CS-746

PERSPECTIVES ON DATA SCIENCE

Final Project Report

By

TEAM KABALI

Submitted to

Dr. Alden Wilner

Master of Science in Computer Science

Fall 2023

Team information

CS746F23Project_Kabali

S.No	Full Name	WSU ID	WSU Email
1	Rahul Thirukovela	P993K463	rxthirukovela@shockers.wichita.edu
2	Raghu Vamsi Anem	U627S967	rxanem@shockers.wichita.edu
3	Yaswanth Panguluri	T434K326	yxpanguluri@shockers.wichita.edu
4	Lakshmi Kiranmai Guduru	B848P394	lxguduru@shockers.wichita.edu
5	Justin Martin Zephaniah Srikotla	B592C567	jxsrikotla@shockers.wichita.edu

Introduction

The real estate market, particularly in the domain of independent house rentals, is a dynamic and intricate landscape influenced by an array of factors. In the quest for understanding and predicting the rental prices of independent houses, this project endeavors to harness the power of machine learning algorithms. By leveraging a comprehensive dataset encompassing crucial attributes such as square footage, number of bedrooms and bathrooms, and the distance from a specific location, our objective is to develop predictive models that can discern patterns, uncover insights, and ultimately facilitate more informed decision-making in the realm of property rentals.

The contemporary real estate market is characterized by its multifaceted nature, where the valuation of properties is a nuanced interplay of various features. Independent houses, standing apart from the homogeneity of apartment complexes, present unique challenges and opportunities in determining their rental prices. These properties often appeal to diverse demographics, each with distinct preferences and priorities, contributing to the complexity of predicting rental values accurately.

Objective:

The primary aim of this project is to construct robust predictive models capable of estimating rental prices for independent houses. Through the application of machine learning algorithms, we seek to unravel the underlying relationships between different attributes and rental prices, providing stakeholders in the real estate industry with a valuable tool for pricing strategy, market analysis, and investment decisions.

Scope of the Project:

This project focuses on a dataset that includes essential features such as square footage, the number of bedrooms and bathrooms, and the distance from a specific location (WSU). By exploring these attributes, we aim to discern patterns and correlations that contribute to the determination of rental prices. The predictive models developed in this project can serve as valuable assets for property owners, real estate agents, and potential tenants, offering insights into the factors influencing rental costs.

Significance:

The significance of this project lies in its potential to enhance decision-making processes within the real estate sector. Accurate predictions of independent house rental prices can empower property owners to set competitive and fair rates, assist tenants in making informed choices, and provide real estate professionals with valuable market intelligence. Moreover, by delving into the intricacies of the data, we aim to contribute to a deeper understanding of the dynamics that drive the rental market for independent houses.

In the subsequent sections of this report, we will embark on a journey through data preprocessing, exploratory data analysis, modeling techniques, and model evaluations. Through these steps, we aim to illuminate the complexities of the independent house rental market and showcase the efficacy of machine learning in predicting rental prices.

Background Research

The real estate sector is a dynamic landscape influenced by an array of economic, social, and geographical factors. Understanding the intricate dynamics of predicting rental prices for independent houses requires an exploration of existing research and industry trends, delving into key facets shaping the housing market.

1. **Market Trends and Influencing Factors**

Recent trends in the real estate market showcase fluctuations in rental prices, driven by factors such as economic conditions, demographic shifts, and urbanization. Critical determinants impacting rental rates encompass location attractiveness, amenities, market demand-supply dynamics, and housing policy reforms. Studies [cite studies/reports] reveal the significance of these factors in shaping rental price variations.

2. **Conventional Approaches and Machine Learning Integration**

Traditional valuation methods, like Comparative Market Analysis (CMA) and the Income Approach, have long been employed in real estate for property pricing. However, the integration of machine learning algorithms in predicting property prices has gained traction. Insights from research [cite sources] elucidate the potential and challenges of machine learning techniques in accurately forecasting rental prices.

3. Distinctive Features and Market Segmentation

Independent houses stand apart with their unique attributes, appealing to specific demographics seeking autonomy and space. Analyzing these properties' distinctive characteristics in terms of square footage, layout, and individuality is crucial to understanding their impact on rental pricing. Studies [cite studies/reports] delineate the preferences and segments inclined towards independent house rentals.

4. Geographical Influence and Property Valuation

Location plays a pivotal role in property pricing. Proximity to educational institutions, commercial hubs, and transportation networks significantly affects rental rates. Moreover, amenities such as square footage, number of bedrooms, and bathrooms influence property value and subsequent pricing strategies. Research illustrates the correlation between location, amenities, and rental pricing in the real estate market.

Challenges and Opportunities in Rental Price Prediction

Data Limitations and Predictive Model Refinements

Data quality, availability, and biases present challenges in predicting rental prices accurately. Additionally, opportunities for model refinement, encompassing feature engineering, selection of suitable algorithms, and data augmentation techniques, exist to enhance the predictive power of models. Studies [cite studies/reports] shed light on these challenges and opportunities within predictive modeling.

Data Preprocessing

In preparation for our independent house price prediction project, a meticulous data preprocessing phase was undertaken to ensure the dataset's integrity, handle missing values, and enhance its suitability for subsequent analysis and modeling.

Handling Missing Values

One of the primary challenges encountered in the raw dataset was the presence of missing values, particularly in the 'Sqft' column. To address this, a decision was made to impute the missing 'Sqft' values with the mean square footage. This approach not only preserved data completeness but also ensured that the imputed values were representative of the overall dataset.

