PHASE 2 GROUP PROJECT

START

GROUP MEMBERS

01 FIDELIS WANALWENGE

02 SYLVIA MURITHI

03 KHADIJA ALI

04 ELIZABETH NYAMBURA

PROJECT OUTLINE

Project overview

In this project, we will make use of everything we have learned about pandas, data cleaning, and exploratory data analysis. We will employ multiple linear regression modeling to analyze house sales data in the King County area. By using statistical techniques, we aim to identify key factors that impact property sales in the region and provide valuable insights to guide our recommendations.

Additionally, we will explore the potential of remodeling specific areas of a property to increase its value and make it more attractive to potential buyers. This approach can help homeowners add functionality and beauty to their property while simultaneously boosting its resale value.

Objectives

1. To create a complex model using several different independent variables that can swifty and effectively achieve pricing estimates closer to realized housing prices

2. To evaluate different models that ultimately lead to selecting our best model for predicting house prices

3. To provide insight on house features that have the biggest impact on sale price

Business Understanding

This project aims to provide valuable insights for a real estate agency operating in King County, Washington, USA. Specifically, the agency seeks to provide accurate advice to homeowners on how home renovations can potentially increase the estimated value of their properties and homes, and by how much. This information will help the agency guide their clients towards making informed decisions on home renovations, which can maximize their return on investment when selling their properties


SOLD

The problem at hand is to provide homeowners with accurate advice on how specific home renovations can impact the estimated value of their properties and homes, and the amount by which it can increase. This information is crucial for the real estate agency to guide their clients towards making informed decisions on home renovations, which, in turn, can help homeowners to maximize their return on investment when selling their properties.

Therefore, the primary objective of this project is to analyze the impact of home renovations on the estimated value of properties and provide recommendations that can help the real estate agency and their clients to make sound investment decisions.

Data Understanding

The King County House Sales dataset, available in "kc_house_data.csv," is the primary data source for this project. However, one of the main challenges we may encounter is the ambiguity or incompleteness of the column names in the dataset. Nonetheless, with thorough research and careful judgment, we can extract the necessary insights to make informed decisions about which variables to use in our analysis.

The dataset contains information on house sales in King County, including the price, design, square footage, location, and more. A comprehensive list of the column names can be found in the "Property Schema".

We will also explore the general areas in which renovations are typically undertaken in properties to identify the potential impact on the estimated value of a property.

Modelling

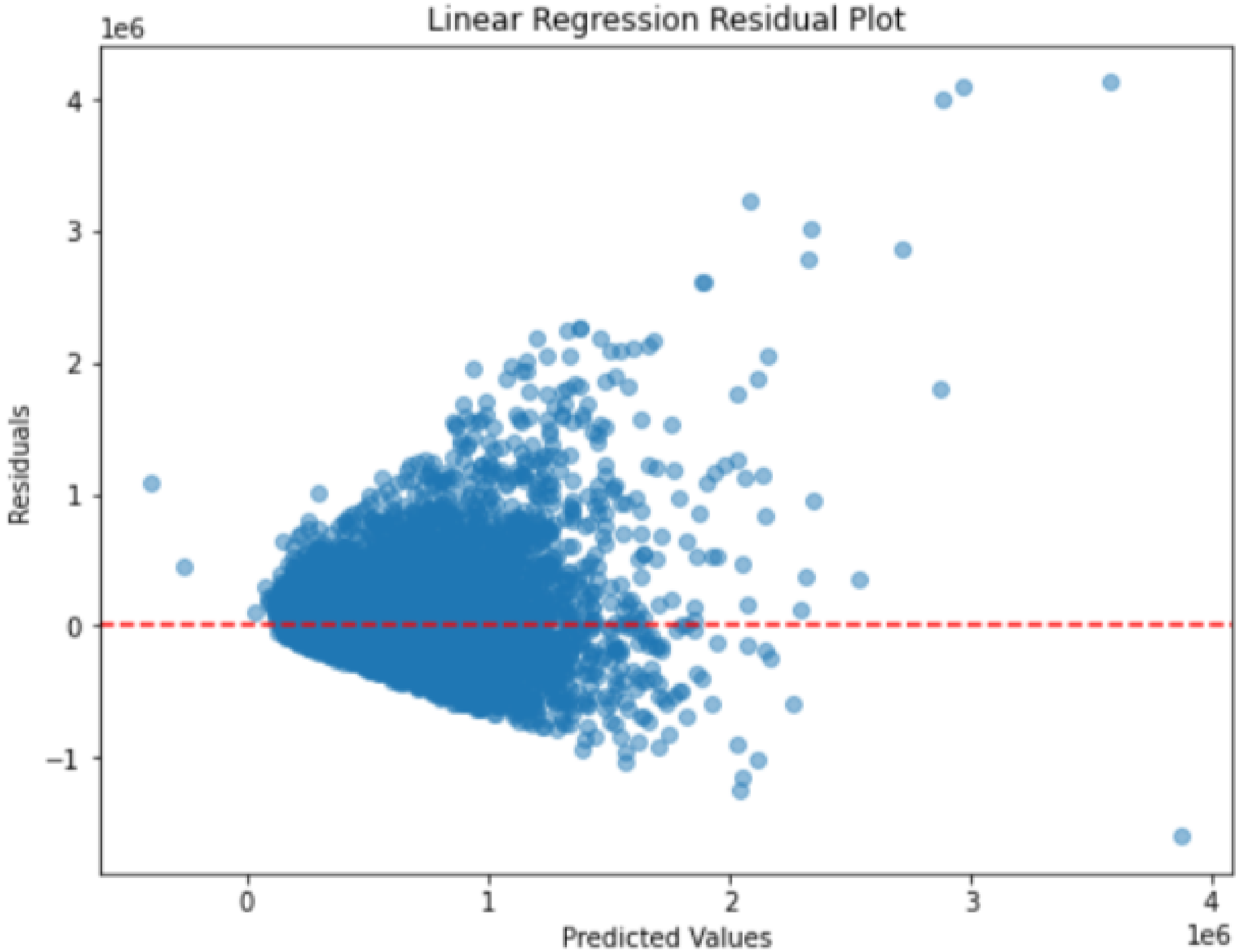
We first import standard packages, load the data into a pandas data frame, check the datatypes, and their descriptive statistics, and check for duplicate and missing values.

