

**BUAN 6356.004**

**Project Report: Project Report**

**SUBMITTED BY**

**TRUNG NGUYEN – TBN140130**

**SUBMISSION DATE: 12/3/2023**

**Summary:**

This study focuses on the analysis of typical house prices in Paris. Our project, which predicts Paris housing prices, uses computer analytics to understand what affects house prices in the city. We are looking for details. For example, house size, number of rooms, and square footage. Floor, city. Location, age of homes, and other factors can determine a home's value.

We want to help people better understand the Paris housing market, whether they are agents, landlords, or looking to buy. We aim to make it easy for everyone to know what is what when buying or selling a house in Paris. By studying these data, we aim to discover hidden patterns of house price performance in Paris. For example, we want to know if new homes or homes near major landmarks are more expensive. We aim to explain why homes are expensive, not just reflect that price. We seek to simplify how the Paris housing market works for professionals and individuals looking for a new home.

The analysis's primary goal is to identify the spatial disparities that affect housing costs in the city. Accessibility to public transportation, the atmosphere of the surrounding communities, and proximity to famous landmarks all become essential factors determining the price of residential homes. By closely examining these geographical patterns, we hope to provide stakeholders with a more nuanced picture of the local dynamics guiding the city's property values' undulating waves. Our work aims to uncover the real meaning behind home prices. We use computer tools to categorize homes into types, like basic or luxury. We dissect data and use sophisticated analytics. We do it; we make sure. The ranking is perfect. It is easy. He has an understanding of the city's housing market.

This analysis goes beyond statistical correlations and numerical values; it is essentially an investigation of Paris's economy, with data serving as our guide to interpreting the stories expressed in housing prices. We want to give stakeholders a road map for navigating Paris's varied and dynamic real estate landscape by providing a thorough insight that goes beyond the surface as we navigate the dataset's many channels

**Project Motivation:**

The motivation behind undertaking a comprehensive analysis of Paris housing pricing data stems from the intrinsic allure of the Parisian real estate market and the desire to unearth actionable insights for various stakeholders. Paris, often regarded as one of the most enchanting and culturally rich cities globally, boasts a real estate landscape that mirrors its diverse and dynamic character.

This project seeks to delve into the wealth of information contained within the Kaggle dataset to provide a nuanced understanding of the factors influencing housing prices in this iconic city.

Paris is renowned for its distinct arrondissements, each with unique charm and character. Investigating the spatial differentials in housing prices allows us to unravel the complex interplay of factors contributing to property valuations.

We aim to identify the key features that significantly impact housing prices by examining the dataset. Insights into the importance of factors such as the number of bedrooms, square footage, and proximity to public transportation can contribute to predictive models and a deeper understanding of buyer preferences and market trends. This knowledge is invaluable for real estate professionals seeking to tailor their offerings to meet evolving market demands.

Temporal Analysis for Informed Decision-Making: The temporal dimension of the data provides a historical narrative of housing price trends over time. Unravelling the temporal dynamics allows us to identify patterns influenced by economic shifts, seasonal variations, and broader market trends. This temporal analysis enhances our predictive capabilities and empowers stakeholders with the foresight needed to make strategic decisions in the ever-evolving Parisian real estate market.

Practical Applications for Stakeholders: The findings from this analysis have practical implications for a range of stakeholders. Investors can identify opportunities and mitigate risks based on spatial and temporal trends. Policymakers can use the insights to inform urban development strategies and housing policies that address affordability challenges in specific areas. Real estate professionals can tailor their marketing strategies to align with the preferences driving housing prices.

In essence, this project is motivated by the desire to uncover the intricacies of the Parisian real estate market, offering a data-driven narrative that goes beyond numerical values to capture the essence of what drives housing prices in this culturally rich and diverse city.

**Data description**

The project's data set was taken from Kaggle.The data set's main attribute is the price, which provides us with the estimated cost of housing in Paris. The data set contains ten numerical and eight categorical attributes. The data set includes housing data with multiple variables impacting the cost of housing in Paris; the data can be utilized to identify the significant variables affecting the cost of housing.