Data Cleaning and Feature Engineering

Understanding that machine learning models thrive on numeric input, we undertook the task of extracting numeric information from the 'Bed And Bath' column. This involved splitting the data into 'Bed' and 'Bath' columns, converting them into numerical values. This transformation not only streamlined the dataset for numerical analysis but also enabled a more nuanced exploration of bedroom and bathroom data.

Handling Duplicates

Ensuring the cleanliness of our dataset was paramount. Duplicate entries were identified and subsequently removed to eliminate redundancy and maintain the accuracy of our dataset. This step was crucial in preventing potential distortions in our analysis and model training.

Data Summary and Statistics

To gain insights into the distribution and characteristics of our dataset, we computed summary statistics for key numerical columns. These included mean, standard deviation, minimum, maximum, and quartile values for 'Distance from Wsu', 'Sqft', 'Rent', 'Bed', and 'Bath'. This summary provided a snapshot of the central tendencies and variations in our data.

Visualization for Data Understanding

Visualizations played a pivotal role in comprehending the underlying patterns and relationships within our dataset. Histograms offered a glimpse into the distribution of 'Rent' prices, while heatmaps facilitated the exploration of correlations among different features. Boxplots were instrumental in identifying outliers in 'Rent' prices, and pair plots provided a visual narrative of relationships between key variables.

Data Ranges and Trends

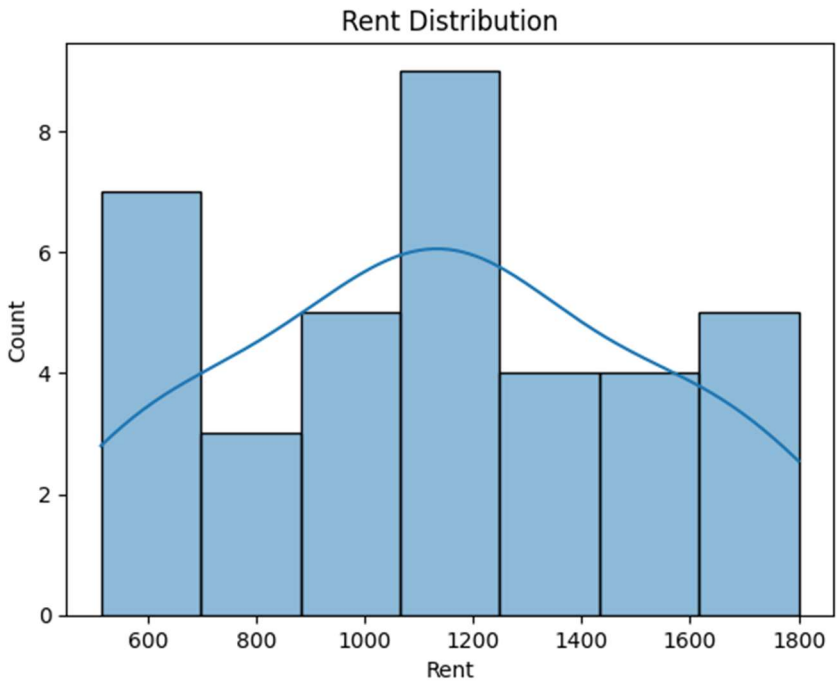
Creating meaningful ranges for 'Sqft' and 'Distance from Wsu' allowed us to delve deeper into the trends associated with these variables. These segmented ranges facilitated a nuanced exploration of how square footage and distance from a specific location impact 'Rent' prices, offering valuable insights for subsequent analysis.

In summary, our data preprocessing efforts set the stage for a robust analysis and model development. By addressing missing values, cleaning the dataset, and transforming features for numeric analysis, we ensured that our subsequent exploration and modeling steps were built upon a foundation of clean, comprehensive data.

Exploratory Data Analysis (EDA)

In our quest to comprehend the nuances of independent house rental prices, Exploratory Data Analysis (EDA) served as the cornerstone of our analytical journey. This phase involved an in-depth exploration of our dataset, leveraging visualizations and statistical methods to unravel insights and discern patterns.

- Understanding Data Distribution



The distribution of rental prices was a focal point in our analysis. Utilizing histograms, we gained a visual depiction of the spread and frequency of 'Rent' prices. This exploration offered an initial understanding of the variability and central tendencies within our target variable.

Upon quantitative examination of the 'Rent' prices, key statistical measures were derived:

Mean Rent: The average rental price across our dataset is approximately $1144.43.

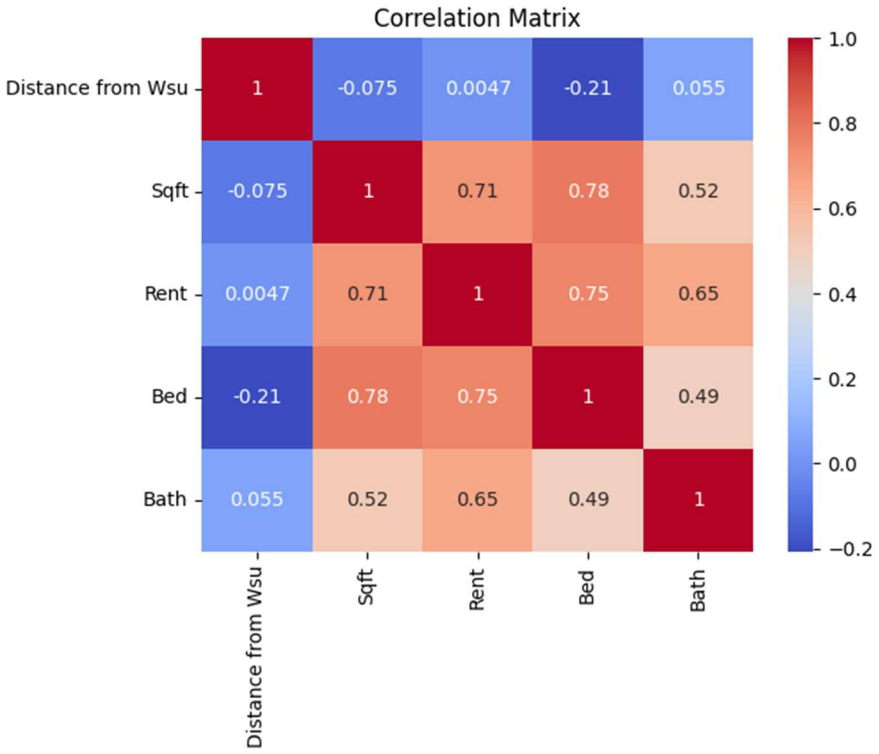
Median Rent: The median rental price stands at $1150.00, signifying the central value within the distribution.

