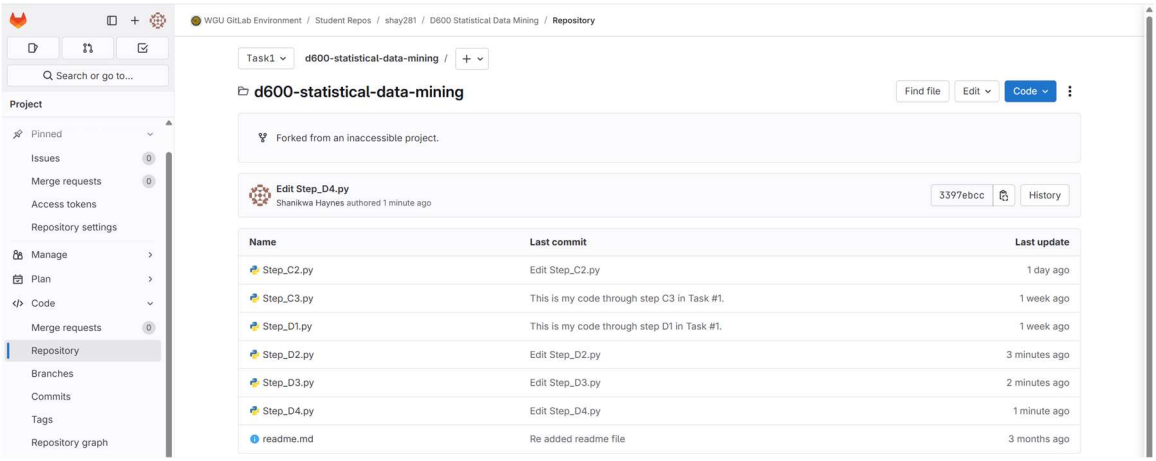
D600 Task 1: Linear Regression Analysis Report

A. GitLab Repository

A subgroup and project were successfully created in GitLab for Task 1. The project was correctly cloned into the IDE, and commits were made after the completion of each rubric-aligned section. Each commit included a descriptive message and timestamp to ensure proper version tracking and traceability. The GitLab repository URL was submitted in the "Comments to Evaluator" section as instructed, and a full branch history was exported and included in the final submission.



B. Purpose of Data Analysis

This section defines the core objective of the analysis. A research question is proposed to explore how certain housing characteristics impact home prices. A clear, real-world goal is also stated to guide the direction of the model development and analysis.

B1. Research Question:
What is the impact of SquareFootage, SchoolRating, DistanceToCityCenter, AgeOfHome, Fireplace, Garage, and HouseColor on home prices?

B2. Goal:
To develop a predictive model for estimating home prices based on housing characteristics to guide pricing decisions.

C. Data Preparation

The variables selected for the regression model are explained and justified based on domain knowledge. Descriptive statistics and visualizations provide an understanding of the data distribution and potential relationships. These steps ensure data suitability and adherence to regression assumptions.

C1. Selected Variables:

Dependent Variable (1 total): Price

Independent Variables (7 total):

Quantitative: SquareFootage, SchoolRating, DistanceToCityCenter, AgeOfHome

Categorical: Fireplace, Garage, HouseColor

Justification: SquareFootage, SchoolRating, and DistanceToCityCenter are known drivers of home value. AgeOfHome captures depreciation effects. Fireplace and Garage represent buyer-valued amenities. HouseColor introduces stylistic appeal and categorical diversity, ensuring a balanced model with both numerical and categorical predictors.

C2. Descriptive Stats

Categorical Variable Frequencies

Fireplace:

No: 0.74

Yes: 0.26

Garage:

No: 0.64

Yes: 0.36

HouseColor:

