House Price Prediction

BY: MD. NAFIUL ISLAM

Contents

- Introduction
- Objectives
- Workflow
- Exploratory Data Analysis
- Data Cleaning and Preprocessing
- Modeling
- Performance Analysis
- Findings
- Recommendation
- Conclusion

Introduction

- Connected to or constructed on land
- Any improvement in relation to the land that rises or lowers the house price
- Ownership and usage rights
- Residential, Commercial, Industrial, Raw Land, and Special use

Introduction (Cont.)

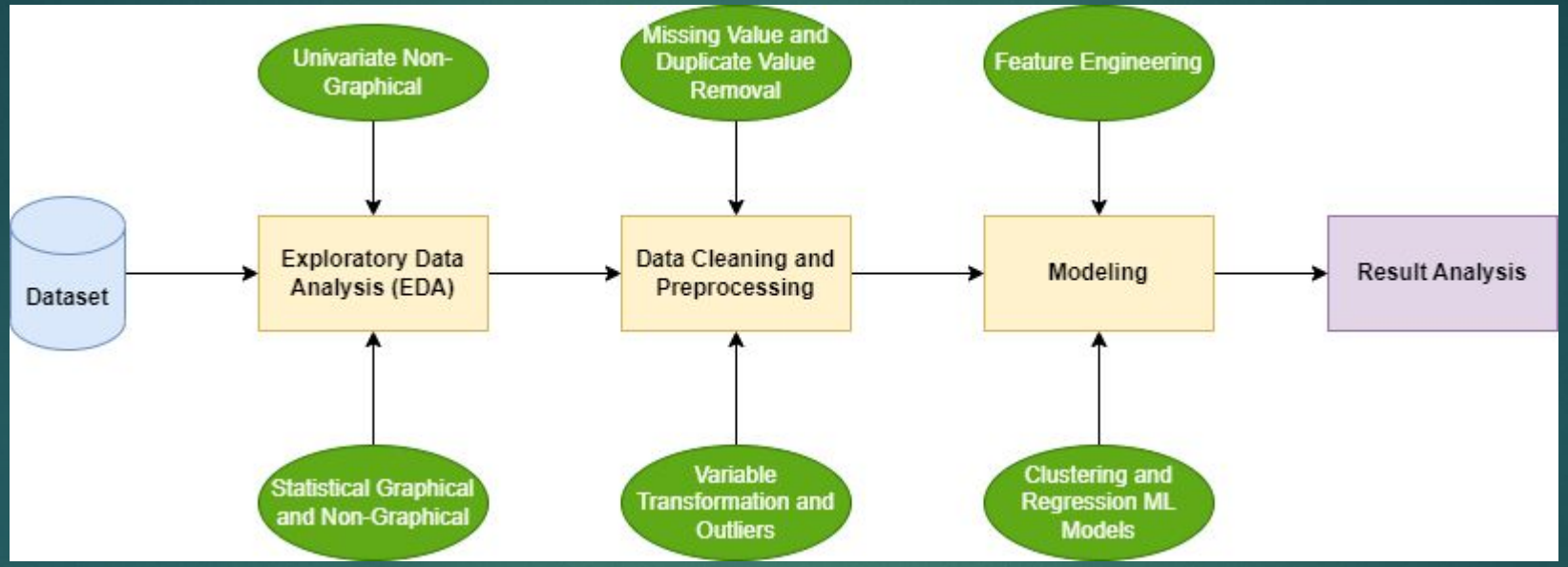
- **Challenges:**
  - No control over the market
  - Stressful to both buyer and seller
  - Setting the price
  - Confliction of unrealistic home buyers

- **Tackles:**
  - Automated system
  - Investigating the most useful features
  - AI and Machine learning

Objectives

- Identifying the essential features influencing the cost of a house using Exploratory Data Analysis
- Understanding the aspects affecting the cluster model for houses and estimate house prices based on the attributes