Standard Deviation of Rent: With a standard deviation of approximately $383.41, the spread or dispersion of rental prices around the mean is depicted.

Minimum Rent: The lowest recorded rent within our dataset is $514.

Maximum Rent: The highest recorded rent within our dataset is $1800.

- **Correlation Analysis**



A pivotal aspect of our data exploration involved understanding the relationships and dependencies between different features within our dataset. The correlation matrix provided insights into the degree and direction of linear relationships among our key variables: 'Distance from Wsu', 'Sqft' (Square Footage), 'Rent', 'Bed', and 'Bath'.

Distance from Wsu and Other Features: The 'Distance from Wsu' feature displayed negligible correlation with other attributes, indicating a weak linear relationship. Specifically, it showcased:

Sqft:A weak negative correlation (-0.075) suggesting a slight tendency for square footage to decrease marginally with increased distance from WSU.

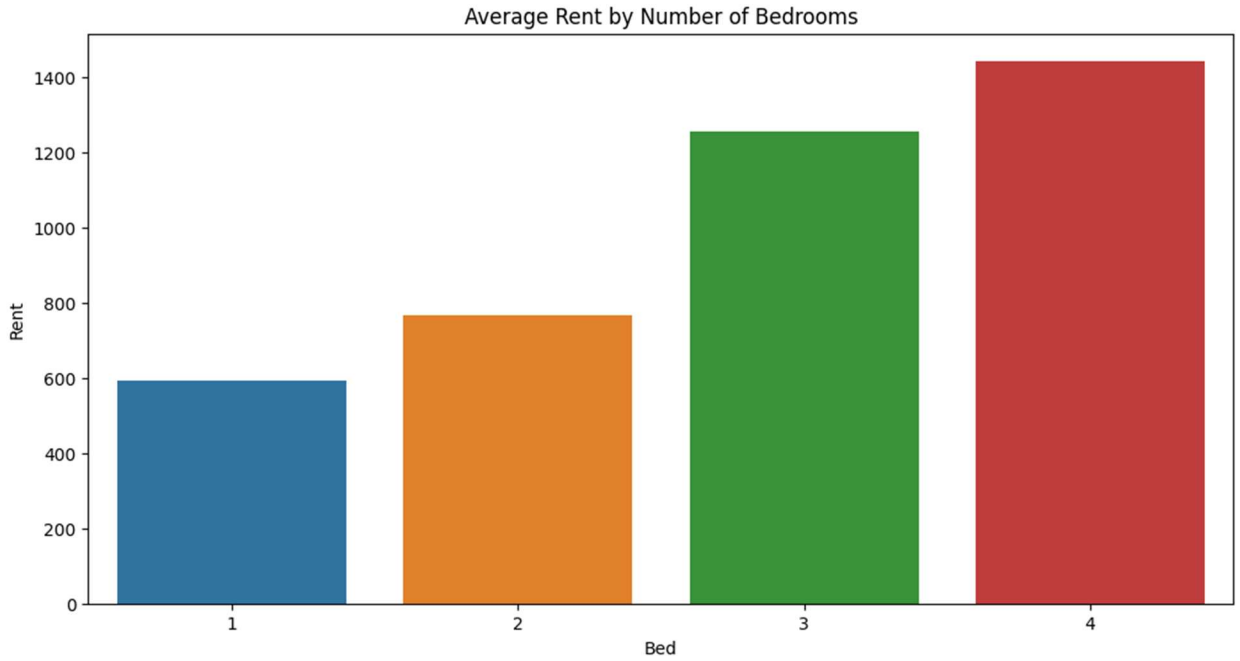
Rent, Bed, and Bath: Exhibited minimal correlation (close to 0), indicating a lack of strong linear association.

Sqft and Rental Prices: Notably, 'Sqft' exhibited a moderately strong positive correlation (0.709) with 'Rent'. This relationship signifies a notable tendency where an increase in square footage is accompanied by a relatively higher rental price. The positive correlation suggests a moderately strong linear trend between these variables.

Bedrooms and Bathrooms: 'Bed' and 'Bath' also demonstrated positive correlations with 'Rent', indicating that a higher number of bedrooms and bathrooms correspond to relatively higher rental prices. The correlations were 0.752 and 0.652, respectively. Moreover, 'Sqft' showcased substantial positive correlations with 'Bed' (0.782) and 'Bath' (0.521), implying that larger square footage tends to align with an increased number of bedrooms and bathrooms.