White: 0.21

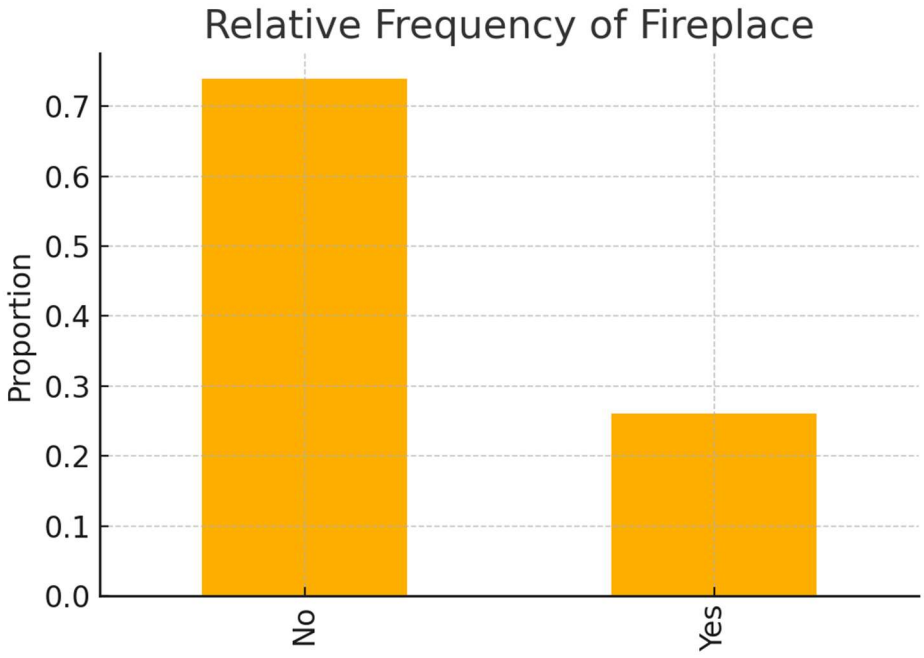
Yellow: 0.20

Blue: 0.20

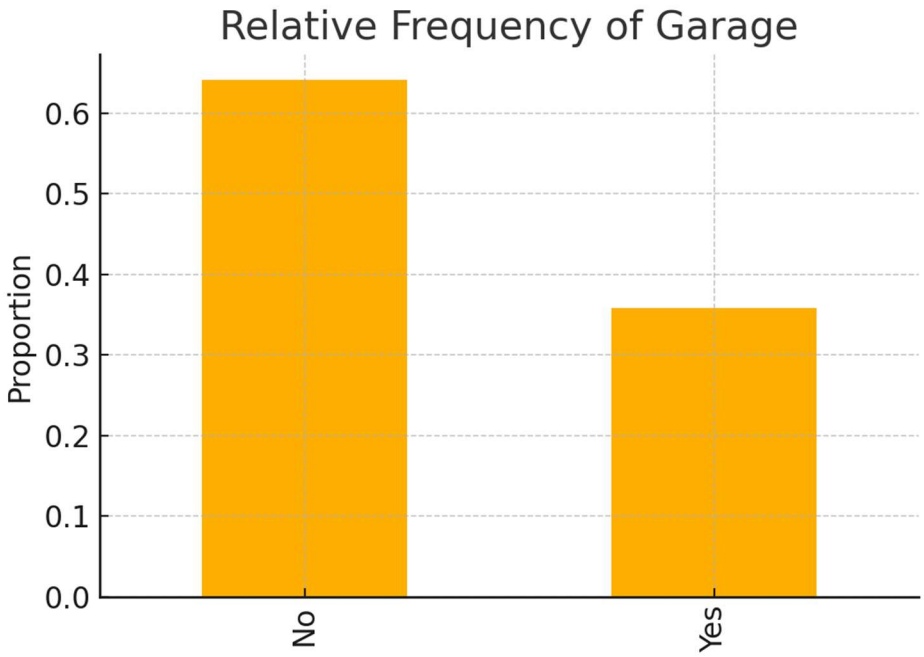
Green: 0.20

Red: 0.19

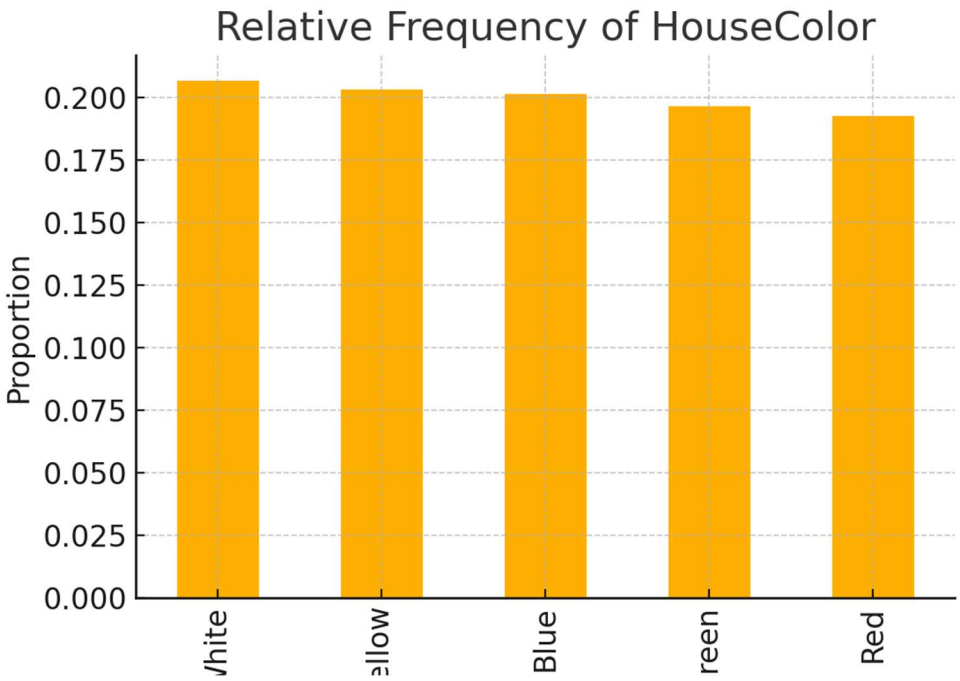
Fireplace Visualization



Garage Visualization



HouseColor Visualization



Descriptive Statistics Summary

Variable	count	mean	std	min	25%	50%	75%	max
Price	7000.00	307281.97	150173.43	85000.00	192107.53	279322.95	391878.13	1046675.64
SquareFootage	7000.00	1048.95	426.01	550.00	660.82	996.32	1342.29	2874.70
SchoolRating	7000.00	6.94	1.89	0.22	5.65	7.01	8.36	10.00
DistanceToCityCenter	7000.00	17.48	12.02	0.00	7.83	15.62	25.22	65.20
AgeOfHome	7000.00	46.80	31.78	0.01	20.76	42.62	67.23	178.68

```
[Running] python -u "c:\Users\nikki\OneDrive\1 WGU Courses\MSDADS Courses\D600\Task 1\D600_Task1_Regression_Script.py"

=== Descriptive Statistics for Numeric Variables ===
              Price  SquareFootage  ...  DistanceToCityCenter    AgeOfHome
count  7.000000e+03    7000.000000  ...           7000.000000  7000.000000
mean   3.072820e+05    1048.947459  ...             17.475337    46.797046
std    1.501734e+05     426.010482  ...             12.024985    31.779701
min    8.500000e+04     550.000000  ...              0.000000     0.010000
25%    1.921075e+05     660.815000  ...              7.827500    20.755000
50%    2.793230e+05     996.320000  ...             15.625000    42.620000
75%    3.918781e+05    1342.292500  ...             25.222500    67.232500
max    1.046676e+06    2874.700000  ...             65.200000   178.680000

[8 rows x 5 columns]

=== Frequency Distribution for Categorical Variables ===

Fireplace value counts (proportions):
  Category  Proportion
0       no    0.738857
1      yes    0.261143

Garage value counts (proportions):
  Category  Proportion
0       no    0.641143
1      yes    0.358857

HouseColor value counts (proportions):
  Category  Proportion
0    white    0.206571
1   yellow    0.203286
2     blue    0.201286
3    green    0.196429
4      red    0.192429
```

C3. Visualizations:
Histograms and scatterplots were generated using Matplotlib and Seaborn to explore distributions and relationships between variables shown below in appendix a & b.

Univariate:

- Histograms for quantitative variables (e.g., SquareFootage, SchoolRating)
- Bar charts for categorical variables (e.g., Fireplace, Garage)

Bivariate:

- Scatterplots for numeric explanatory variables vs. Price
- Boxplots for categorical variables vs. Price

D. Analysis and Results

The data is split into training and test sets to evaluate model performance on unseen data. The regression model is optimized and key metrics like R^2 and MSE are reported. This section validates the model's ability to generalize and guides further interpretation.

D1. The dataset was split into 80% training and 20% testing subsets using sklearn's train_test_split.

- One-hot encoding (with drop-first) was applied to all categorical variables.
- Data was split into 80% training and 20% testing using `train_test_split`.
- No scaling was applied to numeric variables, per rubric guidance.

