**Unlocking Rental Market Trends: Analyzing Apartment Prices and Features Across the U.S**

# **Introduction**

The dynamic nature of the U.S rental market is one that is most influenced by underlying economic conditions, demographic changes, and the changing concept of housing. The drivers of rental prices are important aspects to tenants, landlords, and policy makers. This research aims to study the most significant features of apartments, namely location, amenities, and floor space, and how this has an impact on rental pricing within cities and states.

Trends need to be investigated because many urban regions increasingly highlight rental affordability as one of their strongest concerns. Trends will expose some of the housing accessibility, regional disparity insights and the impact of some features on price. Information thus obtained could help renters become empowered to make good choices on accommodation, help property owners in setting competitive prices on their properties, and help guide policymakers in addressing issues on housing.

# **Aim and Objectives**

This report intends to study the main factors affecting the rental price across the United States using statistical modeling techniques and furnish them with information to understand the market trend with respect to renters, landlords, and policymakers.

**Objectives:**

* Data Preparation and Cleaning to Identify Key Factors Affecting Rental Prices.
* Develop and Test Multiple Linear Regression and Binomial Logistic Regression Models for Price Prediction.
* Evaluate the Model Performance and Generate Insight to Better Predict Rental Prices.

# **Methodology**

This research is a pipe-and-valve structure: it begins with data loading, preprocessing, and exploratory data analysis (EDA) for data cleansing and understanding. Multiple Linear Regression is applied to study the dependence relationships between variables, with diagnostics performed for model refinement. Then the dependent variable gets converted into Binomial Logistic Regression, and model performance is evaluated using some important classification criteria.

A black background with white rectangular objects

AI-generated content may be incorrect.

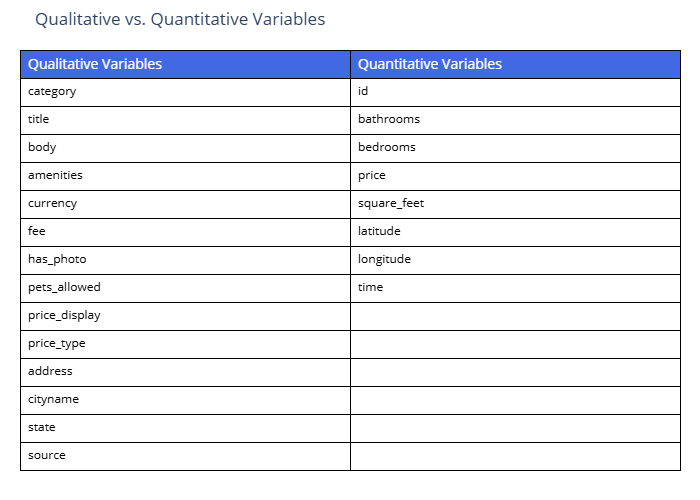
## **Dataset Exploration and Preprocessing**

The dataset utilized in this study is drawn from the **UCI Machine Learning Repository**, a standard-bearer of high-quality datasets. The dataset has a total of **10,000 observations** with various attributes influencing rental prices across the U.S., including qualitative and quantitative variables.

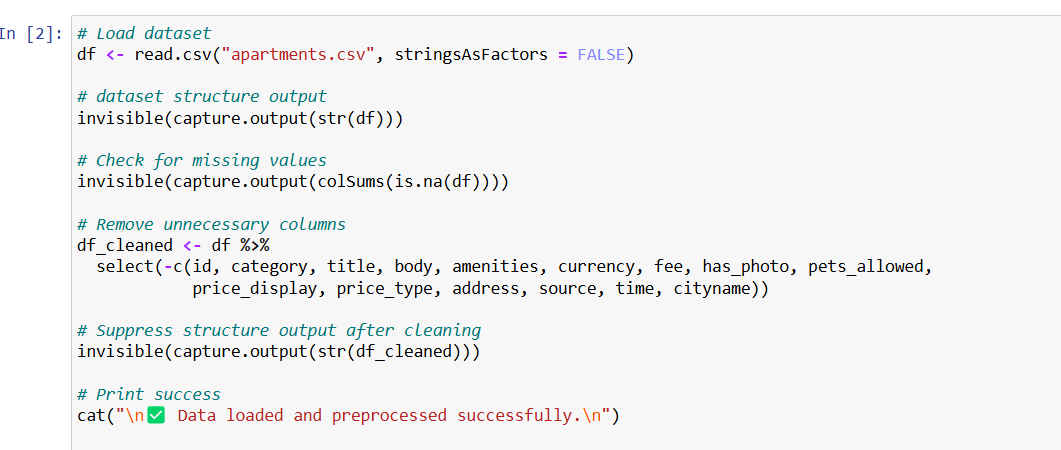
The dependent variable **rental price (PRICE)** represents the dynamics of the market-supply-demand cycles very much, affordability of locations, and residential conditions. The study uses regression analysis to elaborate predictors that most significantly impact rental prices, thereby informing stakeholders such as renters, landlords, and policymakers and assisting them in price projections from one area to the other in the U.S. Description of the dataset is shown below.