**Attributes used in the dataset:**

1. **Id:**
   * Type: Numeric (or String, depending on the format)
   * Description: A unique identifier for each record in the dataset serves as a reference to distinguish one property from another.
2. **squareMeters:**
   * Type: Numeric
   * Description: The total area of the property is in square meters, indicating the size of the living space.
3. **numberOfRooms:**
   * Type: Numeric (Integer)
   * Description: The count of rooms in the property, encompassing bedrooms, living rooms, and other functional spaces.
4. **hasYard:**
   * Type: Binary (0 or 1)
   * Description: A binary indicator denoting whether the property includes a yard providing outdoor space.
5. **hasPool:**
   * Type: Binary (0 or 1)
   * Description: A binary indicator indicating whether the property is equipped with a pool.
6. **floors:**
   * Type: Numeric (Integer)
   * Description: The number of floors in the property indicates its vertical structure.
7. **cityCode:**
   * Type: Categorical (String or Numeric)
   * Description: The zip or city code representing the property's location.
8. **cityPartRange:**
   * Type: Ordinal (Integer)
   * Description: An ordinal variable indicating the exclusivity or prestige of the neighborhood, with higher values corresponding to more exclusive areas.
9. **numPrevOwners:**
   * Type: Numeric (Integer)
   * Description: The number of previous owners of the property, providing insights into its history.
10. **Made:**
    * Type: Numeric (Year)
    * Description: The year the property was constructed, indicating its age and potential impact on condition and value.
11. **isNewBuilt:**
    * Type: Binary (0 or 1)
    * Description: A binary indicator denoting whether the property is newly built.
12. **hasStormProtector:**
    * Type: Binary (0 or 1)
    * Description: A binary indicator indicating whether the property has storm protection measures.
13. **basement:**
    * Type: Numeric
    * Description: The basement is square meters, providing additional usable space.
14. **attic:**
    * Type: Numeric
    * Description: The size of the attic is in square meters, indicating potential additional living space.
15. **garage:**
    * Type: Numeric
    * Description: The size of the garage, providing information on parking space.
16. **hasStorageRoom:**
    * Type: Binary (0 or 1)
    * Description: A binary indicator denotes whether the property has a storage room.
17. **hasGuestRoom:**
    * Type: Numeric (Integer)
    * Description: The number of guest rooms in the property, providing information on additional accommodation space.
18. **price:**
    * Type: Numeric
    * Description: The predicted value of the property serves as the target variable for analysis and prediction.

The values obtained from the dataset can be analysed using algorithms for segmentation, association, prediction, classification, and classification because it includes both continuous and numerical data.

**Project Methodologies:**

The software used for this project is Python, and Anaconda IDE was used as we found it convenient. The software leverages various powerful tools and libraries for efficient data cleaning and manipulation. With its extensive ecosystem, Python employs popular libraries such as Pandas for data wrangling, allowing seamless handling of missing values, outlier detection, and feature engineering. Pandas provides versatile data structures that facilitate the organization and manipulation of housing data, ensuring it is well-prepared for subsequent analysis. Additionally, NumPy is employed for numerical operations, enabling efficient computations across large datasets. Python's sci-kit-learn library offers a rich model selection and evaluation tool suite, encompassing various regression algorithms crucial for predicting housing prices. The versatility of Matplotlib and Seaborn allows for the creation of insightful visualizations, aiding in the exploration of spatial and temporal patterns within the dataset. Altogether, Python's cohesive ecosystem provides an integrated and efficient environment for cleaning and analysing housing data, making it a preferred choice for predictive modelling in real estate.

**Data cleaning:**

The data set used in the project has 22,730 records and 18 attributes. Some of which are null values and are inconsistent data. To maintain data integrity, we clean the data before starting the analysis.

The dataset is cleaned using the procedures listed below.

1) import csv file

2) check for nulls

3) changing categorial data into dummies numeric

4) drop the unneeded data

5) save to new csv named

**BI Model:**

**Exploratory**

**Checking for Null Values:** Null values might negative affect the correctness of the predicting model. As a result, our team has checked for the possibility of having nulls values and figured a way to eliminate the possible interference in the future.