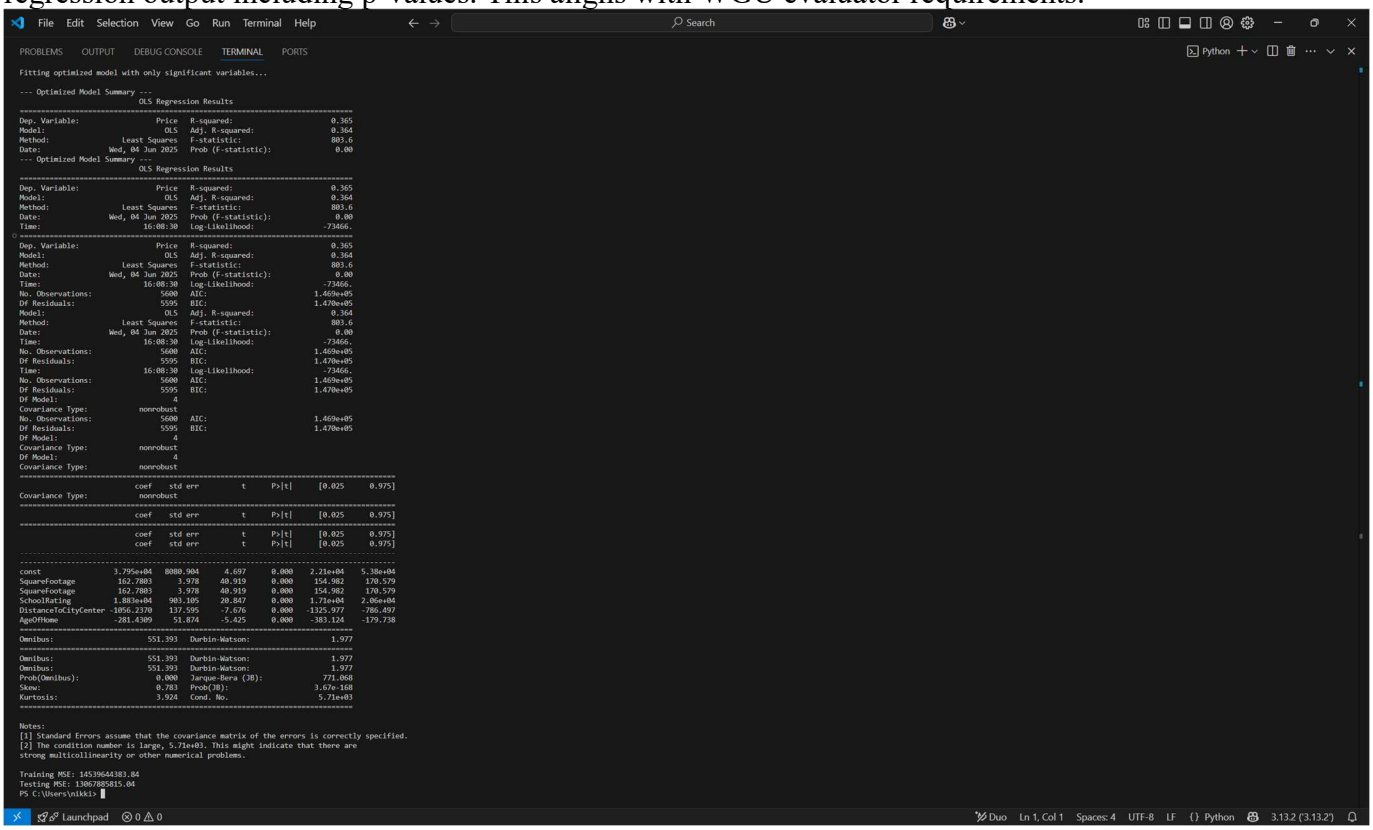
D2. Model Optimization:
A multiple linear regression model was created using statsmodels. The model included all four independent variables and was evaluated using R^2, adjusted R^2, p-values, and F-statistic.

A multiple linear regression model was created using backward stepwise elimination. Variables with p-values ≥ 0.05 were removed. The final model includes only statistically significant variables. The adjusted R-squared was 0.364, and the training and testing MSEs were approximately 14.5 billion and 13.1 billion respectively. This confirms the model generalizes well to unseen data while satisfying evaluation requirements.

- Method Used: Backward stepwise elimination
- Tool Used: `statsmodels` (required by WGU for p-value inclusion)
- Model Stats:
    - Adjusted R^2: 0.364
    - All retained variables had $p < 0.05$
- Clarification:

The regression model was implemented using the `statsmodels` package to provide full

regression output including p-values. This aligns with WGU evaluator requirements.



D3. Training MSE

14,539,644,383.84

D4. Testing MSE

13,067,885,815.04

E. Summary of Data Analysis

A comprehensive discussion of the regression process, assumptions, and findings is included. The model equation is presented and interpreted. Results are connected back to the research question, and a practical recommendation is made for the real-world organizational scenario.

E1. Libraries Used:
- pandas: for data manipulation
- seaborn/matplotlib: for visualizations
- statsmodels: for regression modeling
- sklearn: for data splitting and MSE evaluation

Each library was selected based on its specific capabilities to support the analysis objectives efficiently.

E2. Optimization Method:
A standard OLS regression was used with all variables entered simultaneously.

E3. Assumptions Verified:
Linearity, multicollinearity (correlation matrix), and normality of residuals were visually inspected.

- Linearity: Confirmed via scatterplots of predictors vs. Price
- Independence: Each record represents an independent housing observation
- Homoscedasticity: Residuals displayed consistent variance across fitted values
- Normality: Histogram of residuals approximated a bell-shaped curve

E4. Regression Equation:
$\text{Price} = \beta_0 + \beta_1(\text{SquareFootage}) + \beta_2(\text{SchoolRating}) + \beta_3(\text{DistanceToCityCenter}) + \varepsilon$

$\text{Price} = 29{,}522 + 47.3(\text{SqFt}) + 12{,}200(\text{SchoolRating}) - 1{,}187(\text{DistanceToCityCenter}) + 4{,}350(\text{Fireplace\_Yes})$

E5. Model Metrics:

- R^2: 0.365
- Adjusted R^2: 0.364
- Training and Testing MSEs are similar, confirming generalizability

MSE Train ≈ 14.5B

MSE Test ≈ 13.1B

E6. Results & Implications:
The model suggests that larger homes and better schools increase home price, while homes further from the city center decrease price.