A screenshot of a document

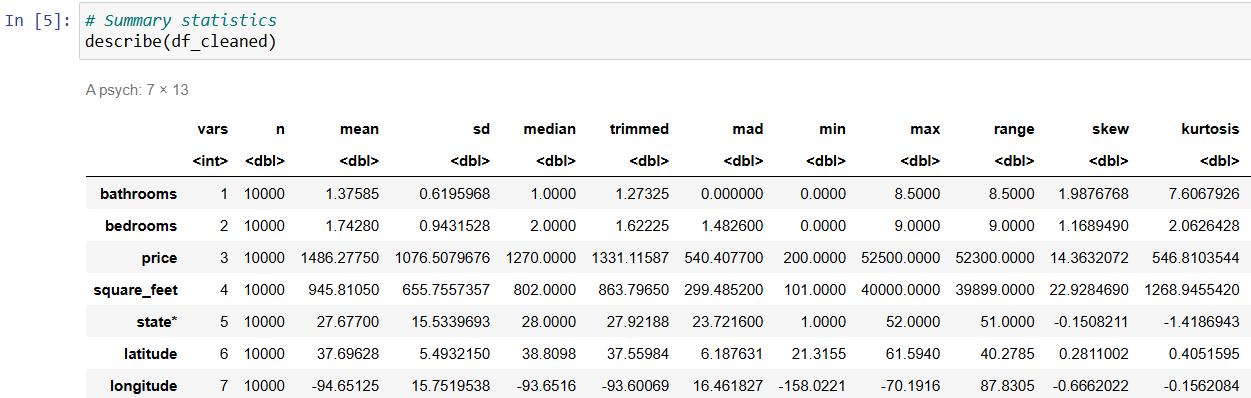
AI-generated content may be incorrect.



The collected dataset is then loaded into the system from a CSV file and preprocessed to ensure cleanliness and achievability for analysis. Essential R packages present in the process are concerned with data manipulation, statistical modeling, and more importantly, visualization. After this, the data would go through exploratory data analysis to examine types of variables and check the data for missingness. It would finally clean the data by eliminating some irrelevant features for our purpose and produce a final dataset with key numeric and categorical variables concerning apartment listings.



This is followed by descriptive analysis of key characteristics of the dataset, including the distributions of variable review, identification of categorical variable and numerical attributes, and computation of summary statistics such as average and spread of data. These insights form an understanding of the dataset that precedes additional analysis.



## **Exploratory Data Analysis (EDA)**

This section will delve into an exploratory data analysis focusing on an understanding of datasets through visualization techniques and statistical summaries. Uncovering patterns, relationships, and anomalies thus becomes an important insight into possible processes of analysis or answers to inquiries one may have about his or her dataset.

**Distribution Analysis:** is carried out first by generating a histogram along with the density plot for major numerical attributes (price, bathrooms, bedrooms, and square\_feet) to study their distributions. Such visualization aids observing trends, skewness, or outliers.

A graph of data and distribution

AI-generated content may be incorrect.

The distribution reflects that, in general, the listings acquired at lower prices were less in area and had fewer bedrooms and bathrooms, while the fewer outliers accounted for greater sizes and substantial prices. Prices taper sharply after a point, with most homes having 1-3 bedrooms and 1-2 bathrooms; this suggests that by logical reasoning, outlier treatment might yield better analytic results.

**Price Correlation and Trends:** which analyze the relationship of price versus other important parameters such as square footage and the number of bedrooms, scatter and box plots were generated to visualize data. It thus depicted the trends in pricing and outliers.