Workflow



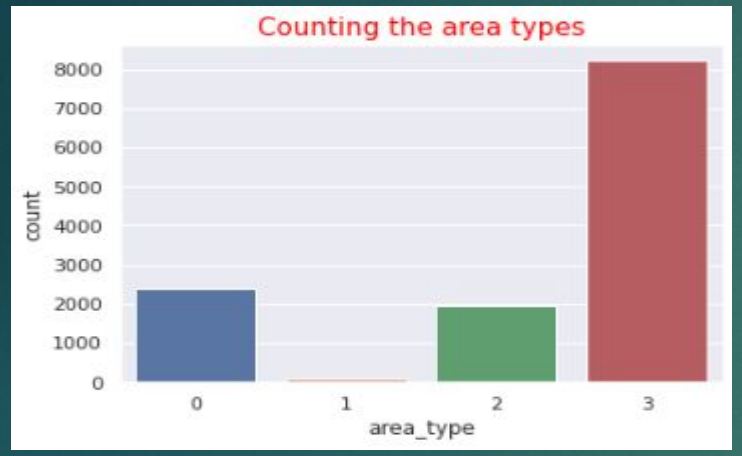
- **Dataset**
  - Collected from the kaggle
  - 3,320 instances with 9 attributes
  - Attributes are: area type, location, society type, availability, room counts, bathrooms, balconies, total square size and price

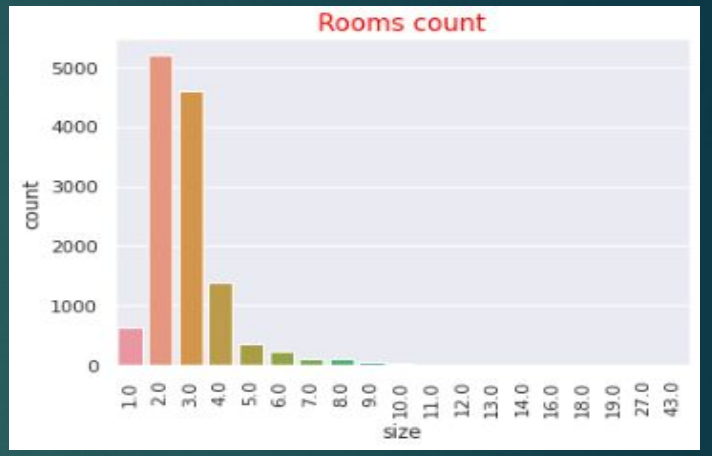
Exploratory Data Analysis

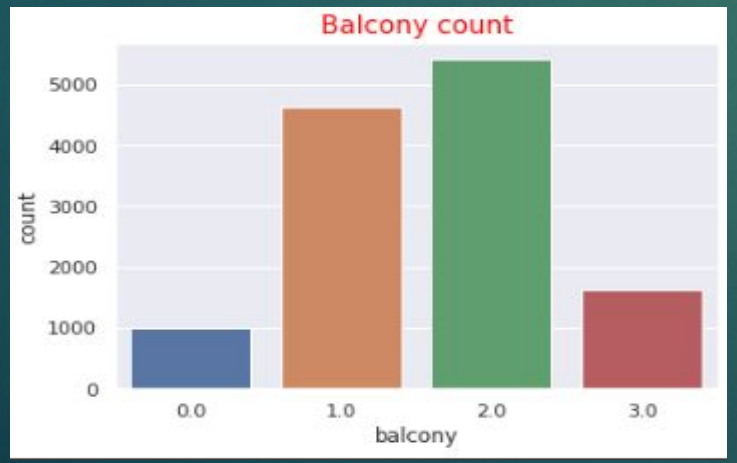
- **Univariate Non-Graphical**
  - Unique value analysis for each columns
  - Skewness of data
- **Statistical Graphical and Non-Graphical**
  - Univariate analysis
  - Bivariate analysis

Exploratory Data Analysis (Cont.)

- **Univariate Analysis**







Exploratory Data Analysis (Cont.)