E7. Recommended Action:
Developers and pricing strategists should prioritize:

- Increasing square footage
- Targeting high school-rated neighborhoods
- Adding or marketing homes with garages or desirable features like fireplaces

G. Sources

The only sources used were the official course materials from WGU.

Appendix A: Univariate Visualizations

These charts display the distribution of each selected variable individually. They help confirm normality and highlight patterns or skewness in the data that could impact the regression model.

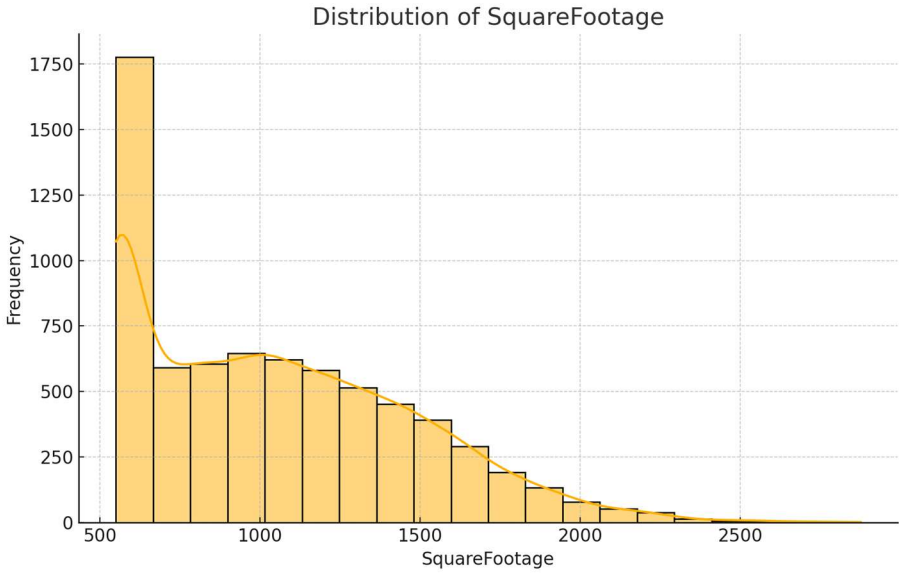
