = Good, 5= Very Good	integer

Data Profile

Consistency checks were performed to verify the absence of duplicates, mixed-type data, missing values, or outliers that are inconsistent with the context of the dataset. The following adjustments were made:

- The **'date'** column's data type was updated to datetime format.

- Checked descriptive statistics of all the variables and made sure that the data is within the acceptable range for all the columns.
- Extracted the year, month from date column and created separated columns for year and month.
- Created new column as apartment_grade using the grade variable.
- created separated column for house_view using the view variable.
- Created a new column as house_condition using the data in condition column.
- checked for missing values, duplicate rows and outliers.

Data Limitations:

The data is from 2014-2015 which means that insights gleaned from an analysis on this data may not reflect current housing price trends. Additionally, with limited geographical coverage, it restricts the ability to generalize findings to other regions.

Ethical Considerations:

Although the dataset is shared under a Public Domain license, granting permission for its use and distribution, there is limited information regarding the original collection and handling of the data. This raises concerns about whether informed consent was obtained from individuals whose data is included. While the open license allows for its ethical use in research and analysis, it leaves unanswered questions about the dataset's origin.

Potential biases, such as collection bias, may exist because the dataset's scope and methodology are unclear. Without transparency, it's difficult to know if certain factors were overemphasized or omitted based on the collector's assumptions. For instance, if the dataset predominantly includes homes sold within a specific price range or focuses on single-family homes but excludes other types like condos, it might not fully represent the housing market. Such biases could distort the conclusions drawn from the analysis.

Approach:

Data Exploration and Cleaning:

- Handle any missing values, incorrect data types, and outliers.
- Explore the distribution of key variables and visualize correlations.
- Create new features (if needed) such as age of the property or price per square foot.

Exploratory Data Analysis (EDA):

- Analyze how each feature impacts property prices.
- Use visualization tools like scatter plots, heatmaps, bar charts, box plots etc. to uncover patterns in the data.

Modeling:

Depending on the nature of data, appropriate predictive model will be selected to predict future house prices.

Interpretation and Insights:

This project aims to deliver a deeper understanding of the real estate market in King County, Washington, by identifying key factors that influence property prices and building accurate predictive models. The analysis will provide valuable insights for both real estate investors and potential homeowners, helping them make data-driven decisions when purchasing or pricing properties. It will also offer recommendations for home buyers or real estate professionals based on findings.