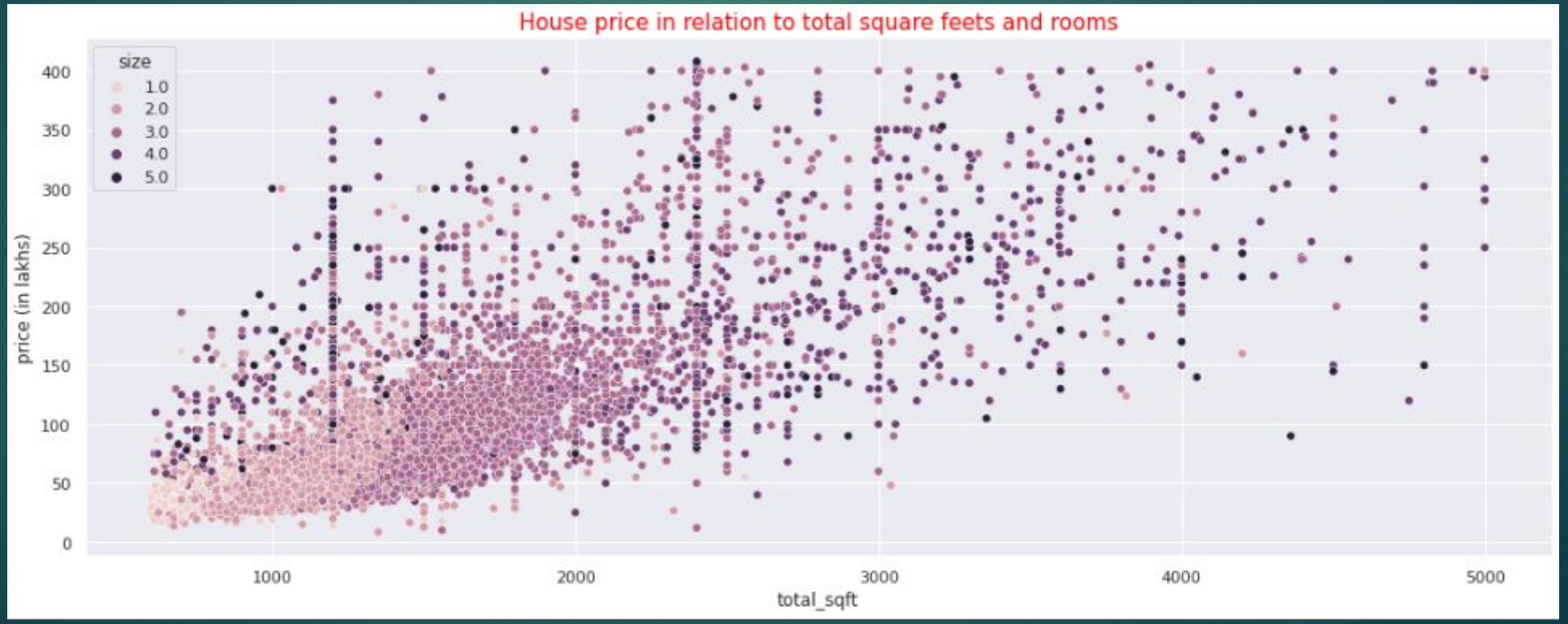
- **Bivariate Analysis**

Data Cleaning and Preprocessing

- Handling the missing data
- Duplicate data removal
- Data transformation
- Outlier detection and handling the outliers

Data Cleaning and Preprocessing (Cont.)

- **After Data Cleaning**



Modeling

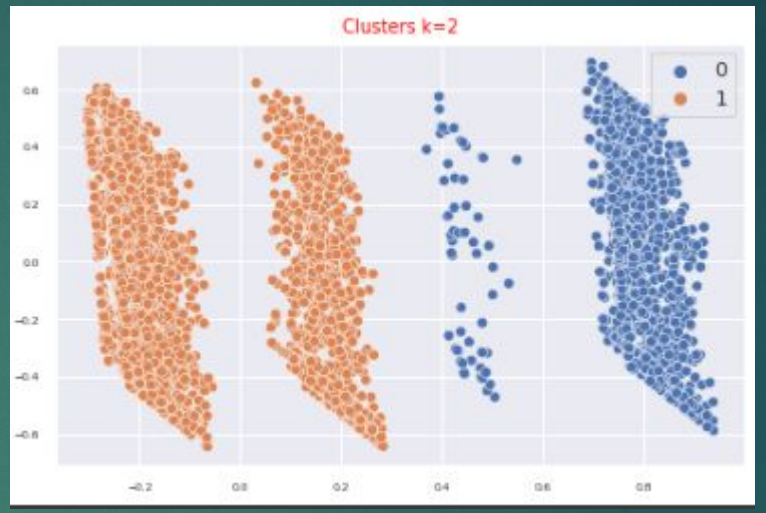
- Clustering
- Regression models of machine learning

Modeling (Cont.)