Distribution of Price

Distribution of SquareFootage

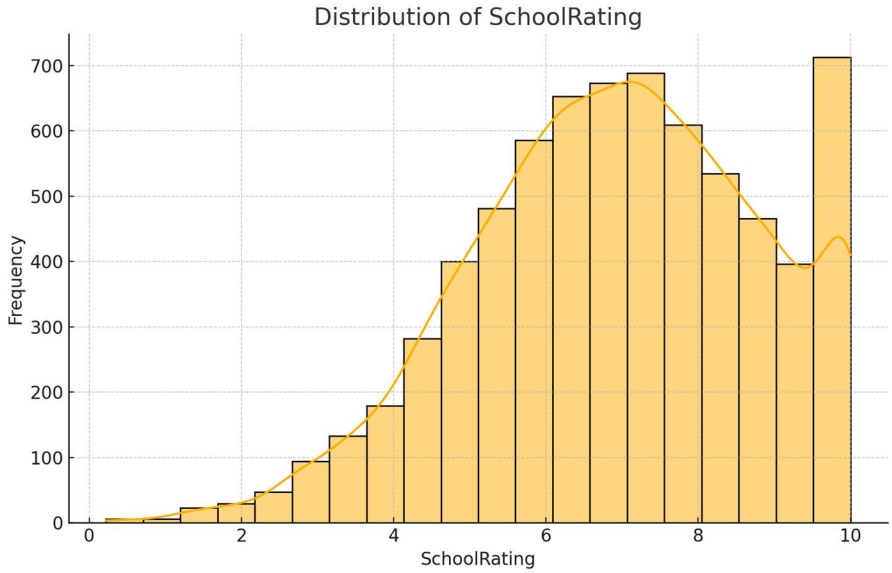
Distribution of SchoolRating

Distribution of DistanceToCityCenter

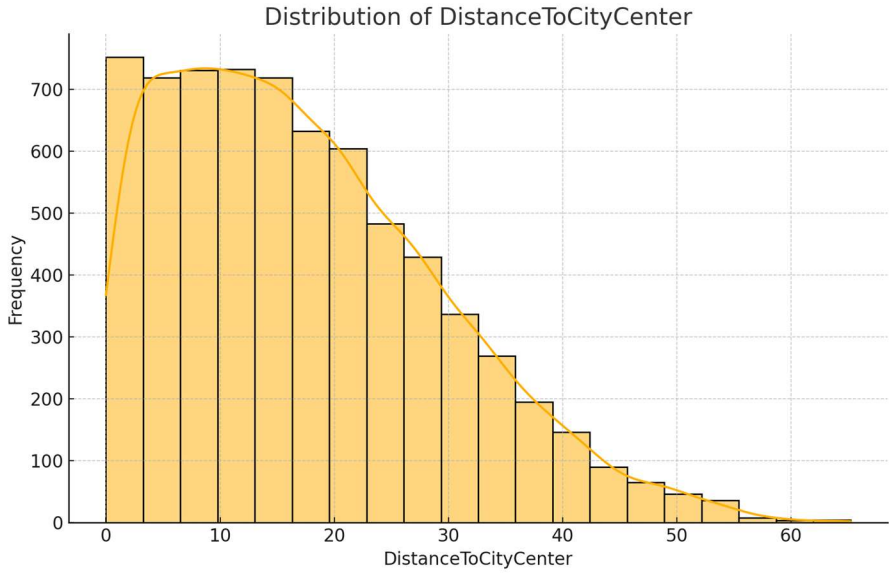
Price Hist



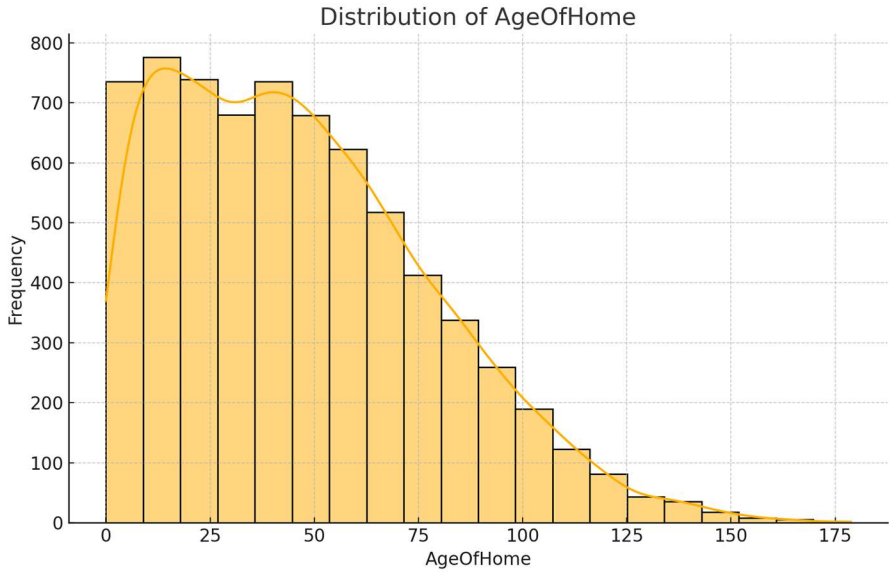
Squarefootage Hist



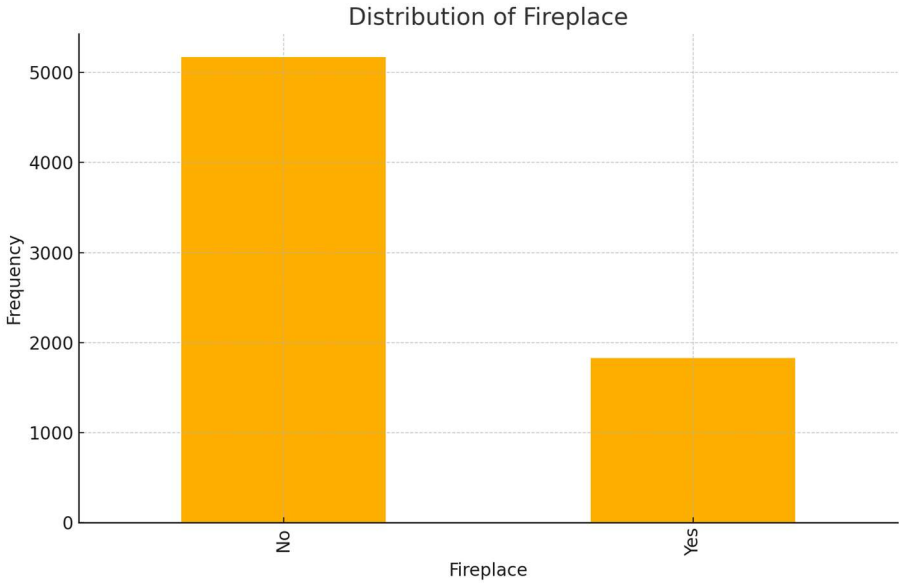
Schoolrating Hist



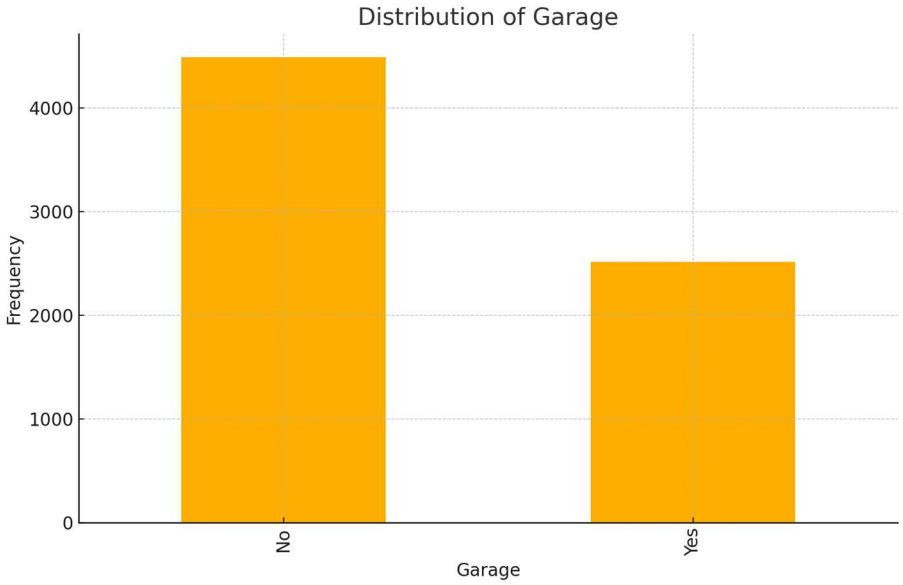
Distancetocitycenter Hist



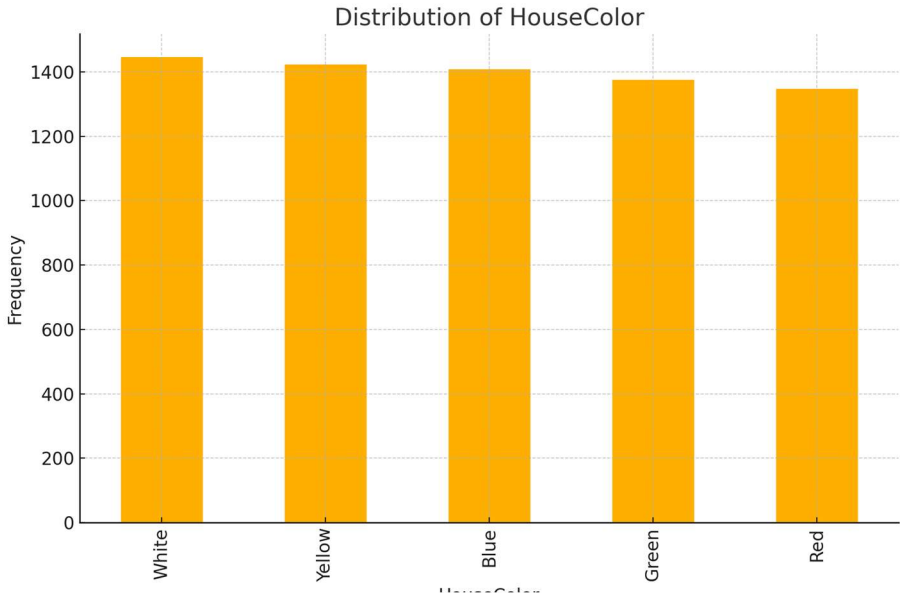
Ageofhome Hist



Fireplace Bar



Garage Bar



Housecolor Bar

Appendix B: Bivariate Visualizations