**A screenshot of a computer code

Description automatically generated**

**Overview the data via boxplot and histogram for a better understanding of the data.**

#1st look at boxplot

A group of blue squares

Description automatically generated

#1st look at histogram

A group of graphs and diagrams

Description automatically generated with medium confidence

**Result**: Some of the diagrams have shown problematic such as hasYard, hasPool, isNewbuilt, hasStormProtector and hasStorageRoom. Under a deeper investigation, the nature of those data is recorded as 1 and 0. As a result, histogram and boxplot have shown an unnatural diagram.

**Additional step** – looking into correlation among variables.

#1st look at the correlation

A graph of numbers and a graph

Description automatically generated with medium confidence

**Result:** Looking into the heatmap, we can obviously notice the strong connection between price and squareMeters, however the other variables tend to have negligent connection to determine the price of a property.

**Modeling**

In order to create a predicting model, our team has devided the dataset into x\_train, y\_train and validation dataset with x\_test, y\_test. It is reasonable to have a test size of 20% before launching a full scale.

**Independent Variables** will contain: squareMeter, numberOfRooms, hasYard, hasPool, floors, cityPartRange, numPrevOwners, made, isNewBuilt, hasStormProtector, basement, attic, garage, hasStorageRoom, hasGuestRoom.

**Dependent variable** will be the price of the property.

#Check for histogram

A group of blue and white graphs

Description automatically generated

#Check for correlation

A graph of numbers and a chart

Description automatically generated with medium confidence

**Normalization**: as the scale of the variables are largely different, our team have decided to normalize the dataset with log function, so that all independent variables will be considered with similar scale, leading to increasing the accuracy of the model.

A screenshot of a computer code

Description automatically generated

#Check data again

A group of blue and white graphs

Description automatically generated

#Recheck correlation

A graph of numbers and a graph

Description automatically generated with medium confidence

The normalization process has significantly highlighted the relationship between squareMeters and price by increasing correlation from 0.55 to 0.86.

**Choosing predicting method**

* 1. Linear regression

Linear regression is the first model that our team tried to predict the price as it is simplicity.

A close-up of a check

Description automatically generated

For the linear regression, we have generated a model that predicts the data with 73.4% accuracy.

* 1. RandomForest

Additional technique could be used is using randomForest. In this method, there are many models will be generated and the weighted average will help to enhance the accuracy of the model.

A screenshot of a computer

Description automatically generated

With the model that we have increased the accuracy to 99.4%.

**Our findings and managerial implications/conclusions:**

Our examine delved into the Paris housing marketplace and discovered several key elements affecting housing prices: house size, room count, square footage, floor level, location, age of homes, proximity to landmarks, and accessibility to public transport.

We used diverse statistical methods to determine the values. Traditional regression gave us a decent rating of 73.41%, however when we switched to random forest regression, the accuracy jumped to an excellent 99.40%. This suggests that the random forest regression approach captures the connection between housing and rate in Paris thoroughly.

Therefore, the usage of random forest regression is the nice method for predicting housing costs out there. It is going past the traditional method and gives marketers, landlords, or customers with a more reliable device for correct pricing.

*Here are managerial implications:*

**Better choice making**: Agents and owners can use this model to make smarter pricing choices.

**Better verbal exchange**: Buyers and sellers can better apprehend the impact of housing expenses, resulting in higher communique.

**Understanding the market**: This observe enables all people better understand the Paris housing market.

**Identifying the first-rate investments**: Successful fashions help discover worthwhile investment possibilities within the marketplace.

**Planning contribution**: Planners can use this observe to optimize housing techniques for distinct regions of Paris.

In quick, the usage of random forest regression is a robust and correct approach for forecasting Parisian housing prices. It can help all stakeholders recognize the determinants of real estate inside the city.

**References:**

* 1. *Data Mining for Business Analytics: Concepts, Techniques,and Applications in R*, by Galit Shmueli, Peter Bruce, Inbal Yahav, NitinPatel, and Kenneth Lichtendahl. Wiley, ISBN-10: 1118879368, ISBN-13:978-1118879368
  2. Kaggle. (n.d.). Paris Housing Data. Retrieved from <https://www.kaggle.com/datasets/mssmartypants/paris-housing-price-prediction>