yr_renovated has the most missing values, followed by waterfront and view. waterfront and view are categorical, so we will replace those null values with 0s. I will assume null values in yr_renovated mean that the house has not been renovated and will be replaced with 0s.

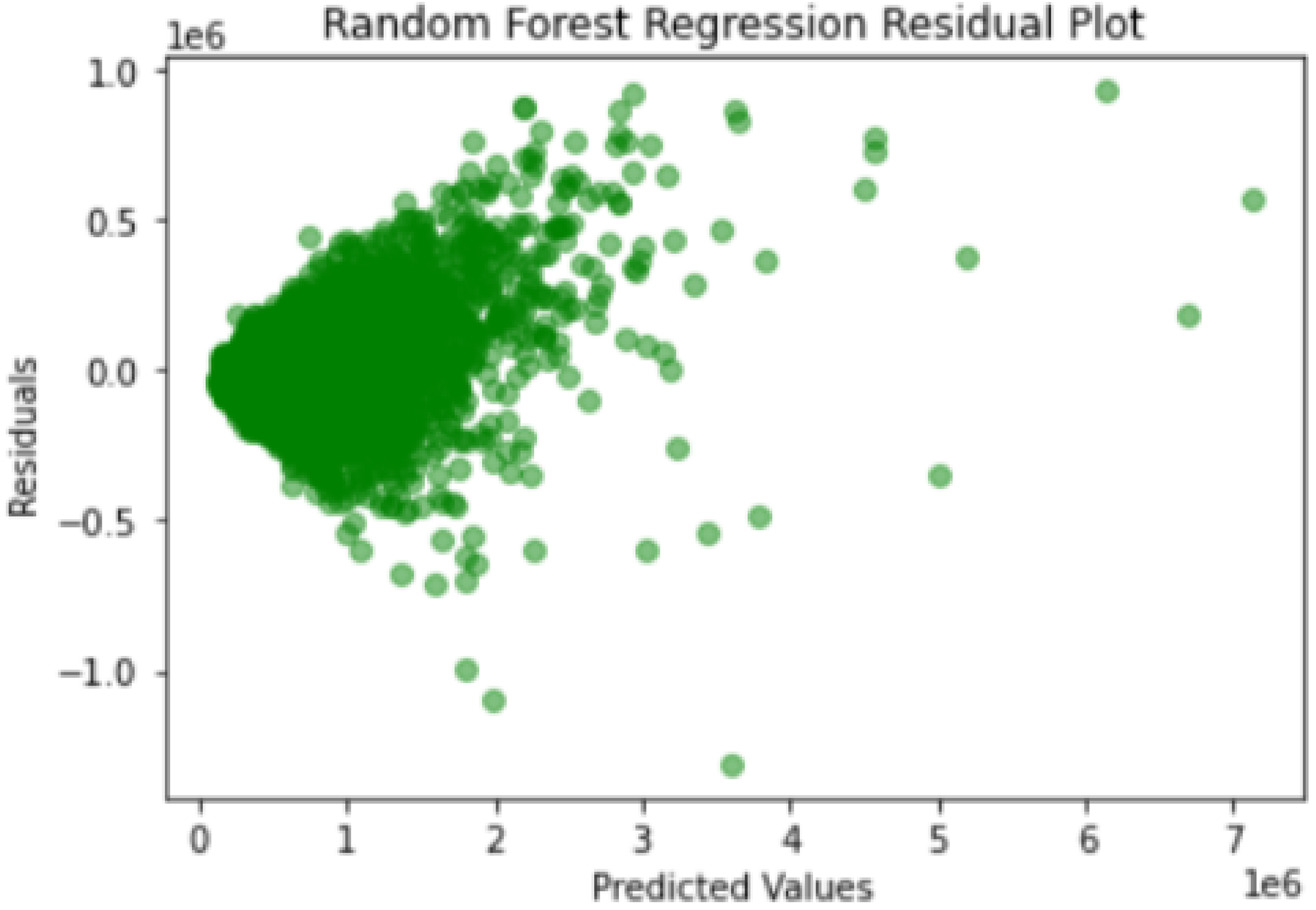
The date, waterfront, view, condition, grade, and sqft_basement columns are stored as type objects, we will need to remove or convert them to numerical data type before modeling.