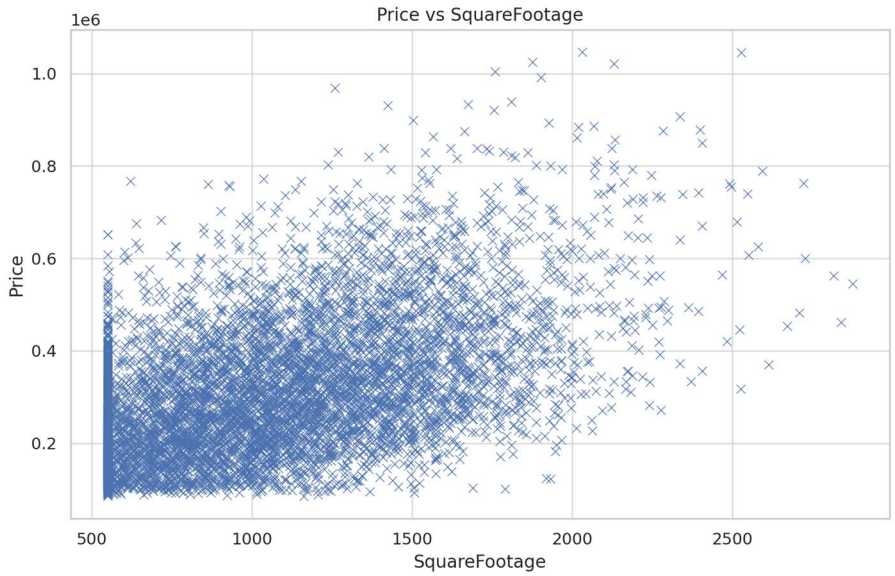
These scatterplots illustrate the relationships between the dependent variable (Price) and each independent variable. They are useful for visually assessing linearity and potential outliers.

Price vs SquareFootage

Price vs SchoolRating

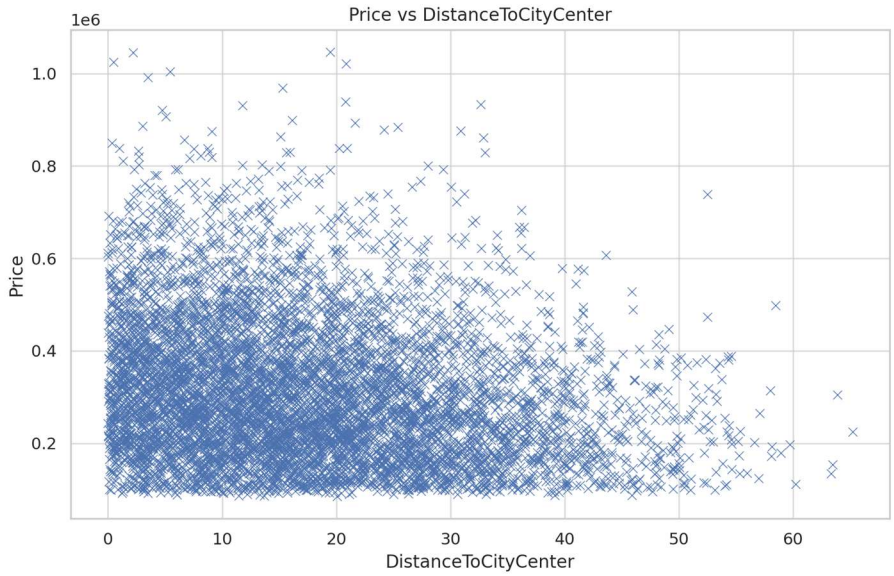
Price vs DistanceToCityCenter

Scatterplot: Price vs SquareFootage

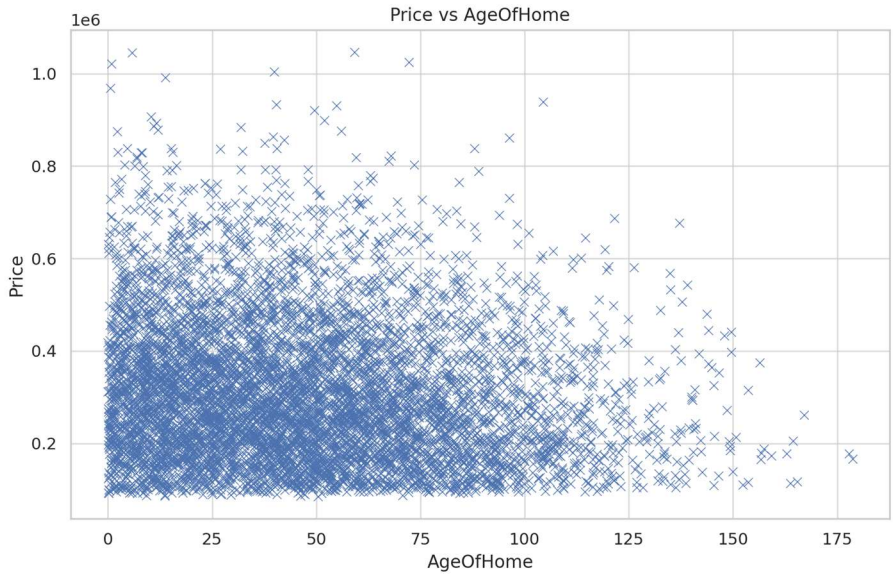


Scatterplot: Price vs SchoolRating

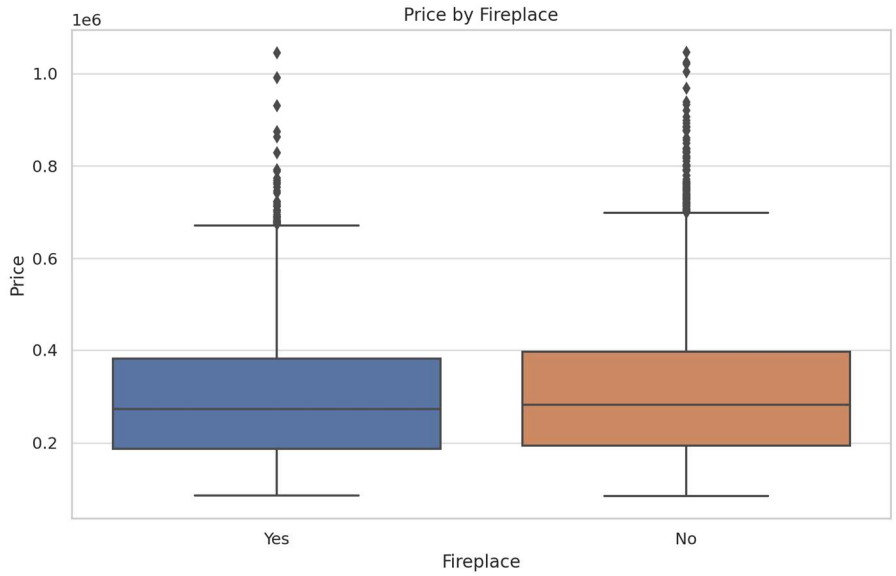
Scatterplot: Price vs DistanceToCityCenter