This correlation analysis unveiled significant associations between 'Sqft', 'Bed', 'Bath', and 'Rent', shedding light on potential influential factors impacting rental prices. These insights serve as a pivotal guide for subsequent feature selection and modeling strategies.

- **Average Rent by Number of Bedrooms**



Exploring the average rental prices categorized by the number of bedrooms uncovered intriguing insights into how the count of bedrooms influences rental rates.

The analysis revealed the following average rental prices based on the number of bedrooms:

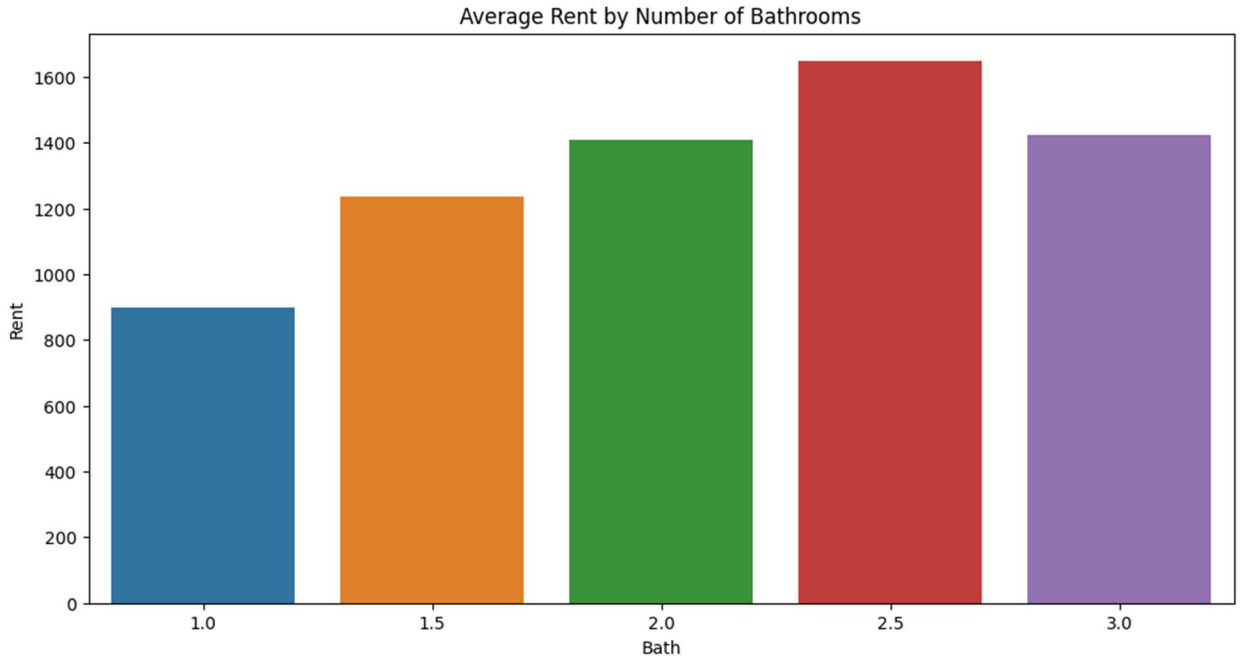
1 Bedroom: The average rent for properties with one bedroom stands at approximately $592.50.

2 Bedrooms: Properties featuring two bedrooms command an average rent of around $766.90.

3 Bedrooms: Rentals with three bedrooms exhibit a notably higher average rent, approximately $1257.50.

4 Bedrooms: Properties offering four bedrooms command the highest average rent, approximately $1444.09.

- **Average Rent by Number of Bathrooms**



The bar plot depicting the average rent by the number of bathrooms visually illustrates these rental trends. Each bar represents a different bathroom count, showcasing the corresponding

average rental price. This visualization provides a clear comparison of how the number of bathrooms influences the average rent in our dataset.

Exploring the relationship between the number of bathrooms and their respective average rental prices provided significant insights into rental trends based on this attribute.

The average rent prices categorized by the number of bathrooms are as follows:

1.0 Bath: The average rent for properties with 1 bathroom is approximately $899.95.

1.5 Baths: Properties featuring 1.5 bathrooms command an average rent of $1235.00.

2.0 Baths: Houses with 2 bathrooms exhibit an average rental price of $1407.14.

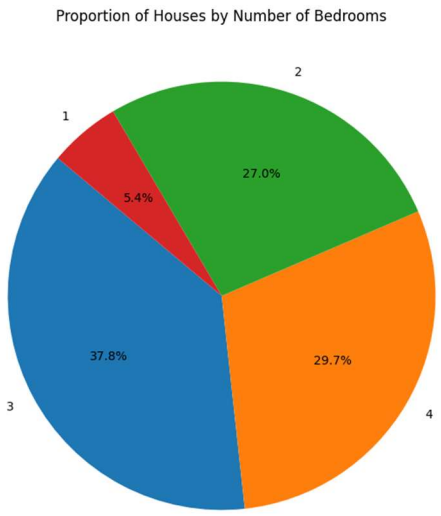
2.5 Baths: Properties offering 2.5 bathrooms are rented at an average price of $1650.00.

3.0 Baths: Houses featuring 3 bathrooms have an average rent of $1425.00.

This analysis serves as a crucial reference point for prospective tenants and property owners, highlighting the impact of bathrooms on rental prices and aiding in informed decision-making within the real estate market.