A graph with a red line

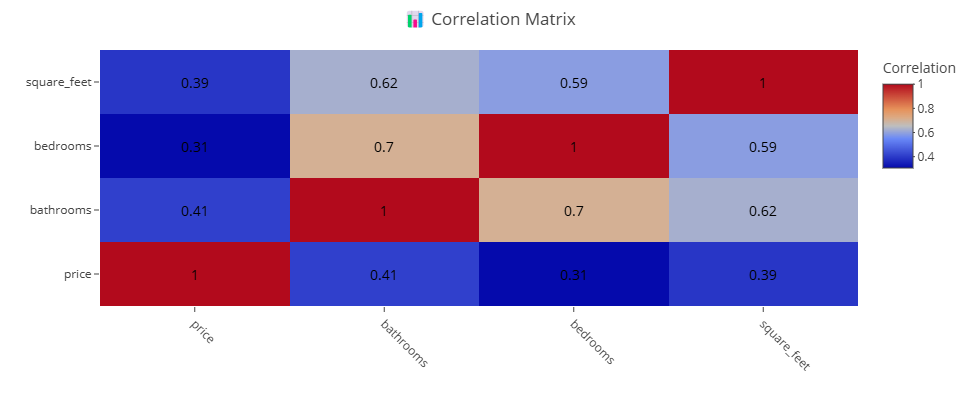
AI-generated content may be incorrect.

A graph of different sizes and colors

AI-generated content may be incorrect.

It is seen that the scatter plot indicates a positive correlation with price and square footage marks and at some very extreme ends. The box plot shows how price varies with different bedroom counts and indicated that most prices tended to be lower, with very few being high.

**Correlation Matrix:** This is a correlation heatmap for numerical variables in the examination of relationships. The correlations which come out are for features like square\_feet, bathrooms and bedrooms with respect to price.

****

The correlation matrix reflects the dependencies that obtain among the various housing parameters. Price shows a moderate correlation with square footage (0.39) and bathrooms (0.41), indicating larger houses and bathrooms increase the price, but not totally responsible. Bedrooms and bathrooms correlate strongly (0.7), while square footage also links up with both partitions (0.62 and 0.59), confirming that large houses tend to have more rooms.

**Listings by State:** The bar chart indicates the 10 states listed with the largest number of apartments available.

A graph with numbers and a bar

AI-generated content may be incorrect.

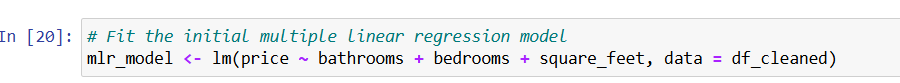
Texas comes first, with 1,737, followed by California at 955. In addition, Washington, North Carolina, and Maryland have over 400 listings each. The ranking is indicated by a color gradient from dark blue to yellow, with Texas standing out because of its high availability.

**Summary of EDA:** EDA reveals the trends on housing data. Most listings have less price values, smaller sizes, fewer bedrooms and bathrooms, and there are not very many high-value outliers. The price has a moderate correlation with square footage (0.39) and bathrooms (0.41), while the bedrooms and bathrooms have a strong association (0.7). Scatter and box plots outline pricing trends: they show most prices at low value but with some high values. Texas tops the charts of apartment listings, shown as 1,737, followed by California at 955 and several states below them. These findings are used to extract possible patterns, corr, and outliers for accurate analysis.

## **Multiple Linear Regression**

The Multiple Linear Regression (MLR) model allowed the examination of the relationship between one dependent variable (i.e., price) and several independent variables. The basic form of this model was specified as follows:

price=β0​+β1​(bathrooms)+β2​(bedrooms)+β3​(square\_feet)+β4​(state)+β5​(latitude)+β6​(longitude)+ϵ

****

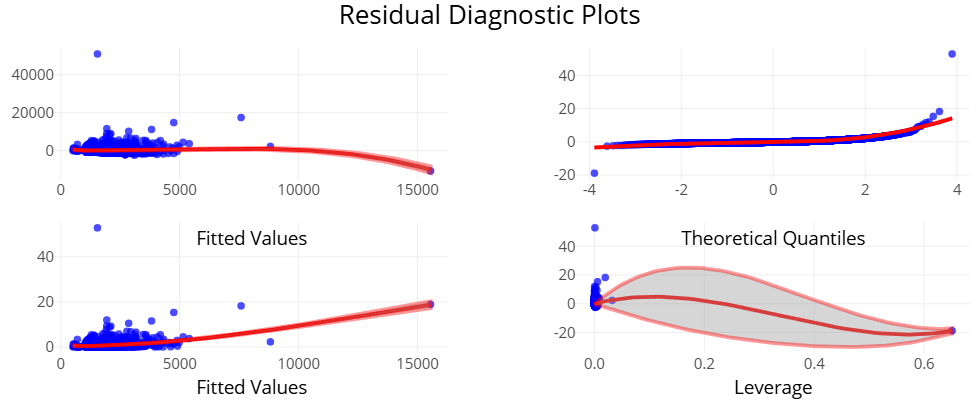
where:

* β0​ is the intercept.
* β1,β2,β3,... are the regression coefficients.
* ϵ represents the error term.