We then clean the dataset by removing irrelevant columns to my analysis and trimming the dataset of null values. We also remove outliers and convert the remaining categorical columns containing strings into numeric datatypes.

We employ a Linear Regression model, a foundational technique for predicting a continuous target variable based on predictor features. It offers clear interpretability of feature coefficients, aiding in extracting meaningful insights.


Linear Regression Residual Plot
1e6
4
3
2
1
0
−1
Residuals
0
1
2
3
4
Predicted Values
1e6

Introduce a Random Forest Regressor, a sophisticated ensemble technique that leverages multiple decision trees for predictions. Unlike Linear Regression, it excels in capturing complex, non-linear relationships within the data. This can be crucial when the underlying patterns are intricate and challenging to model with simpler techniques. By implementing this, we aim to explore potential improvements in predictive accuracy compared to the initial Linear Regression model.


Random Forest Regression Residual Plot
1e6
1.0
0.5
0.0
−0.5
−1.0
Residuals
0
1
2
3
4
5
6
7
1e6
Predicted Values

Regression Results

The model's performance hinges on two key metrics: Root Mean Square Error (RMSE) and R-squared (R^2). With an RMSE of approximately 257,093, our predictions tend to deviate within this range from actual prices. Simultaneously, the R^2 value of 0.51 signifies that 51% of the variability in house prices is elucidated by our model. Among the features, bedrooms wield the most significant impact, with each additional bedroom decreasing the price by about 65,738.

Conversely, an extra bathroom contributes an increase of roughly 7,967. Living area positively influences prices, with each additional square foot resulting in an increase of approximately 318. However, lot size has a minimal impact and the extra floor reduces the price by around 2,096. The Random Forest model bolsters accuracy, yielding a lower RMSE of about 91,818, indicating enhanced predictive capability.

In choosing statistical analyses over basic data visualization, we opt for a more nuanced understanding of the intricate relationships within our dataset. While graphs offer visual representation, regression coefficients provide precise quantification of each feature's impact on house prices. This level of detail is paramount in the complex realm of real estate, where numerous factors converge to determine property values. Through regression analysis, we can discern subtle effects and interactions, offering a comprehensive assessment for our data science audience. Our modeling process was guided by these statistical insights, ensuring the model's accuracy and effectiveness.

Limitations and Recommendations

For a data science audience, it is crucial to acknowledge potential limitations. Assumptions inherent in linear regression may not perfectly align with the intricate dynamics of the real estate market. Outliers and influential data points could potentially skew results, warranting vigilant consideration. To augment the model's efficacy, stakeholders should supplement it with expert judgment and market awareness. While the model excels at providing price ranges, it does not encompass broader market trends or unforeseen external influences.

Therefore, an iterative approach involving an expanded dataset and exploration of advanced modeling techniques, such as Random Forest Regressors, will refine the model's accuracy and applicability in real-world scenarios.

Next Steps

More research is required to have a more integrated and informative dataset finding more factors that influence the price.

More time would be required to fine-tune our findings and model results.

Using datasets from other countries to be able to better advise our customers by comparing the dataset results.

The agency may have a questionnaire to identify their strengths, weaknesses, opportunities, and threats and use this information to prioritize recommendations that would help address their weaknesses and take advantage of their opportunities and strengths.

It is also important for the agency to continuously evaluate the effectiveness of the strategies they implement and make adjustments as necessary. This could involve tracking metrics like website traffic, this model, social media engagement, and lead generation to assess the impact of their efforts and identify areas for improvement.

THANK YOU