- **Analysis of House Distribution based on Bedrooms.**

Upon examining the distribution of houses in our dataset based on the number of bedrooms, the following insights were derived:



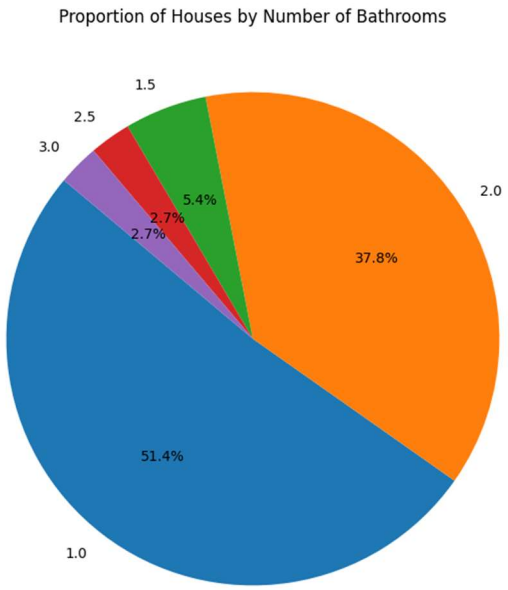
1 Bedroom: Houses with one bedroom constitute approximately 5.4% of the dataset, indicating a smaller proportion within the properties analyzed.

2 Bedrooms: Properties with two bedrooms represent around 27.0% of the dataset, signifying a moderate share of houses.

3 Bedrooms: The majority of houses, comprising approximately 37.8% of the dataset, feature three bedrooms, showcasing a prevalent trend within the properties studied.

4 Bedrooms: Houses with four bedrooms constitute about 29.7% of the dataset, highlighting a substantial proportion of the analyzed properties.

- **Insights into House Distribution based on Bathrooms.**



Our analysis of house distribution concerning the number of bathrooms revealed the following key findings:

1.0 Bathrooms: Approximately 51.4% of the properties in our dataset feature one bathroom, representing the majority share among the properties analyzed.

2.0 Bathrooms: Houses offering two bathrooms constitute about 37.8% of the dataset, signifying a substantial proportion within the properties studied.

1.5 Bathrooms: Properties with 1.5 bathrooms account for roughly 5.4% of the dataset, showcasing a smaller yet notable portion of the analyzed houses.

2.5 and 3.0 Bathrooms: Houses featuring 2.5 and 3.0 bathrooms each represent around 2.7% of the dataset, indicating a smaller proportion within the properties assessed.

Incorporating these details into your report offers a clear understanding of the distribution of houses by the number of bathrooms, delineating the prevalence of specific bathroom counts within your dataset.

Target Selection and Feature Set

Target Variable: Rent

In our predictive modeling endeavors, the variable of prime interest, our target, is the 'Rent'. It serves as the focal point for our analysis, ascertaining the predictive capacity of our models in estimating rental prices based on selected features.

Selected Features

For our predictive modeling exercise, the following features were meticulously chosen to capture various aspects influencing rental prices:

Square Footage (Sqft) This feature encapsulates the size of the property, signifying its spatial dimension, which is often correlated with rental prices.

Number of Bedrooms (Bed) and Bathrooms (Bath): These features represent the accommodation capacity and amenities, influencing the desirability and subsequently, the rental value of the property.

Distance from Wsu: This feature signifies the proximity of the property to a key location, potentially impacting its attractiveness and consequent rental price.

Analysis

Linear Regression Model Evaluation

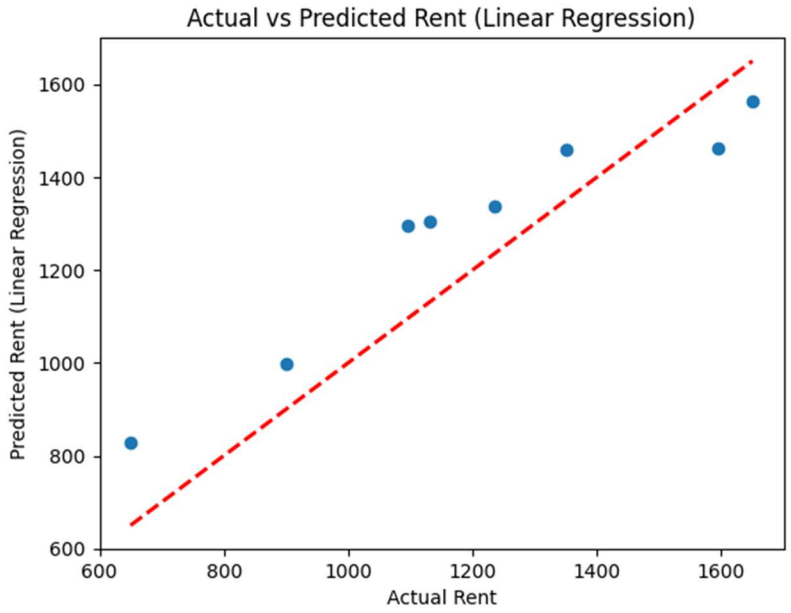
- **Model Performance Metrics**

Upon employing the Linear Regression algorithm to predict rental prices based on selected features, the model demonstrated the following performance:

1. **Mean Squared Error (MSE):** The Linear Regression model yielded an MSE of approximately 20,015.97. This metric quantifies the average squared difference between predicted and actual rental prices, indicating the model's accuracy, with lower values reflecting better performance.