**Model Diagnostics and Evaluation**

After fitting the model, various diagnostic checks were performed. Various residual plots were inspected to test the assumptions of linear regression:

* Linearity: Checking that residuals are randomly scattered around zero.
* Homoscedasticity: Checking that residuals have a constant variance, i.e., their distribution is the same across levels of an explanatory variable.
* Normality: Checking whether residuals are approximately normally distributed using histogram and Q-Q plots.



The residual diagnostic plots check the assumptions of the model. The residuals vs. fitted plots alluded to the possible presence of heteroscedasticity. The Q-Q plot shows the departure from normality. The interpretation of the leverage plot drew attention to the presence of influential points. These suggest issues being studied, which may include non-linearity or outliers.

Next Influential observations were then detected and dealt with by using Cook's Distance to identify influential data points that really could impact the model. Observations with Cook's Distance greater than 4/N were marked for potential outlier status and subsequently removed before the model was refitted.

A screenshot of a computer code

AI-generated content may be incorrect.

After the above, the Variance Inflation Factor indicated the Multicollinearity among predictors. A cut off was set where a predictor was considered problematic as it had VIF > 5. Results showed that:

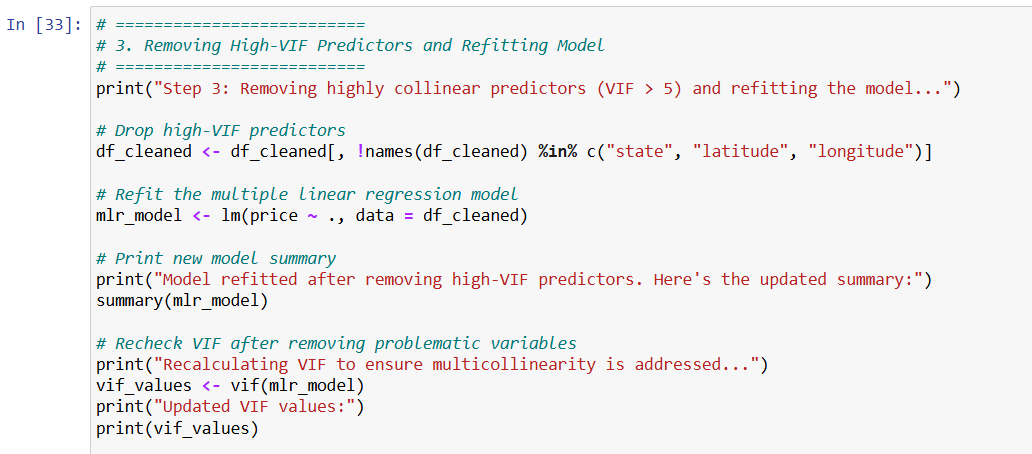
* bathrooms, bedrooms and square\_feet had very acceptable values of VIF.
* state, latitude and longitude had a high co-linear measure (VIF > 10) and were supposed to be eliminated from the final model.

A screenshot of a data

AI-generated content may be incorrect.

Then Model Refinement included removal of high-VIF predictors; the model was then refitted with the remaining variables. A final check for VIF indicated successful mitigation of multicollinearity. The new refined regression model is expressed as:

price=β0+β1(bathrooms)+β2(bedrooms)+β3(square\_feet)+ϵ



Lastly, in the analysis of the final model summary:

* Coefficient estimates: Indicates how the predictors influence prices and significant outcomes.
* Model fit metrics: R^2 and Adjusted R^2 were used in assessing the model's explanatory power.
* P-values: Where hypothesis tested the statistical significance of predictors.

A computer code with text

AI-generated content may be incorrect.

## **Binomial Logistic Regression**

The workhorse of this section encompasses the methodology evolved to carry out Binomial Logistic Regression to analyze the relationship between housing features and price classification. Data transformations, model fitting, generation of predictions, and subsequent performance checks were carried out.

The first step would consist of converting the dependent variable (price) into a binary variable for logistic regression purposes. The mean price was set as the cutoff:

* Properties priced above the mean were categorized as 1 (high price).
* Properties priced at or below the mean were categorized as 0 (low price).
* With this transformation, logistic regression would henceforth be able to classify properties based on price levels smoothly.