- **Clustering**
  - Silhouette analysis and sum of squared distance to find the optimal clusters
  - Elbow plot
  - K-means clustering algorithm
    - Sklearn module
    - Own function
  - Hierarchical clustering
    - Dendrogram

Modeling (Cont.)

- With 2 clusters



- With 60 clusters

Modeling (Cont.)

- **Regression Models**
  - Checking multicollinearity
    - Heat map
    - Variation Inflation Factor
    - Creating final features
  - ML regression models
    - 15 regression models
    - Linear models
    - Ensemble models
    - Hyper-parameter tuning

Modeling (Cont.)

- **Heat map**

Modeling (Cont.)

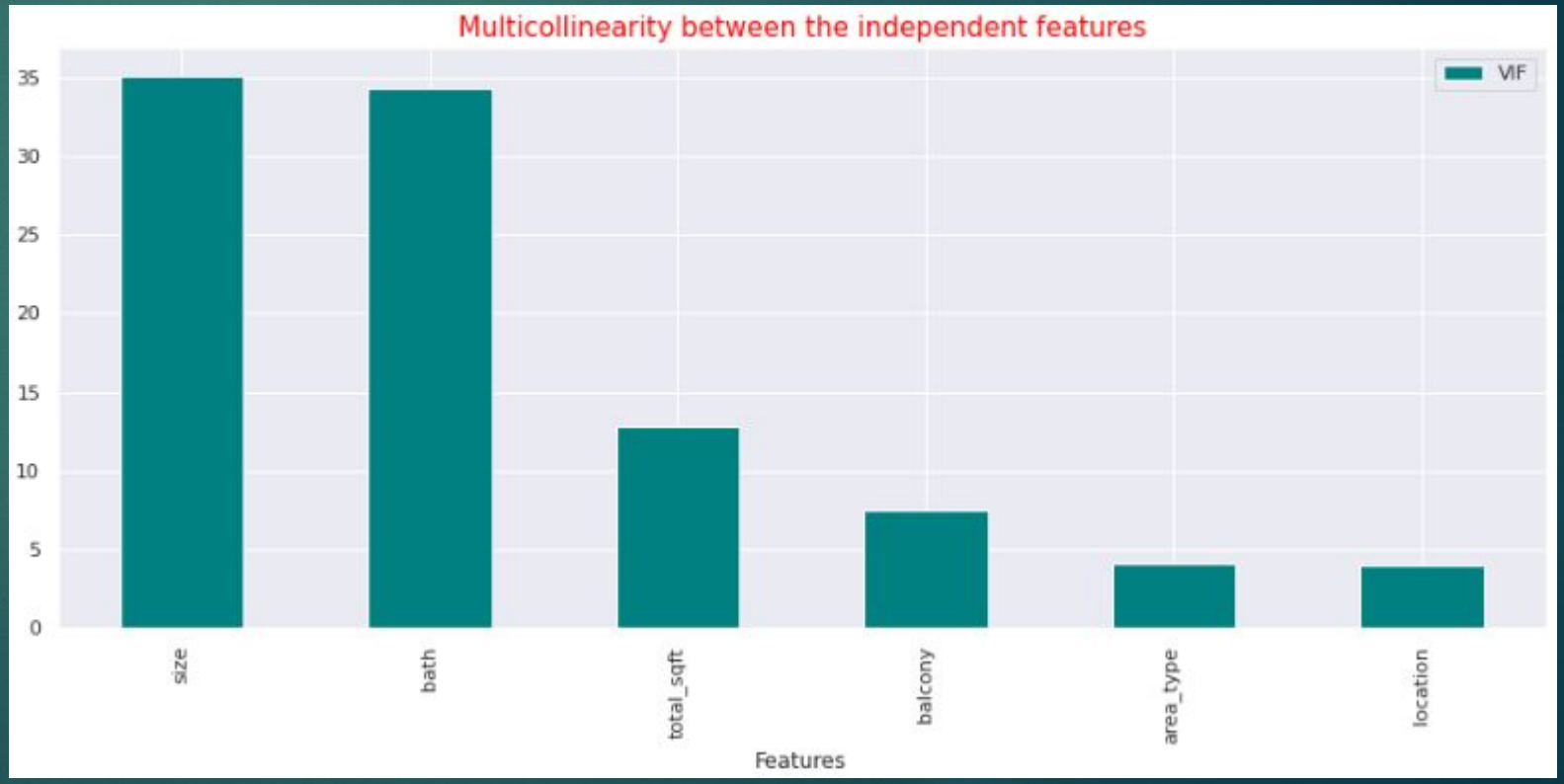
- **Variation Inflation Factor**

Final features are:

Size, Total square size,

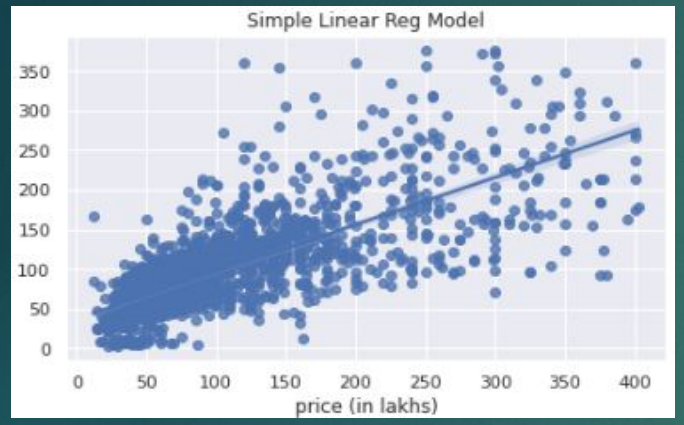
Number of balconies,

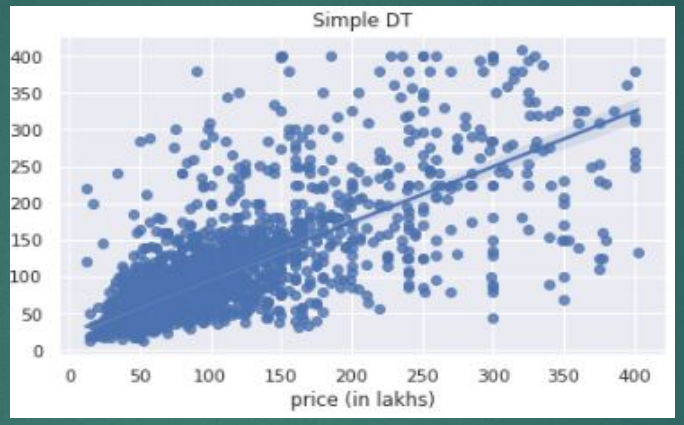
Area type and location

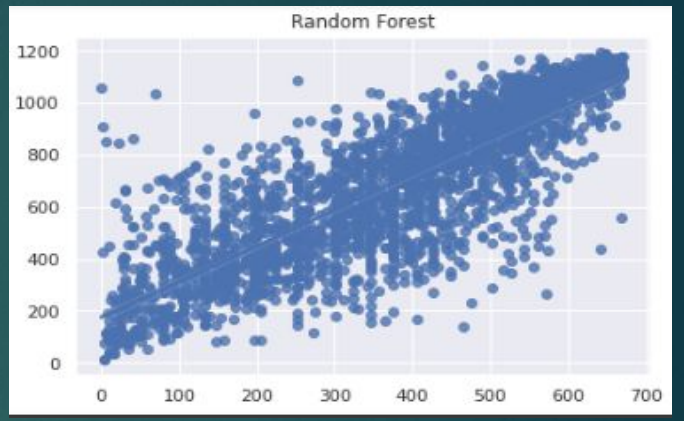


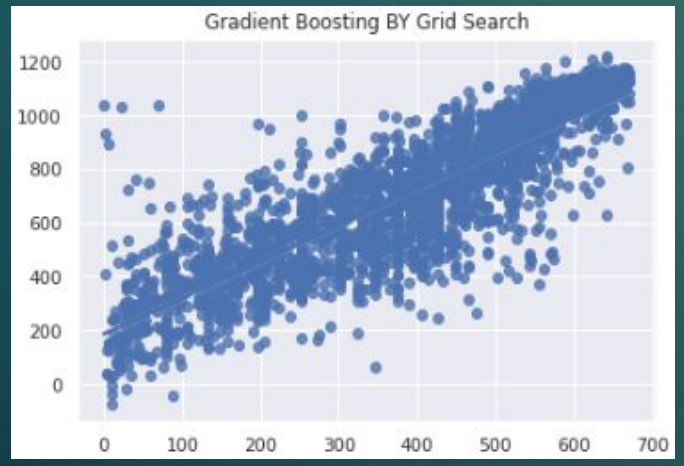
Modeling (Cont.)

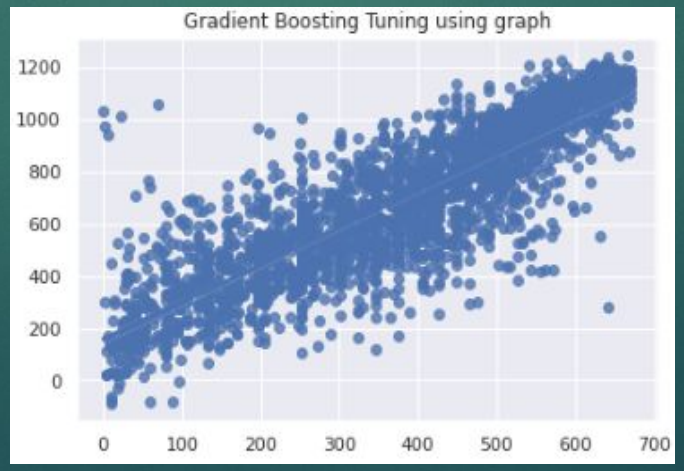
- **Relationship between true and predicted value for unseen data for some model**