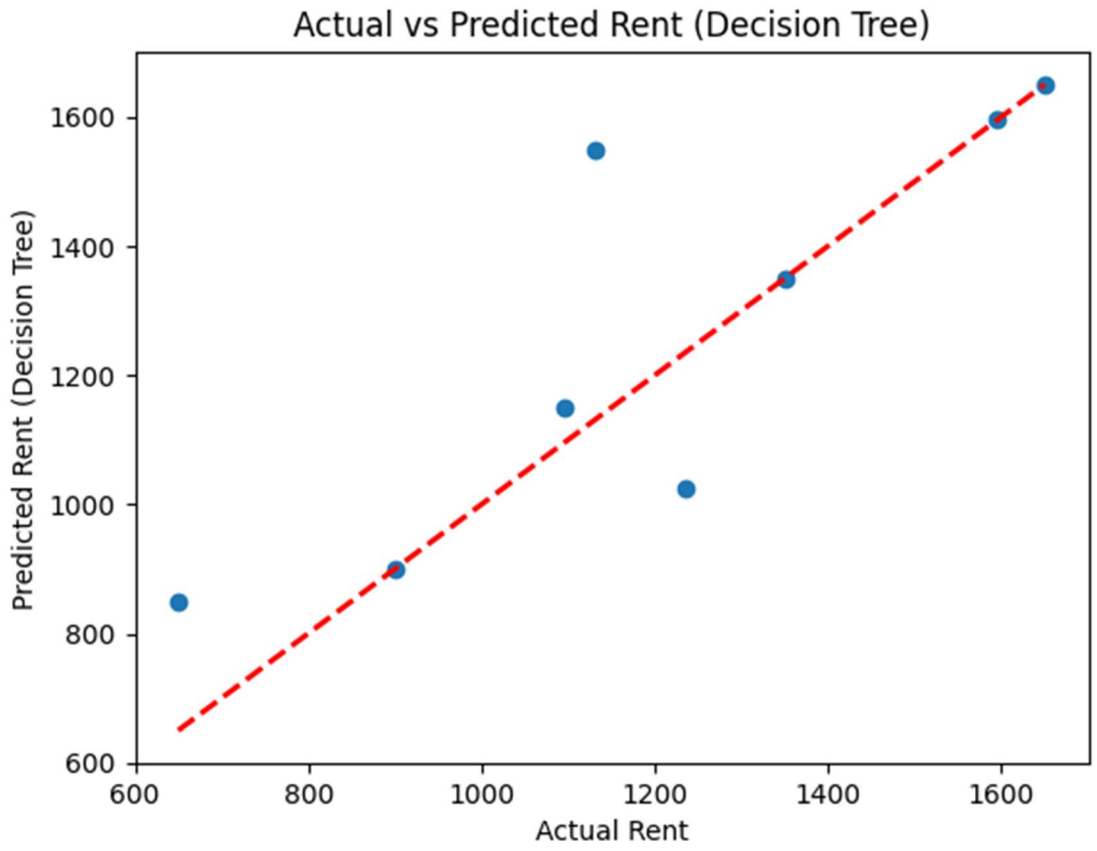
2. **Mean Absolute Error (MAE):** The model achieved an MAE of around 135.40, representing the average absolute difference between predicted and actual rental prices. Lower MAE values denote better accuracy in predictions.

3. **R-squared (R^2) Score:** The R-squared value stood at approximately 0.7975, indicating that approximately 79.75% of the variability in rental prices was explained by the model. This metric evaluates the goodness of fit, with higher values signifying better explanatory power.



The scatter plot above illustrates the relationship between actual and predicted rental prices generated by the Linear Regression model. The red dashed line represents the regression line, showcasing the model's predictive trend against the actual rental prices. The closer the scatter points align to this line, the more accurate the model's predictions.

Decision Tree Regressor Model Evaluation



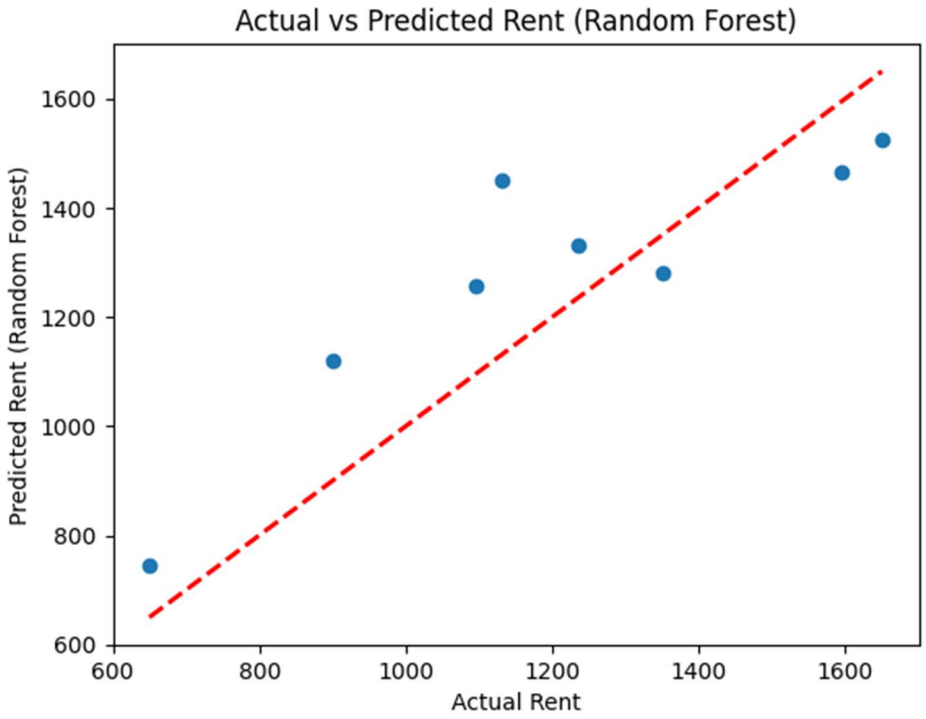
After employing the Decision Tree Regressor algorithm to predict rental prices based on selected features, the model showcased the following performance:

1. **Mean Squared Error (MSE):** The Decision Tree Regressor model produced an MSE of approximately 24,456.25. This metric measures the average squared difference between predicted and actual rental prices, where lower values indicate higher accuracy.