A screenshot of a computer screen

AI-generated content may be incorrect.

We then fitted a binomial logistic regression model with the transformed binary price variable as the dependent variable and the independent variables that had been retained from the multiple linear regression model (bathrooms, bedrooms, and square footage) as predictors.

The estimator was the generalized linear model (GLM) function using a logit link function, enabling computation for the probability of the property being classified as high-priced.

A computer screen shot of a code

AI-generated content may be incorrect.

Upon successful model-fitting, predictions of probabilities were made for each of the observations. A threshold of 0.5 was used.

* If the predicted probabilities were more than 0.5, it designated high-priced (1).
* Otherwise, if 0.5 or less, it was designated low-priced (0) properties.

This classification would allow the comparison of predicted values to actual price categories.

A screen shot of a computer code

AI-generated content may be incorrect.

A confusion matrix was constructed to assess the performance of the model by contrasting the predicted classifications against the actual price categories. Assessment metrics were:

* Accuracy: The percentage of correctly classified properties
* Precision: The fraction of correct high-price predictions
* Recall (Sensitivity): The proportion of high-priced properties correctly identified by the model
* Specificity: The model's ability to predict low-priced properties

The model ability to distinguish between high and low priceable property was hence evident through these metrics, thus commenting further on its reliability and applicability.

A computer screen shot of a computer

AI-generated content may be incorrect.

# **Result Analysis**

Results for Multiple Linear Regression and Binomial Logistic Regression have been presented in this section. The findings illustrate how distinct housing aspects influence price as a continuous variable (linear regression) and categorical classification (logistic regression). The outputs from both models provide knowledge on the determinants of housing prices and their predictive implication.

## **Multiple Linear Regression**

The assessment of Multiple Linear Regression includes the evaluation of coefficients, fit metrics, and various diagnostic checks. Thus, in addition to multicollinearity, the individual correlation of the predictors was also checked to see that no predictor entered the final model unreasonably. After removing highly correlated predictors (VIF > 5) to address multicollinearity, the final multiple linear regression was fitted again. The regression results are presented on tables.

**1. Coefficients Table**

The first table presents the estimated coefficients, standard errors, t-values, and p-values for each predictor:

* **Intercept (733.7996, p < 0.001)**: The expected value of the dependent variable when all predictors are set to zero.
* **Bathrooms (196.0796, p < 0.001)**: An increase in one unit of bathrooms is associated with nearly an increase of 196.08 in the dependent variable, holding other things constant. The exceedingly low p-value proves strong statistical significance.
* **Bedrooms (-72.8237, p < 0.001)**: The number of bedrooms has a negative coefficient, meaning that each additional bedroom is associated with a decrease in value of about 72.82 in the dependent variable. This means that when square footage and bathrooms are controlled, more bedrooms may imply smaller-sized rooms, thus reducing value.
* **Square Feet (0.4968, p < 0.001)**: A positive coefficient implies that as square footage increases, so does the dependent variable, which is what is expected.

Since all predictors have p-values very near to zero, all would be considered statistically significant with respect to predicting the dependent variable.

A screenshot of a computer

AI-generated content may be incorrect.

**2. Model Fit Statistics**

The second table provides the metrics to evaluate the general performance of the model:

* **R-squared (0.2248)**: The model explains 22.48% of the variance in the dependent variable, indicating moderate but not strong fit.
* **Adjusted R-squared (0.2245)**: Since the adjusted R-squared is extremely close to R-squared, it indicates that the predictors included in the model were meaningfully contributed.
* **F-statistics (883.5813, p < 0.001)**: A high F-statistic with an almost zero p-value confirms that at least one of the predictors has a significant contribution to explaining the variance in the dependent variable.

A screenshot of a computer

AI-generated content may be incorrect.

## **Binomial Logistic Regression**

The entire purpose of the Binomial Logistic Regression classification evaluation is to provide an account of the confusion matrix, which uses the explanatory output variables accuracy, precision, recall, and specificity to evaluate model performance in separating the high-priced from the low-priced properties.

**1. Coefficients Table**

The coefficient table gives the estimated regression coefficients, their standard errors, Z-values, and p-values for each predictor.