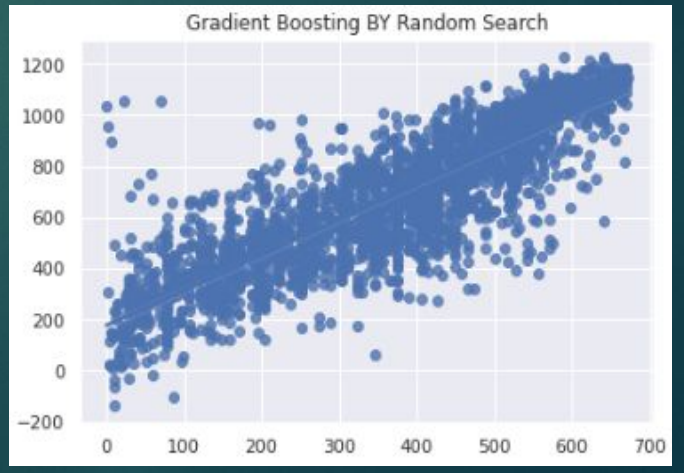












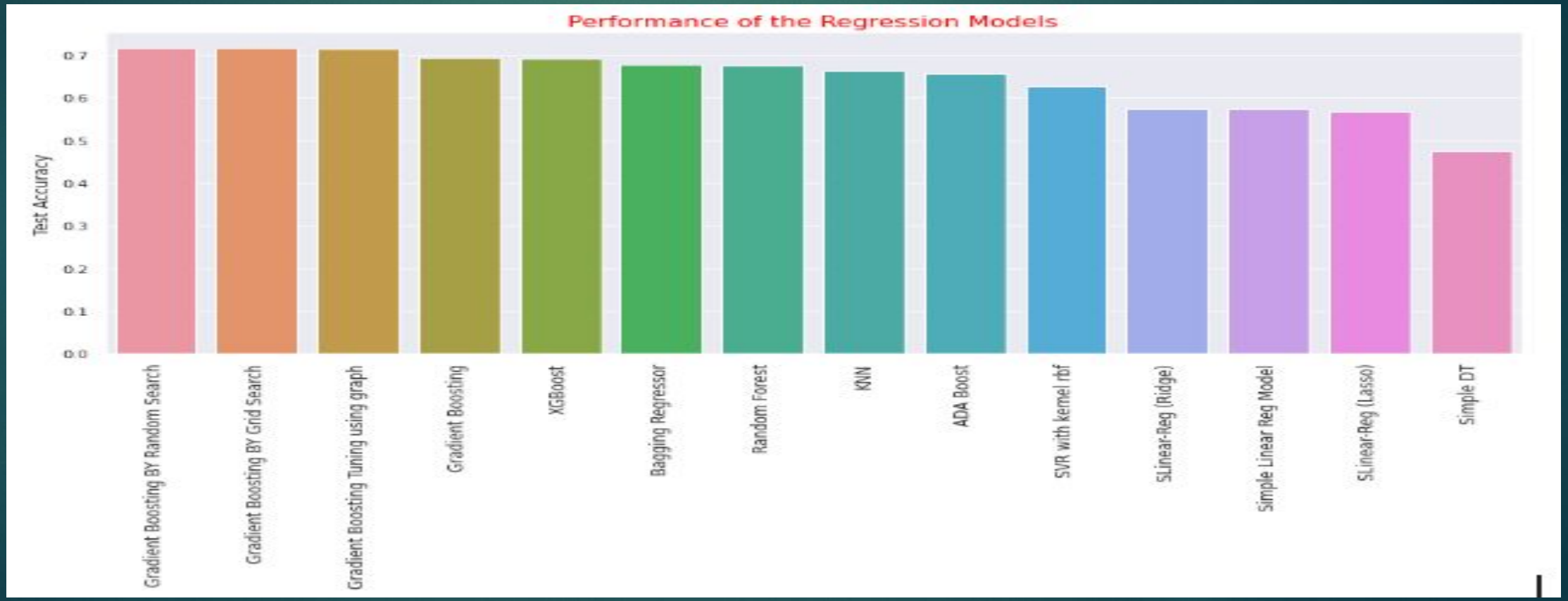
Performance Analysis

- Gradient Boosting with Random Search gives best performance
- Logistic Regression gives the worst performance

	Algorithm	Test Accuracy
0	Gradient Boosting B'	0.7170229359
1	Gradient Boosting B'	0.7163811035
2	Gradient Boosting Tı	0.7133390202
3	Gradient Boosting	0.6945983633
4	XGBoost	0.6917708802
5	Bagging Regressor	0.6783504946
6	Random Forest	0.6758614657
7	KNN	0.6633681174
8	ADA Boost	0.6560014896
9	SVR with kernel rbf	0.6273025868
10	SLinear-Reg (Ridge)	0.5744952807
11	Simple Linear Reg M	0.57338649
12	SLinear-Reg (Lasso)	0.5653820661
13	Simple DT	0.4742778974
14	Logistic Regression	-1.887993018