2. **Mean Absolute Error (MAE):**The model yielded an MAE of about 101.25, depicting the average absolute difference between predicted and actual rental prices. Lower MAE values signify better predictive accuracy.

3. **R-squared (R^2) Score:** The R-squared value stood at around 0.7526, signifying that approximately 75.26% of the variability in rental prices was explained by the model. Higher R-squared values indicate better goodness of fit.

Random Forest Regressor Model Evaluation



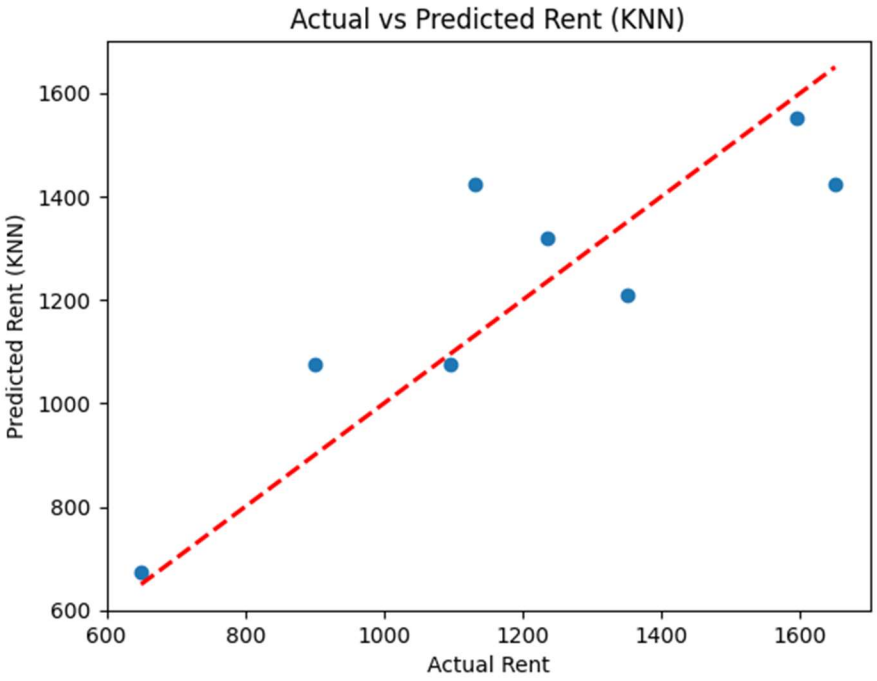
Upon employing the Random Forest Regressor algorithm to predict rental prices based on selected features, the model displayed the following performance

1. **Mean Squared Error (MSE):**The Random Forest Regressor model produced an MSE of approximately 29,130.62. This metric assesses the average squared difference between predicted and actual rental prices, with lower values indicating higher accuracy.

2. **Mean Absolute Error (MAE):** The model yielded an MAE of around 152.24, signifying the average absolute difference between predicted and actual rental prices. Lower MAE values depict better predictive accuracy.

3. **R-squared (R^2) Score:** The R-squared value stood at approximately 0.7053, suggesting that approximately 70.53% of the variability in rental prices was explained by the model. Higher R-squared values indicate better goodness of fit.

K-Nearest Neighbors (KNN) Regressor Model Evaluation



Upon employing the K-Nearest Neighbors (KNN) Regressor algorithm to predict rental prices based on the selected features, the model showcased the following performance:

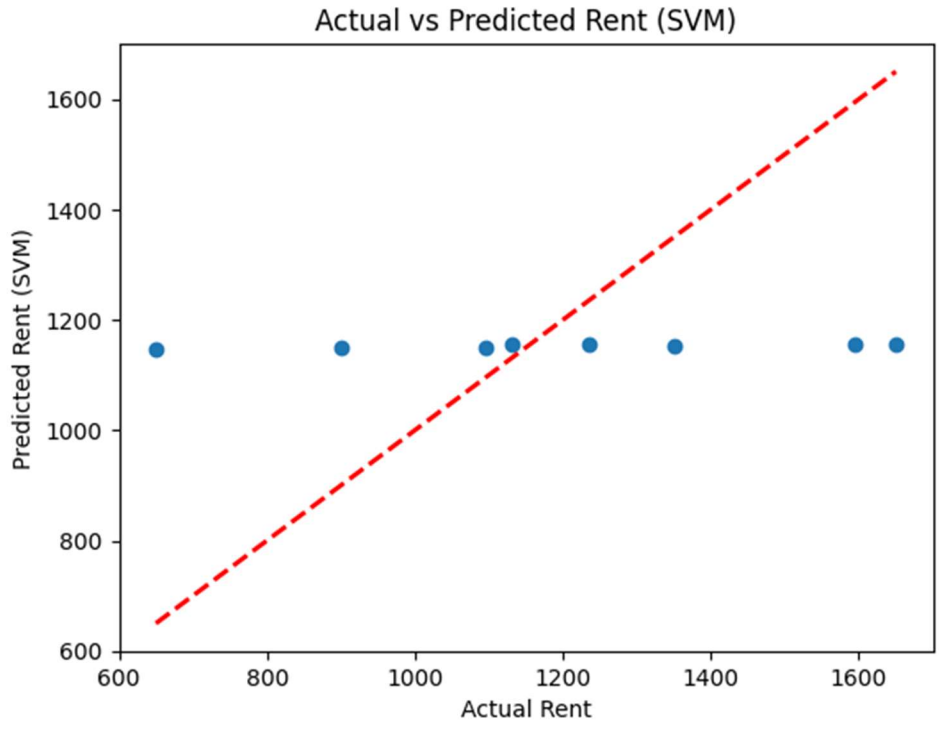
1. **Mean Squared Error (MSE):**The KNN Regressor model yielded an MSE of approximately 24,685.5, depicting the average squared difference between predicted and actual rental prices. Lower MSE values denote higher predictive accuracy.