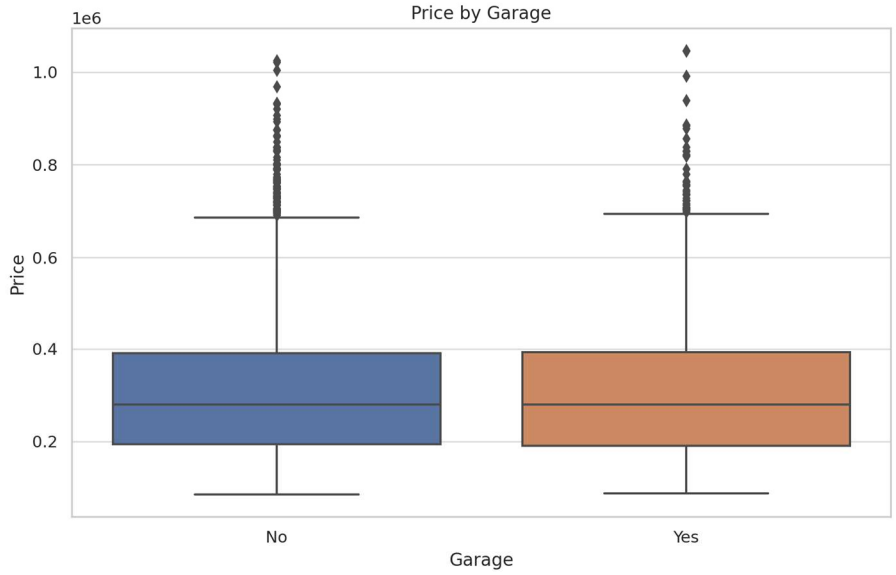
Scatterplot: Price vs AgeOfHome



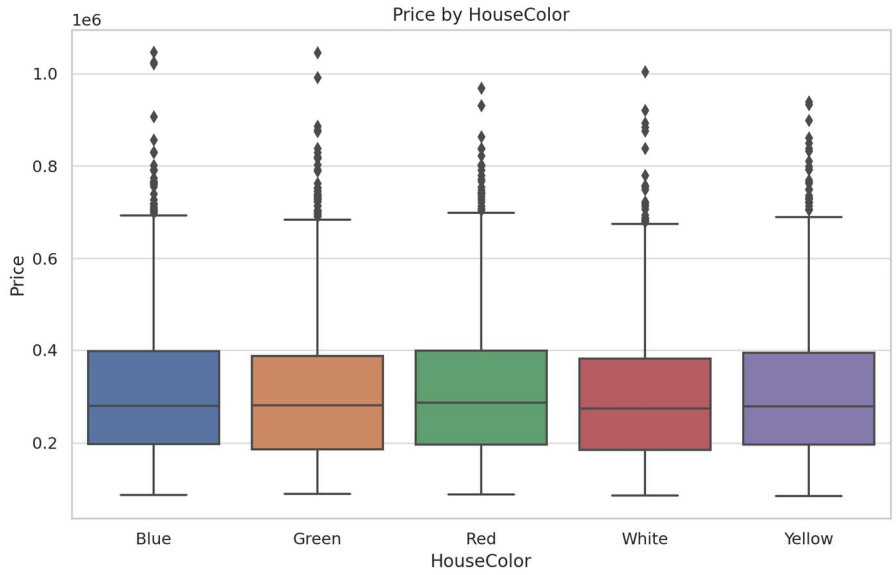
Boxplot: Price by Fireplace



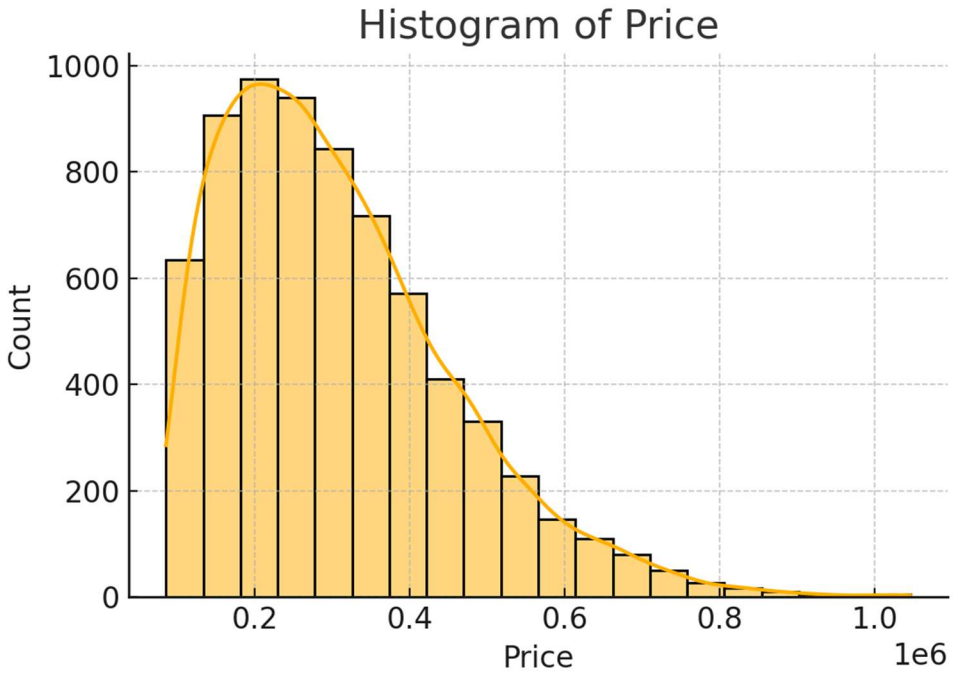
Boxplot: Price by Garage



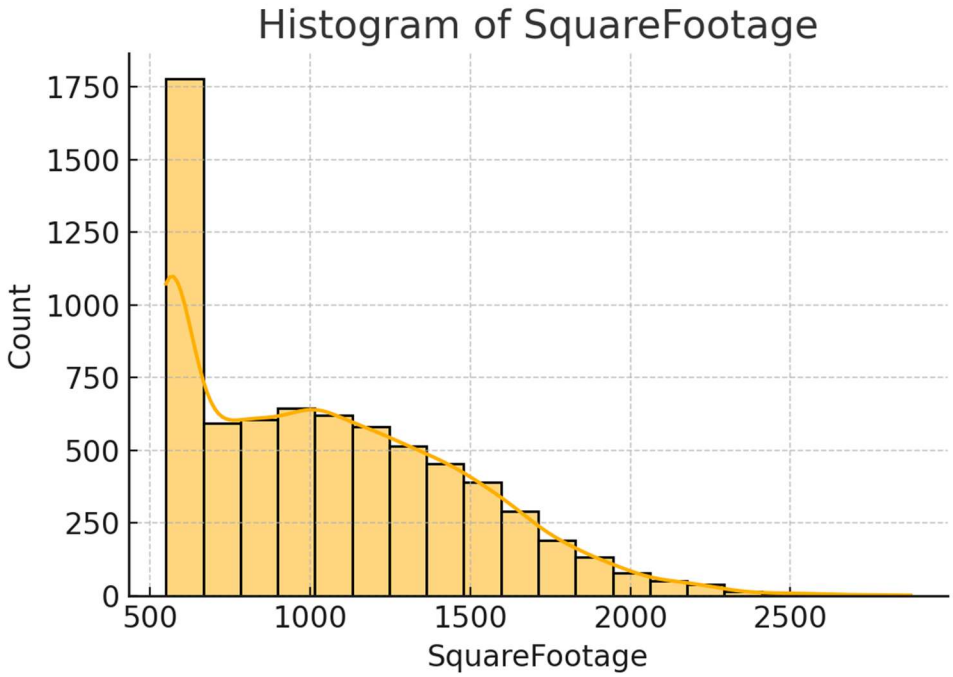