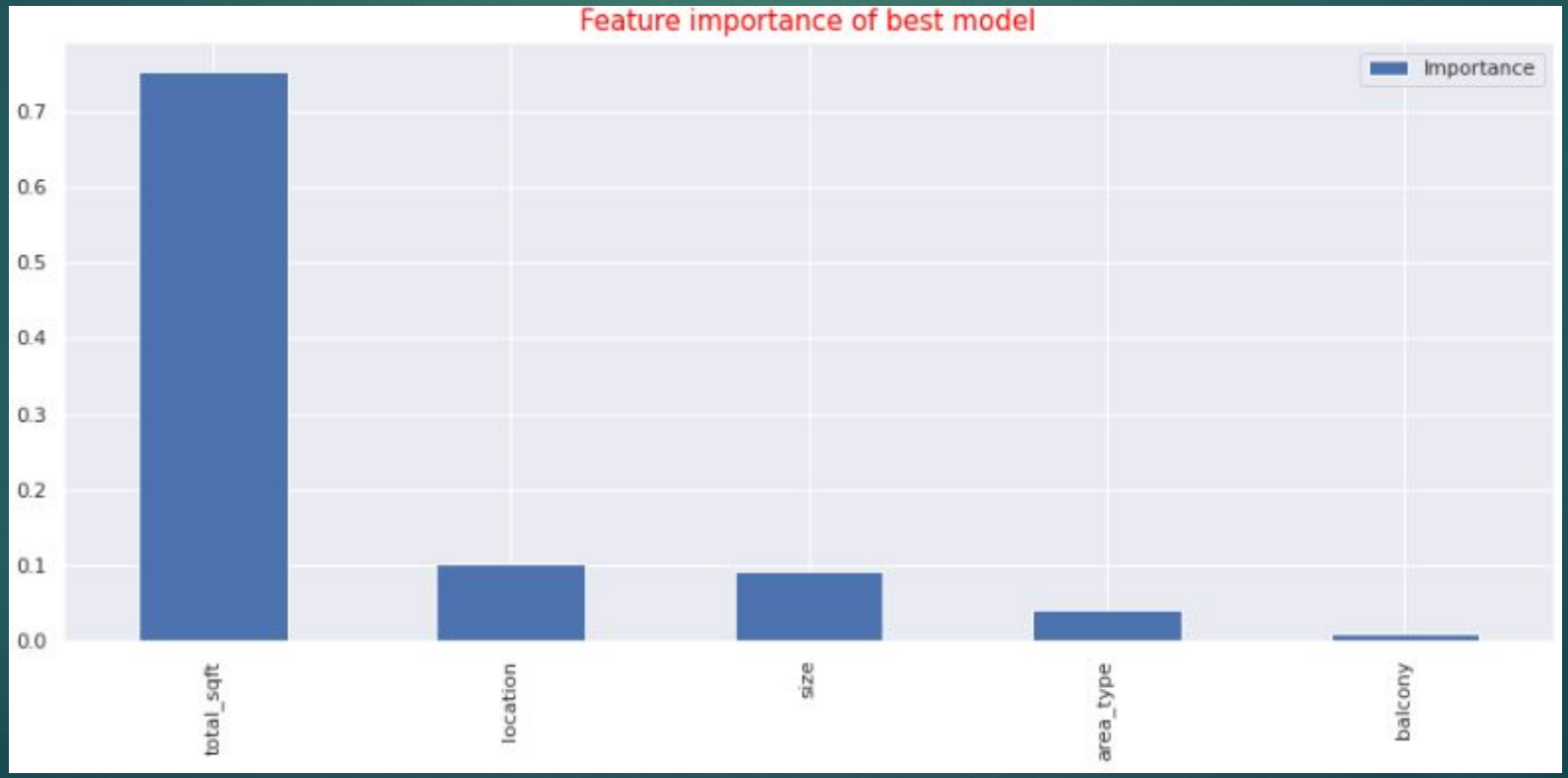
Performance Analysis (Cont.)

- **Performance comparison excluding the Logistic Regression as it gives (-ve) accuracy value**



Performance Analysis (Cont.)

- **Feature importance of the best model**



Findings

- **Clustering**
  - The silhouette approach generates two clusters, but the sum of squared method generates thirty to sixty or more clusters. However, the silhouette method is more efficient
  - Hard to find the optimal clusters. With the 60 clusters all the clusters almost same in kind

Findings (Cont.)

- **Machine Learning Regression Models**
  - Simple ml models like linear regression, KNN, SVR, decision tree performed poor but ensemble models performed better
  - Gradient boosting algorithm is performed well by performing the hyper parameter tuning using randomized search cv

Recommendation

- **Clustering**
  - DBSCAN, TSNE methods could be better
  - RICA or SFT to apply unsupervised feature learning to input data
  - Agglomerative clustering method can improve the performance

Recommendation (Cont.)

- **Machine Learning Regression Models**
  - Hyper parameter tuning of neural networks or simple linear models can improve the performance
  - Can be model how two or model independent factors combined to interact with the house price
  - Polynomial regressions could be used to improve the performance

Conclusion

- A rigorous process of data analysis to clean and ready the data
- Tried to create the clusters carefully
- No multicollinearity
- 10-fold cross validation used for the regression models
- Simple linear to advanced ensemble method are performed to estimate the cost of the house