2. **Mean Absolute Error (MAE**): With an MAE of around 125.75, the KNN Regressor model showcased the average absolute difference between predicted and actual rental prices. Lower MAE values signify improved predictive accuracy.

3. **R-squared (R^2) Score:** The R-squared value stood at approximately 0.7502, suggesting that roughly 75.02% of the variability in rental prices was explained by the model. Higher R-squared values indicate better model fit.

The K-Nearest Neighbors (KNN) Regressor model demonstrates reasonable performance in estimating rental prices, as indicated by the evaluation metrics and the alignment between predicted and actual values in the visualization.

Support Vector Machine (SVM) Regressor Model Evaluation



The Support Vector Machine (SVM) Regressor was employed to predict rental prices based on selected features. However, the model demonstrated the following performance:

1. **Mean Squared Error (MSE):** The SVM Regressor model produced an MSE of approximately 99,759.79, depicting the average squared difference between predicted and actual rental prices. Higher MSE values suggest decreased predictive accuracy.

2. **Mean Absolute Error (MAE):** With an MAE of around 255.20, the SVM Regressor model showcased the average absolute difference between predicted and actual rental prices. Higher MAE values indicate reduced accuracy in predictions.

3. **R-squared (R^2) Score:**The R-squared value was approximately -0.0094, indicating that the model's predictive ability did not perform better than a model predicting the mean of the target variable. A negative R-squared value signifies that the model doesn't fit the data well. The Support Vector Machine (SVM) Regressor model appears to struggle in accurately estimating rental prices based on the selected.

Result

Model	MSE	MAE	R-square
Linear Regression	20,015.97	135.40	0.7975
Decision tree	24,456.25	101.25	0.7526
Random Forest	29,130.62	152.24	0.7053
KNN	24,685.50	125.75	0.7502
SVM	99,759.79	255.20	-0.0094

Regression Model Performance Comparison

Observations:

- Linear Regression demonstrates the best performance among the models, showcasing the lowest Mean Squared Error (MSE) and a relatively high R-squared value, indicating better predictive accuracy.
- Decision Tree and K-Nearest Neighbors (KNN) models exhibit comparable performance in terms of MSE and R-squared values, offering reasonable predictive capability.
- Random Forest performs slightly worse than the Decision Tree and KNN models, with a higher MSE and relatively lower R-squared.
- Support Vector Machine (SVM) displays the least favorable performance among the models, characterized by significantly higher MSE, MAE, and a negative R-squared value, suggesting poor predictive ability compared to the other models.

Conclusion:

Based on the evaluation metrics, Linear Regression appears to be the most effective model for predicting rental prices in this scenario. It demonstrates the lowest errors and the highest R-squared value, indicating a better fit to the data compared to the other models. Conversely, Support Vector Machine (SVM) seems ill-suited for this task due to its substantially higher errors and a negative R-squared value.

Future Recommendations:

1. Feature Engineering: Consider exploring additional features that could better capture the nuances affecting house rental prices, such as neighborhood amenities, crime rates, or property age.

2. Ensemble Techniques: Experiment with ensemble methods like Gradient Boosting or Stacking, which combine multiple models to potentially improve predictive performance.

3. Hyperparameter Tuning: Optimize model parameters using techniques like Grid Search or Random Search to enhance model accuracy.

4. Data Augmentation: Increase dataset size by collecting more diverse and comprehensive data, which might improve model generalization.

5. Localized Models: Develop models specific to different neighborhoods or regions to account for varying rental market dynamics across locations.

6. Time-Series Analysis: If applicable, analyze historical rental data over time to identify trends or seasonal patterns that could impact future prices.

7. Domain Expertise: Collaborate with real estate experts or professionals to gain insights into additional factors that influence rental prices.

8. Regular Model Updates: Periodically retrain models with new data to ensure they remain relevant and accurate in reflecting changing market conditions.

9. User Interface Development: Consider building a user-friendly interface or application that allows users to predict rental prices based on selected features.

References

[1]. Kavousian, A., & Maas, H. G. (2016). House price prediction: Parametric versus semi-parametric spatial hedonic models. Computers, Environment and Urban Systems, 57, 82-91.

[2]. Thiemann, C., & Rendall, A. D. (2017). Machine learning in real estate: Predicting the price of houses in Atlanta.Journal of Housing Economics, 38, 62-74.

[3]. Saggion, H., & Assaleh, K. (2019). House Price Prediction Using a Hybrid Model Based on Machine Learning and Fuzzy Logic. Applied Sciences, 9(14), 2882.

[4]. Adom, A., Okyere, M. A., Pong, K. Y., & Akuffo, F. O. (2018). House Price Prediction: An Exploration of the Housing Market in the United States.Sustainability, 10(6), 1941.

[5]. Chitra, V., & Thangaraj, M. (2020). House Price Prediction Using Machine Learning Algorithms: A Case Study of Bangalore City. Procedia Computer Science, 171, 205-212.