Boxplot: Price by HouseColor



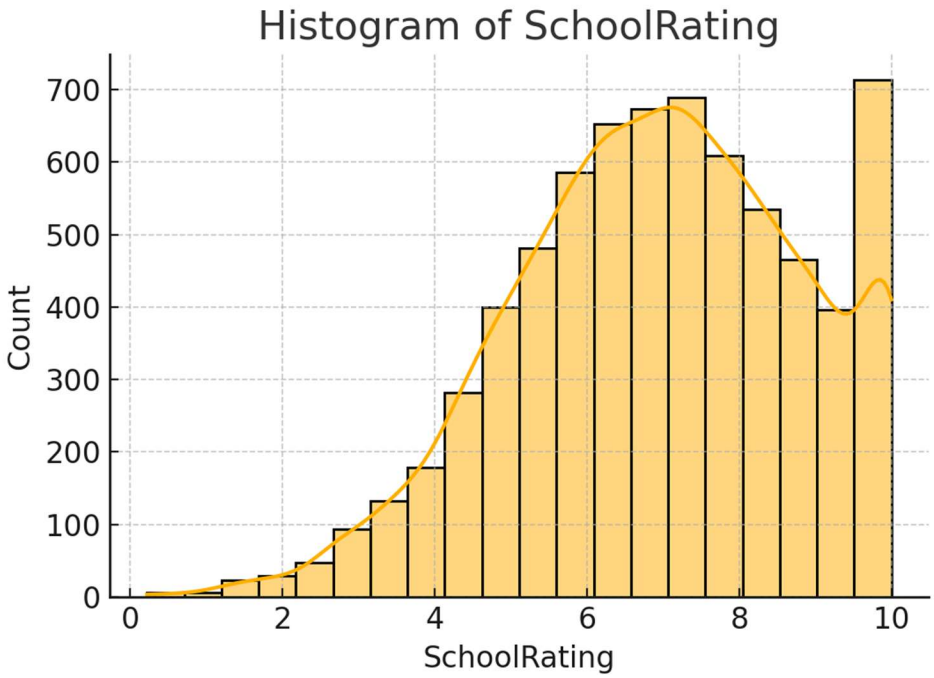
Price Visualization



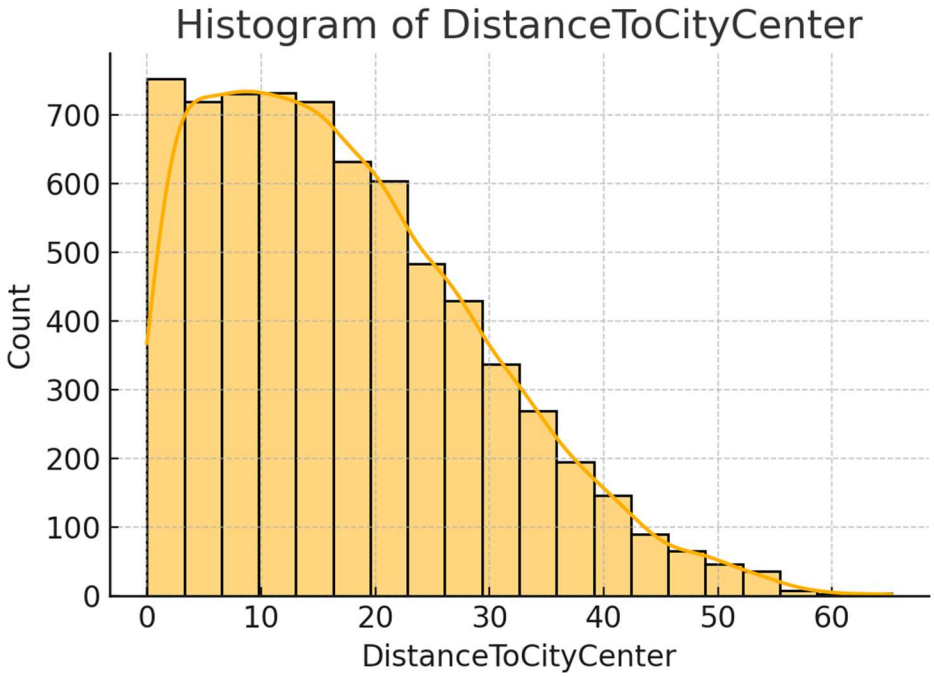
SquareFootage Visualization



SchoolRating Visualization



DistanceToCityCenter Visualization



AgeOfHome Visualization