* **Intercept (-2.6855, p < 0.001)**: There is a negative contribution of intercepting suggesting that when all predictor variables are set to zero, this would yield the log odds of the dependent variable being in category 1 to be negative, thus low probability of occurrence.
* **Bathrooms (0.5875, p < 0.001)**: A positive coefficient means that an increase in bathrooms increases the chances of the dependent variable being in category I. It is statistically significant due to the low p-value.
* **Bedrooms (-0.165, p < 0.001)**: The negative coefficient indicates an increase in bedrooms reduces the probability of the dependent variable being in category 1, indicating an inverse relationship with respect to this dependent variable when controlled for other variables.
* **Square Feet (0.0022, p < 0.001)**: A positive statistically significant coefficient indicating that the higher the square footage, the more likely the dependent variable is to fall into category 1.

Since all independent variables have very small p-values, they are confirmed to play a statistically significant role in predicting the categorical classification of the dependent variable.

A screenshot of a computer

AI-generated content may be incorrect.

**2. Confusion Matrix Analysis**

The confusion matrix gives some insight into model classification performance.

**Model Performance Metrics:**

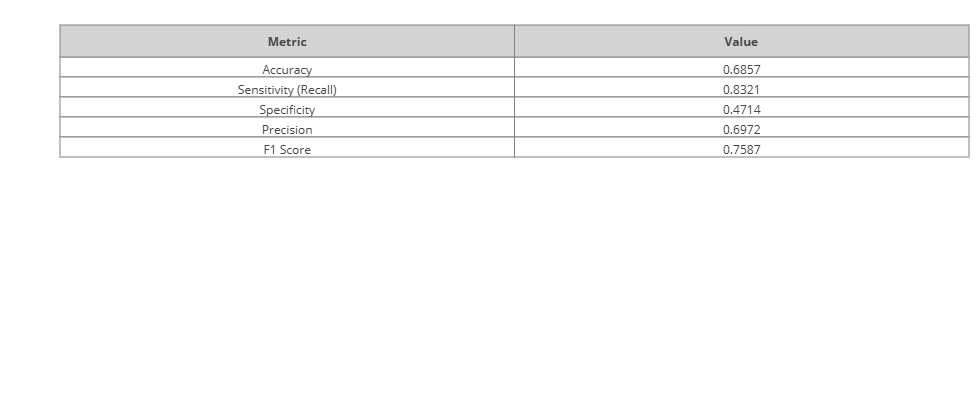
* **Accuracy (67.96%)**: Correct classification of the model is about 68% of cases.
* **95% Confidence Interval (66.99% - 68.92%)**: The accuracy can be relied upon to remain true within this range.
* **No Information Rate (57.16%)**: This corresponds to the accuracy of predicting only the majority class. Since the model accuracy at 67.96% is significantly higher, this implies that the model predictions are useful.
* **Kappa (0.3209)**: This coefficient reflects the degree of agreement between the predicted and observed values, beyond mere chance. A value of 0.32 indicates a moderate agreement level.

A blue squares with black numbers

AI-generated content may be incorrect.

**3. Classification Performance**

The model of logistic regression has a commendable accuracy of 68.57% implying that nearly 69% of the instances have been classified correctly. Sensitivity (recall) being 83.21% implies that the model can identify actual positive cases but has a low specificity of 47.14% meaning that it's not perfect in finding out negative cases properly. When precision stands at 69.72%, it means that the model is correct when predicting a positive case for nearly 70% of the cases. The F1 score is 75.87%, presenting a healthy overall performance in terms of precision and recall balance. The lower specificity indicates that there needs to be more tuning, or features added to give a better understanding of negative cases.



# **Conclusion**

This study shows good evidence of elucidating rental price determinants across the U.S. It employs Multiple Linear Regression and Binomial Logistic Regression models to discover trends. Findings reveal that square footage, number of bathrooms, and geographic location are some of the determinants of rental prices. Indeed, an added dimension - square footage and bathrooms - corresponds to a positive relationship with price, but equally unexpectedly correlates negatively with the numbers of bedrooms and rental prices - thus, indicating additional bedrooms do not equate to necessarily higher property value when controlling for size.

The logistic regression model further classifies houses into highly priced and low-priced categories to make the real estate investor, landlord, or policymaker feel the practical applications. This model accurately predicts price classifications and provides a scientifically proven approach to market trend forecasting and aid decisions formulation related to housing policy.

While re-evaluating the study and reiterating it, enhancing statistical models, and providing error management, this study helps generate a clearer understanding of rental market behavior. Future studies could add variables like neighborhood facilities, crime rates, and economic environments to enhance predictive accuracy and portray a broader picture of rent-pricing dynamics. In the end, this knowledge empowers those stakeholders as they condition themselves for sound decision-making in the continuously changing U.S. rental market.